Document Image Layout Analysis via Explicit Edge Embedding Network

Xingjiao Wu[1]*, Yingbin Zheng[2]*, Tianlong Ma[1]†, Hao Ye[2], Liang He[1]†
[1]East China Normal University, Shanghai, China [2]Videt Lab, Shanghai, China

Abstract

Layout analysis from a document image plays an important role in document content understanding and information extraction systems. While many existing methods focus on learning knowledge with convolutional networks directly from color channels, we argue the importance of high-frequency structures in document images, especially edge information. In this paper, we present a novel document layout analysis framework with the Explicit Edge Embedding Network. Specifically, the proposed network contains the edge embedding block and dynamic skip connection block to produce detailed features, as well as a lightweight fully convolutional subnet as the backbone for the effectiveness of the framework. The edge embedding block is designed to explicitly incorporate the edge information from the document images. The dynamic skip connection block aims to learn both color and edge representations with learnable weights. In contrast to the previous methods, we harness the model by using a synthetic document approach to overcome data scarcity. The combination of data augmentation and edge embedding is important toward a more compact representation than directly using the training images with only color channels. We conduct experiments using the proposed framework on three document layout analysis benchmarks and demonstrate its superiority in terms of effectiveness and efficiency over previous approaches.

1. Introduction

Document layout analysis (DLA) aims to divide a document image into different regions, such as text, figures, and tables. Analysis of the layout from the document image plays an important role in document content understanding and information extraction applications, such as document understanding [48], knowledge extraction [7, 39], handwriting recognition [6], and biomedical event extraction [50]. A modern DLA system usually consists of page segmentation and logical structure analysis steps, and great progress has been achieved in recent years [5].

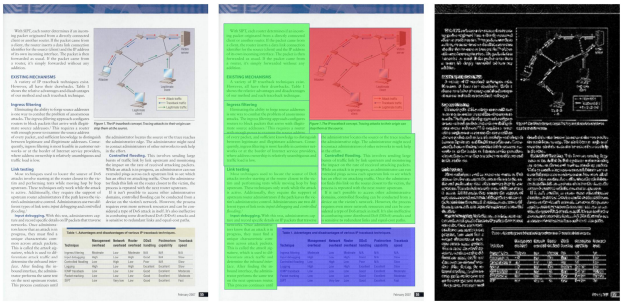


Figure 1: Left: the original document images. Middle: ground-truth of the layouts (segmentation label colors are: figure, table, and text). Right: edges extracted by the Laplacian edge detectors [40].

Accurately estimating the content categories in a document is still a challenging task due to the gap between the high-level semantic and low-level visual contents of the documents. While many existing methods focus on learning knowledge with convolutional networks directly from color channels, we argue the importance of high-frequency structures in document images, especially edge information. The edges can provide skeleton information that is useful to understand the document structure. Specifically, the edges usually contain the classification attributes of image regions and make the characteristics of document layout more prominent. An example is shown in Fig. 1, where the edge of text regions has a dense texture, the edge of figures is relatively smooth, and the edge of tables contains more straight lines. Inspired by these observations, this paper works toward an effective layout analysis framework by incorporating explicit edge knowledge. We design the *Explicit Edge Embedding Network* (E^3Net), which superimposes the edge information onto the image channel to generate a more efficient image input block.

We employ the Fully Convolutional Network (FCN, [24]) as the backbone. FCN is composed of layers that represent high-level and low-level information through the encoding part and then superimposes these features from these response maps onto the decoder. The low-level feature maps tend to contain detailed information, while the high-level feature maps have more semantic information. Explor-

*These authors contributed equally to this work.
†Co-corresponding authors.

ing a network structure that can effectively connect high-level features and low-level features is also crucial. Inspired by the adaption branches strategy [44], we consider dynamically learning the connection structure from the edge representation, and thus we propose the dynamic skip connection block. The core idea of the dynamic skip connection block is to calculate the information gain of the encoding features and add them to the decoding layer by using differently weighted overlaps. We report the impacts of different components as well as the comparison with the state-of-the-art approaches on three challenging DLA benchmarks, *i.e*., DSSE-200 [47], CS-150 [9], and ICDAR2015 [2].

In addition, we augment the data to ensure the universality of the models better. The LaTex is used as a composition engine with contents (images, tables, and text) we prepared to synthetic data. However, the LaTex style cannot generate unnormal printing scale texts. The difference from the previous work is that we use some images that include unconventional texts to replace text to overcome the limitation of the LaTex. To generate more image styles, we use the MS COCO dataset [22] as the image material. Since the MS COCO images are annotated, we can easily obtain the corresponding image description to easily generate the title and description of the image. We use the data synthetic method to generate many samples and provide a closed-loop iteratively update the sample library to realize automatic learning of the model.

Our contributions are summarized as follows.

- For the layout task, we propose explicitly embedding edge information onto the image channels to generate a more efficient image input module. To focus on feature learning, we utilize the edge by generating three edge channels. We propose explicitly embedding edge information onto the image channel to generate a more efficient image input module for the layout task.
- To obtain a universal and effective layout analysis model, we employ the dynamic skip connection on the FCN backbone for the learning of edge representation and improve the data synthetic method for training data generation.
- Extensive evaluations demonstrate the superior performance of the proposed E^3Net. Notably, we achieve state-of-the-art results on three document layout analysis datasets compared with the existing methods. We also conduct an ablation study to evaluate the effect of the edge embedding block and the dynamic skip connection block. Our whole system can process approximately 8 document images per second.

The rest of this paper is organized as follows. Section 2 introduces the background of document layout analysis. Section 3 discusses the model design and network architecture in detail. In Section 4, we demonstrate the qualitative and quantitative study of the framework. Finally, we conclude our work in Section 5.

2. Related Work

Early DLA work can be divided into two categories, *i.e*., top-down and bottom-up strategies [5].

The top-down strategy iteratively divides pages into columns, blocks, lines of text, and words. The representative works belonging to the top-down strategy include texture-based analysis [3], run-length smearing [35], DLA projection-profile [30] and white space analysis [31]. The bottom-up strategy [37, 27, 25, 38] dynamically obtains document analysis results from a small, granular data level. It first uses some local features inherent from the text (such as black and white pixel spacing or connection spacing) to detect individual words and then groups the words into lines of text and paragraphs. The top-down methods and the bottom-up methods deal with common rectangular image layouts that are successful. However, for complex layouts, these methods do not seem to be as effective.

With recent advances in deep convolutional neural networks, several methods based on neural networks have been proposed [14, 46, 20, 41, 52, 51, 34, 19, 45, 43]. For example, He *et al*. [14] used a multiscale network for semantic page segmentation and element contour detection based on three types of document elements (text blocks, tables, and figures). Recently, the DLA task can also be considered a semantic segmentation task, which is to perform a pixel-level understanding of the segmentation object [46, 51, 34, 19]. Xu *et al*. [46] trained the multitask FCN to segment the document image into different regions, and Soullard *et al*. [34] used the FCN for historical newspaper images. Zheng *et al*. [51] further included a deep generative model for graphic design layouts to synthesize layout designs. Li *et al*. [19] proposed the cross-domain DOD model to learn the model for the target domain using labeled data from the source domain and unlabeled data from the target domain. Here many of these papers employ the FCN [24] for semantic segmentation of the document pages. With the help of a full convolution structure, FCN can adapt to any size of the image with the pooling operation, which balances the speed and accuracy. However, it also causes the spatial information from the image to be weakened during the propagation process. To compensate for this problem, we use a skip connection structure to enhance spatial information.

Data augmentation. In addition to improving the network structure, many researchers focus on the expansion of data. Some large-scale datasets with additional tools are proposed, and good results have been achieved through the migration of these datasets. Yang *et al*. [47] proposed a

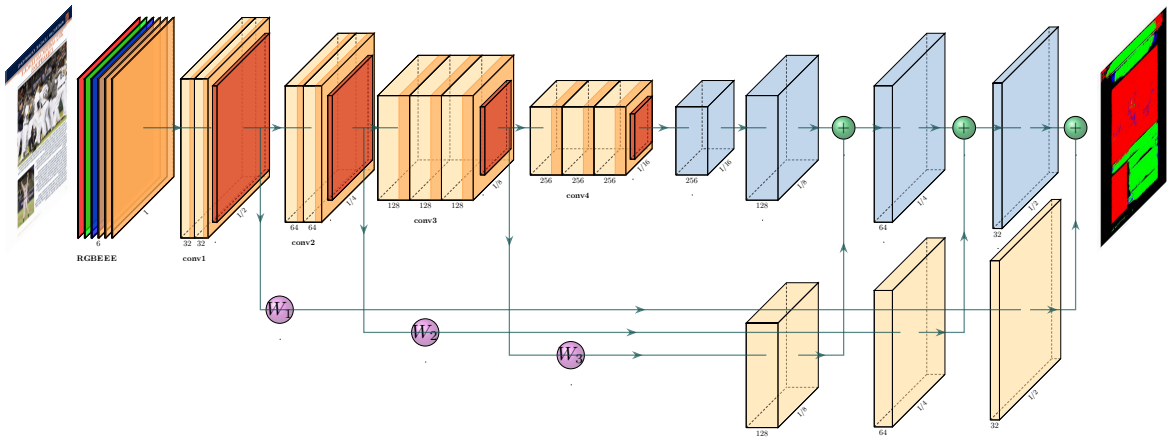


Figure 2: Architecture of the Explicit Edge Embedding Network (E^3Net).

synthetic dataset and an end-to-end multimodal FCN with text embedding for extracting semantic structures from documents. Andreas *et al*. [16] introduced a very challenging dataset of historic German documents of the task of recognizing handwritten documents. Li *et al*. [18] used LayoutGAN to augment the data by generating a different layout. Haurilet *et al*. [13] introduced the SPaSe (slide page segmentation) dataset, which contains dense, pixelwise annotations of 25 classes for 2000 slides. Siegel *et al*. [32] proposed a method to induce high-quality labels to leverage auxiliary data from arXiv and PubMed with no human intervention.

At present, data enhancement methods are mainly divided into the following types, and one is generated by using existing auxiliary data. Such data mainly come from scientific documents, such as arXiv and other network resources. The other is generated through the Generative Adversarial Network (GAN), and finally generated by the LaTex. Using LaTex generation is a simple and effective method, but the text generated in this way is relatively simple due to the LaTex limitation. The older font generation method ignores the relatively large font, and the font of the text is limited. We use some novels as our text source; in this way, we can use the constraint that our basic unit of text is word-level, and we can control the position of the label box more precisely during the process of text synthesis.

3. Method

As shown in Fig. 2, E^3Net is composed of four parts: the edge embedding block (EEB), an encoder structure, a decoder structure, and the dynamic skip connection block (DSC). We show the encoder structure, the decoder structure, and the dynamic skip connection block in Table 1. In this section, we first introduce the edge embedding block and then the dynamic skip connection block. Finally, we introduce the strategy of data synthesis.

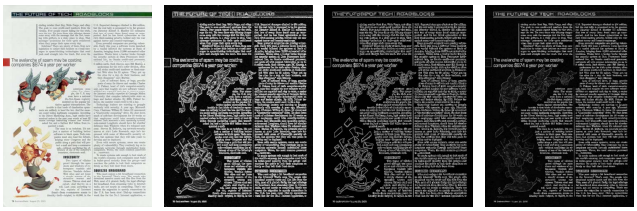


Figure 3: The edge information from edge extraction methods that are used in the E^3Net. From left to right: original image, edge maps extracted by Sobel, Canny, and Laplacian.

3.1. Edge Embedding Block

The edges are direct characterizations of the image and include some categorical properties. The use of edge information for improved image processing/analysis has attracted many researchers to explore due to its excellent performance [23, 42, 1, 36, 12, 26].These edge extraction methods can suppress the noise and ringing artifacts and smooth the staircases.

To embed edge knowledge, we propose the edge embedding block (EEB), which superimposes edge information on the image channel to build a more effective image input block. Different edge extraction algorithms focus on different edge information and edge strengths. As shown in Fig. 3, the edges extracted by one operator cannot represent the overall edge information. In addition, to balance the number of channels, we propose using three different edge extraction operators for edge extraction. In this paper, we use the Sobel edge detector [15], Laplacian edge detector [40] and Canny edge detector [11] to locate sharp intensity changes and to find object boundaries in an image. Previous work, such as [23], achieved good edge detection effects. However, the ablation study in the experiments shows that our combination is more suitable for this architecture.

Algorithm 1: Augment input channels with edge

Input: $I\ (w \times h \times 3)$: Original image channels
Output: $E\ (w \times h \times 6)$: Augmented channels

Obtain image RGB channels I_R, I_G, I_B ;
Obtain the grayscale $I_g = Gray\ (I_R, I_G, I_B)$;
Edge map by Sobel $I_S = Sobel\ (I_g)$;
Edge map by Laplacian $I_L = Laplacian\ (I_g)$;
Edge map by Canny $I_C = Canny\ (I_g)$;
Augmented channels
$E = cat\ \{I_R, I_G, I_B, I_S, I_L, I_C\}$;
return E.

Adding edge information can reduce the image's dependence on color to let the model focus on the learning of features. The input channels are enhanced by explicitly appending edge information to the original image. Algorithm 1 shows the step to generating the augmented channels. We first obtain the RGB channel of the image and process the image into a grayscale image. Then, we use the Sobel, Laplacian, and Canny edge detectors to locate sharp intensity changes and to find object boundaries in an image. Finally, we superimposed the RGB channel and the three edge maps as a 6-channel to accomplish the input. The EEB consists of 6 channels, three of which are the RGB channels of the image, and the remaining three channels are the edge information of the image.

3.2. FCN with Dynamic Skip Connection

It is an important task of learning edge representation knowledge effectively. From the perspective of feature learning, low-level feature maps contain more detailed information, while high-level feature maps contain more semantic information. Traditional FCN cannot explicitly represent features because it only connects the encoder and the decoder for superimposing the feature information [21]. We propose a dynamic skip connection block (DSC) to tackle this problem. The core idea of DSC is calculating the information gain of the encoder feature and adding the information gain to the decoder layer by using different weights. Our method focuses on pixelwise segmentation with a fully convolutional network that uses an edge embedding block and a dynamic skip connection block. We use a lightweight model as the backbone to maintain the model processing speed, and the backbone parameter amounts to only $1/6$ of VGG16 [33]. The backbone is divided into two parts, the encoder and the decoder. The structure of the encoder is shown in the first column of Table 1. It uses the 3×3 convolution kernel and uses four max-pooling, and the encoder will reduce the image to 1/16 of the original. The decoder structure is shown in the third column of Table 1. The decoder order is 256-128-64-32-16 from bottom to top, and each decoding layer is composed of the deconvolution, the ReLU activation function, and the batch regularization.

The dynamic skip connection block is a learnable connection operation added based on U-Net [29]. The high-level features and low-level features will carry different information intensities, but the traditional U-Net directly connects to the encoder and decoder. This connected method cannot distinguish high-level information and low-level information well. For more effective use of information in different dimensions, we connect the high-level features and the low-level features using a learnable unit. The dynamic skip connection block is structured with three parallel pathways. The first pathway is designed for low-level feature fusion with the structure is: GAP(1)-FC(32, 4)-RELU-FC(4, 32)-Sigmoid. The second pathway structure is: GAP(1) - FC(64, 8) - RELU - FC(8, 64) - Sigmoid. The third pathway is designed for high-level feature fusion , and its structure is: GAP(1)-FC(128, 16)-RELU-FC(16, 128)-Sigmoid. Where 'GAP' represents a Global Average Pooling layer, 'RELU' represents a rectified linear unit, 'FC' represents the fully connected layer and 'Sigmoid' represents the sigmoid activation function. The numbers in the parentheses of FC are the input parameter and the output parameter. Each pathway outputs a regularized weight, and the sum of the weights of the three pathways is 1. After obtaining the weight coefficients, we multiply the feature layer obtained by the encoder by the corresponding weight coefficient and superimpose it on the corresponding feature layer of the decoder.

Loss. Cross-entropy loss is used as the loss function:

$$\begin{aligned}\mathcal{L}(x, label) &= -w_{label} \log \frac{e^{x_{label}}}{\sum_{j=1}^{N} e^{x_j}} \\ &= w_{label} \left(-x_{label} + \log \sum_{j=1}^{N} e^{x_j} \right),\end{aligned} \tag{1}$$

where $x \in \mathbb{R}^N$ is the activation value without softmax, N is the feature dimension of x, $label \in [0, C-1]$ is the scalar of the corresponding label, C is the number of classifications to be classified, and $w \in \mathbb{R}^C$ is the label weight.

3.3. Synthetic Document Data

The prerequisite for training a universal model is to provide enough data. At present, the annotation data of the document layout analysis task are limited, so we improved the data synthesis method proposed in [47]. Compared with previous work, our data synthesis method introduces more text elements. We add some special text images to make the generated samples more natural and realistic. In addition, we propose a semiautomatic man-machine hybrid labeling mode to provide more diverse data sources.

The document synthetic can be seen as a simple jigsaw puzzle, and we will add table, figure, and text to an A4 format document. We use LaTex to generate pdf by combining

Table 1: Configuration of the backbone. All convolutional layers use padding to maintain the previous size. The convolutional layer parameters are denoted as conv-(kernel size)-(number of filters)-(dilation rate), and max-pooling layers are conducted over a 2 pixel window with stride 2. Here, deconv, conv, pool, and FC represent the deconvolution layer, convolution layer, max-pooling layer, and fully connected layer, respectively.

encoder ⇓	dynamic skip connection	decoder ⇑
conv3-6-1 conv3-32-1 conv3-32-1 max-pooling	GAP FC-32-4 RELU FC-4-32 Sigmoid	deconv3-32-1 RELU BN(64) deconv3-16-1 conv1-4-1
conv3-64-1 conv3-64-1 max-pooling	GAP FC-64-8 RELU FC-8-64 Sigmoid	deconv3-64-1 RELU BN(32)
conv3-128-1 conv3-128-1 conv3-128-1 max-pooling	GAP FC-128-16 RELU FC-16-128 Sigmoid	deconv3-128-1 RELU BN(64)
conv3-256-1 conv3-256-1 conv3-256-1 max-pooling		deconv3-256-1 RELU BN(128)

images, tables, and texts. We prepare the necessary material by collecting figures and tables from web resources. Moreover, to enrich the image information, we use some images from MS COCO [22] and randomly add the corresponding image title. Due to the limitations of the type of text generated by LaTex, we not only directly use the novel as text sources to generate pdf but also include unconventional text images as text sources. Using these text sources can overcome the limitation of LaTex. Using the novel material to constrain the minimum unit is wordlevel. Constraint minimum units can avoid format problems that appear by using some network resources (for example, a paragraph without spaces or some meaningless text resources). Using LaTex to generate the pdf, we can easily obtain the corresponding label. We synthesized images, as shown in Fig. 4, and we can see that our synthesis data are very similar to the real document images.

The E^3Net has been able to obtain good results in practical problems. Furthermore, we want to provide a better user experience. Therefore, we propose a semiautomatic hybrid data annotation strategy using a human-machine hybrid. Our training process can regard as a closed-loop. We

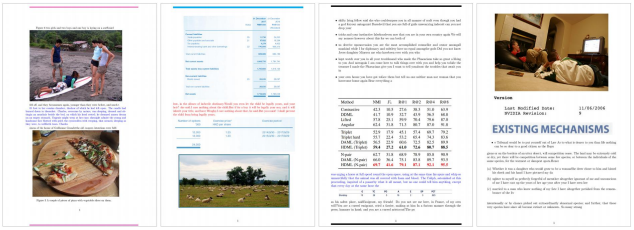


Figure 4: Sample synthetic documents.

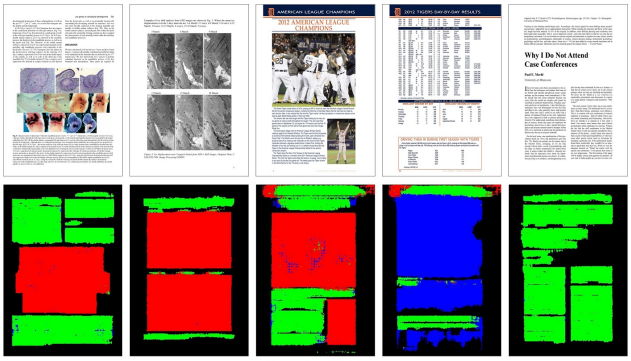
use synthetic data to train E^3Net and test it after a random epoch. The test data are unlabeled; furthermore, we cannot obtain the specific indicators, but we can distinguish the table using edge detection. We use edge detection to obtain a table area, and we use the table area as the masking label. We will compare it with the classified prediction map. We will choose high error rate images and annotate this image. We split these images into different elements (table and figure) and put them to the data source to generate new data samples for retraining. Specifically, we first input unlabeled data into the layout model and obtain the predicted result. In addition, the unlabeled data will be input to a nonartificial intelligence algorithm (the table detection algorithm base rule) to obtain the table area. We will compare the table information predicted by the two algorithms. We will choose those inconsistent data (*i.e.*, data with a degree of difference of more than 60%) and manually label those data into the data pool. We split out the elements (table and figure) in these data and put them into the data generation model to generate more new data.

4. Evaluation and Discussion

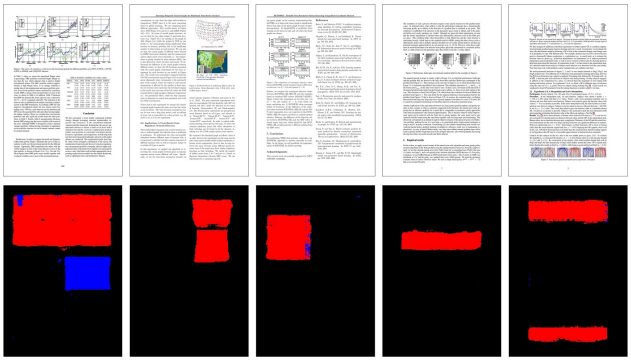
We evaluate the proposed E^3Net on three document layout analysis benchmarks: DSSE-200 [47], CS-150 [9], and ICDAR2015 [2]. We first introduce the experimental configuration. Then we show the qualitative results and compare E^3Net with prior works. Finally, we consider the ablation study on DSSE-200 to evaluate the effect of the dynamic skip connection block and the edge embedding block.

4.1. Configurations

Categories and model training. There is no unified standard on the classification for layout at present. Many previous works divided the layout into three categories: figure, tables, and others. Some works are divided into the following seven categories: figure, table, paragraph, background, caption, lists, and section. However, if only the figure and tables are classified, we cannot effectively use the text and background information. If there are too many categories, it will be more cumbersome for the layout work. This paper makes a trade-off and considers the following four categories: text, figure, table, and background. We fine-tune the



(a) DSSE-200



(b) CS-150

(c) ICDAR2015

Figure 5: Example real documents (top) and their corresponding segmentation predictions (bottom) on three datasets. Segmentation label colors are: figure, table, text, background (for DSSE-200 and ICDAR2015) and non-text (for CS-150).

model by randomly select 10% of the target dataset as the train data, and then we reduce the learning rate to 1/10 of the original learning rate. To prevent overfitting due to too little data, we split the elements (tables and figures) of the data and then put these elements into the LaTex document synthesis engine for data expansion. We use these data to fine-tune the model.

Metric. Several metrics are used to evaluate the performance. We first define M as the $n \times n$ confusion matrix with n categories. *Accuracy* (*Acc*) is the ratio of the pixels that are correctly predicted in a given image, *i.e.*,

$$Acc = \frac{\sum_i M_{ii}}{\sum_{ij} M_{ij}} \tag{2}$$

Precision (*P*) is the ratio that is actually a positive example in the example that is divided into positive examples, *i.e.*,

$$P = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad P_i = \frac{M_{ii}}{\sum_j M_{ji}} \tag{3}$$

Recall (*R*) measures the coverage. There are multiple positive examples of metrics that are divided into positive examples, *i.e.*,

$$R = \frac{1}{n}\sum_{i=1}^{n} R_i \qquad R_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{4}$$

F_1 is an indicator used to measure the accuracy of a binary model. It also takes into account the accuracy and recall rate of the classification model. The F_1 score can be seen as a weighted average of model accuracy and recall:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

MIoU is the mean intersection-overuunion of each foreground category, *i.e.*,

$$MIoU = \frac{1}{n+1}\sum_{i=0}^{n} \frac{M_{ii}}{\sum_{j=0}^{n} M_{ij} + \sum_{j=0}^{n} M_{ji} - M_{ii}} \tag{6}$$

Datasets. We employ three benchmarks for evaluation. *DSSE-200* [47] is a comprehensive dataset that includes various dataset styles. It contains 200 images, including pictures, PPT, brochure documents, old newspapers, and scan files with light changes. *CS-150* [9] is a dataset consisting of 150 papers. *CS-150* is divided into three categories, images, tables, and others, consisting of 1175 samples. *ICDAR2015* focuses on appearance-based regions [2]. It consists of a magazine and journal that contain 7 training sets and 70 tests. *ICDAR2015* is not a simple rectangular segmentation and is directly embedded in the paragraph. The sample dataset are illustrated in the top rows of Fig. 5.

4.2. Qualitative Results

DSSE-200. We use the synthetic dataset to train the network and make predictions on the DSSE-200 dataset. The overall performance of E^3Net is accuracy 0.82, precision 0.79, recall 0.73, F_1 0.76, and MIoU 0.57; the confusion matrix is illustrated in Fig. 6-Left. We can observe that

Table 2: Per-category comparison based on IoU scores (%) on the DSSE-200. FT indicates the model with fine-tuning.

Method	background	figure	table	section	caption	list	paragraph	mean
MFCN [47]	83.9	83.7	79.7	59.4	**61.1**	68.4	**79.3**	73.3
E^3Net	95.9	88.8	90.7	**89.8**	41.6	**71.2**	56.7	76.3
E^3Net (FT)	**96.5**	**96.1**	**93.0**	77.0	50.4	60.6	68.3	**77.4**

Table 3: Comparing E^3Net with previous network structures on the DSSE-200 and CS-150 datasets.

Method	#Parameters	DSSE-200					CS-150				
		Acc	P	R	F_1	MIoU	Acc	P	R	F_1	MIoU
SegNet [4]	29M	0.76	0.71	0.72	0.71	0.49	0.76	0.71	0.72	0.71	0.49
PANet [17]	168M	0.79	0.74	0.72	0.73	0.53	0.96	0.82	0.91	0.87	0.52
PSPNet [49]	46M	0.72	0.69	**0.79**	0.74	0.51	0.96	0.84	0.97	0.90	0.63
DV3+ [8]	53M	0.78	0.72	0.75	0.73	**0.64**	0.96	0.81	0.97	0.88	0.63
E^3Net	3M	**0.82**	**0.79**	0.73	**0.76**	0.57	**0.96**	0.85	**0.97**	**0.91**	**0.64**

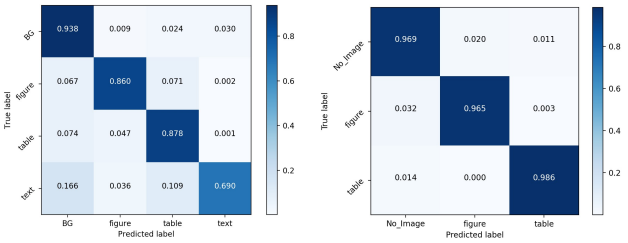


Figure 6: Confusion matrices for DSSE-200 (left) and CS-150 (right).

Table 4: Per-category comparison based on CS-150.

Method	figure			table		
	P	R	F_1	P	R	F_1
Praczyk *et al.* [28]	0.624	0.500	0.555	0.429	0.363	0.393
Clark *et al.* [10]	0.961	0.911	0.935	0.962	0.921	0.941
Clark *et al.* [9]	0.980	0.961	0.970	**0.979**	0.963	0.971
E^3Net	0.938	**0.972**	0.956	0.834	**0.988**	0.905
E^3Net (FT)	**0.986**	0.970	**0.978**	0.971	0.977	**0.973**

E^3Net has excellent recognition rates for backgrounds, figures, and tables, as the edge information is used to improve background discrimination. However, the ability to recognize text is slightly lower than other categories, *e.g.*, some text pixels are recognized as the table, probably because the contents for text and table are quite similar. The sample documents and their corresponding predictions in DSSE-200 are shown in Fig. 5(a). In general, the document layouts are correctly extracted, and the border of the regions can be refined with postprocessing steps such as connected component analysis.

CS-150. We follow the same step in DSSE-200 to conduct an experiment of CS-150, and the results are accuracy 0.96, precision 0.85, recall 0.97, F_1 0.91, and MIoU 0.64. The performance is good in CS-150 for both the overall metrics and the per-category results (as shown in the confusion matrix of Fig. 6-Right). The CS-150 dataset is entirely composed of scientific papers, and the layout is relatively simple. We demonstrate some document images and the corresponding predictions of CS-150 in Fig. 5(b).

Table 5: Per-category comparison based on IoU scores (%) on ICDAR2015.

Method	non-text	text	figure	mean
MFCN [47]	**94.5**	**91.0**	77.1	87.53
E^3Net	81.6	79.1	85.0	81.87
E^3Net (FT)	90.1	88.3	**93.5**	**90.59**

ICDAR2015. We also conduct qualitative results for ICDAR2015, as illustrated in Fig. 5(c). With the help of an edge embedding network, our methods can successfully classify most of the pixels into layout categories with different backgrounds and visual contents. In addition, we can see that E^3Net has a successful effect of dealing with the figure that is directly embedded in the text paragraph.

4.3. Comparison with Prior Arts

To evaluate our model, we compare it with state-of-the-art document layout analysis methods, which also make use of image content as input. We follow the settings and evaluation protocols of [47] (for DSSE-200 and ICDAR2015) and [9] (for CS-150).

For the **DSSE-200** dataset, we can see that our results are more effective than those in [47] (Table 2). It is worth

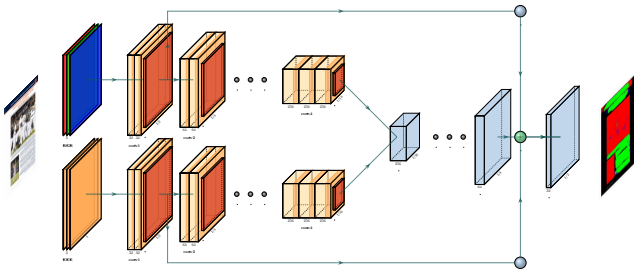


Figure 7: Network for two-stream fusion.

Table 6: Comparison with two-stream fusion.

Method	Acc	P	R	F_1	MIoU
$E^3Net_{\text{w/oEdge}}$	0.70	0.75	0.62	0.68	0.46
Two-stream fusion	0.74	0.72	0.66	0.69	0.53
E^3Net	0.82	0.79	0.73	0.76	0.57

noting that the edge contains more discriminative information for background regions; therefore, E^3Net can obtain good results for background recognition even without fine-tuning. E^3Net has a better recognition effect for figures, tables, and sections, and the mean score improves 4% compared with [47]. For the **ICDAR2015** dataset, as listed in Table 5, E^3Net with fine-tuning also achieves a 3% mean score improvement over [47]. In the comparison for **CS-150**, we use the precision, recall, and F_1 as the metric, and the results in Table 4 show that ours outperform previous approaches for both the figure and table categories. Specifically, when we use E^3Net trained from synthetic documents without fine-tuning, the overall performance is comparable, and the recall rate is high. Fine-tuning brings the distribution of the CS-150 data and improves the precision of the whole network.

As mentioned in the previous sections, E^3Net is designed based on an FCN-like backbone. Here, we also compare our approach with the state-of-the-art network for the general semantic segmentation task. Table 3 reports the performance on the DSSE-200 and CS-150 datasets with different metrics. We can observe that the proposed E^3Net achieves better results, while the parameter size is much smaller than others.

4.4. Ablation Study

In this section, we perform an ablation study on the DSSE-200 dataset. We start by exploring the variations of the network architecture to find the optimal set of fusion strategies. Then the components in our frameworks, such as the edge embedding block and the skip connection, are evaluated. Throughout the experiments, we use $E^3Net_{\text{w/oX}}$ to represent the network of E^3Net without component X for presentation simplicity.

Table 7: Evaluation of different edge embedding settings. LSB indicates the edge embedding block by Laplacian, Sobel, and bilateral filter.

Method	Acc	P	R	F_1	MIoU
E^3Net	0.82	0.79	0.73	0.76	0.57
$E^3Net_{\text{w/oEdge}}$	0.70	0.75	0.62	0.68	0.46
E^3Net (Sobel)	0.78	0.73	0.77	0.75	0.58
E^3Net (LSB)	0.78	0.73	0.76	0.75	0.51

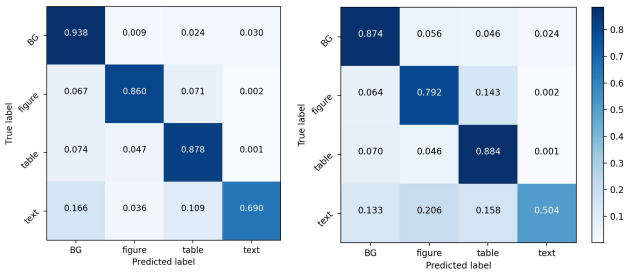


Figure 8: Confusion matrix for DSSE-200. Left: E^3Net, Right: E^3Net (LSB).

Table 8: Evaluation of the dynamic skip connection.

Method	Acc	P	R	F_1	MIoU
E^3Net	0.82	0.79	0.73	0.76	0.57
$E^3Net_{\text{w/oDSC}}$	0.73	0.69	0.66	0.67	0.50
$E^3Net_{\text{w/oEdge}}$	0.70	0.75	0.62	0.68	0.46
$E^3Net_{\text{w/oEdge\&DSC}}$	0.67	0.63	0.64	0.64	0.48

Model architecture. The design of an effective network is of great importance for model learning. In this paper, we fuse the color channels and the edge explicitly into the augmented input, and another potential approach is to treat them as two independent steams and fuse them in the last few layers. Fig. 7 demonstrates the architecture of two-stream fusion, which consists of two branches, two independent encoders, and one decoder for the fusion. We list the comparison in Table 6. Adding edge information under a two-stream framework improves the performance (for accuracy, F_1, and MIoU). However, compared with E^3Net, the effect is limited. This demonstrates that fusion at an early stage can take full advantage of the complementarity of color and edge clues.

Edge embedding. In this section, we verify the effect of the EEB block. As shown in Table 7, removing the edge embedding causes a drop in all metrics, *e.g.*, 12% for accuracy and 11% MIoU. We also compare different edge embedding settings. The first is to use the single-channel Sobel edges (E^3Net (Sobel) in Table 7). The results show that E^3Net outperforms E^3Net (Sobel) by using more edge represen-

tation, probably because of the complementarity of different edge detectors. The second group of experiments involves changing the edge detector in the EEB to other detectors. Here, we replace Canny with the bilateral filter and build the model of E^3Net (LSB). From the table, we can see that this combination is not as good as the original E^3Net. We make the confusion matrix for both networks (Fig. 8) and find that the E^3Net with Laplacian, Sobel, and Canny edge detectors has significantly better representation for the texts and figures.

Dynamic skip connection. Incorporating the skip connection has been proven to be useful for many computer vision tasks, and we wonder whether it can promote a document layout analysis system. As shown in Table 8, the substantial performance gains over $E^3Net_{\text{w/oDSC}}$ confirm the effectiveness of using the dynamic skip connection for the DLA task. Although adding the DSC into a traditional FCN without edges also improves the performance (Table 8, rows 3 and 4), the network combined with DSC and edge embedding has been improved on a larger scale and is able to show the powerful descriptive ability of document layouts.

Speed. The proposed framework is trained and evaluated on a GPU. Inference using an image with a size of 512×384 pixels takes 0.12 seconds with a single Nvidia Titan Xp, meaning that our whole system can generally process approximately 8 document images per second.

5. Conclusions

In this paper, we presented a novel solution for constructing a model of universal document layout analysis. Our approach explored the use of the dynamic skip connection block and edge information to improve the model structure and the construction of a complete synthetic data scheme. We present a dynamic skip connection block that can be dynamically provisioned based on specific instances. We use the edge embedding block to let the model more focused on text content. In addition, we discuss the feasibility of the fusion strategy with the edge. Experimental comparisons with the state-of-the-art approaches on DSSE-200, CS-150, and ICDAR2015 showed the effectiveness and efficiency of our proposed E^3Net for the document layout analysis task.

References

[1] D. Acuna, A. Kar, and S. Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11075–11083, 2019. 3

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. ICDAR2015 competition on recognition of documents with complex layouts-rdcl2015. In *IAPR International Conference on Document Analysis and Recognition*, pages 1151–1155, 2015. 2, 5, 6

[3] A. Asi, R. Cohen, K. Kedem, and J. El-Sana. Simplifying the reading of historical manuscripts. In *IAPR International Conference on Document Analysis and Recognition*, pages 826–830, 2015. 2

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 6

[5] G. M. Binmakhashen and S. A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Computing Surveys*, 52(6):109, 2019. 1, 2

[6] G. M. BinMakhashen and S. A. Mahmoud. Historical document layout analysis using anisotropic diffusion and geometric features. *International Journal on Digital Libraries*, pages 1–14, 2020. 1

[7] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020. 1

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018. 6

[9] C. Clark and S. Divvala. Pdffigures 2.0: Mining figures from research papers. In *ACM/IEEE on Joint Conference on Digital Libraries*, pages 143–152, 2016. 2, 5, 6, 7

[10] C. A. Clark and S. Divvala. Looking beyond text: Extracting figures, tables and captions from computer science papers. In *Workshops at AAAI Conference on Artificial Intelligence*, 2015. 7

[11] L. Ding and A. Goshtasby. On the canny edge detector. *Pattern Recognition*, 34(3):721–725, 2001. 3

[12] Z. Fu, T. Ma, Y. Zheng, H. Ye, J. Yang, and L. He. Edge-aware deep image deblurring. *arXiv:1907.02282*, 2019. 3

[13] M. Haurilet, Z. Al-Halah, and R. Stiefelhagen. Spase-multi-label page segmentation for presentation slides. In *IEEE Winter Conference on Applications of Computer Vision*, pages 726–734, 2019. 3

[14] 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 *IAPR International Conference on Document Analysis and Recognition*, 2017. 2

[15] J. Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42, 1983. 3

[16] A. Kölsch, A. Mishra, S. Varshneya, M. Z. Afzal, and M. Liwicki. Recognizing challenging handwritten annotations with fully convolutional networks. In *International Conference on Frontiers in Handwriting Recognition*, pages 25–31, 2018. 3

[17] H. Li, P. Xiong, J. An, and L. Wang. Pyramid attention network for semantic segmentation. In *British Machine Vision Conference*, 2018. 6

[18] J. Li, J. Yang, A. Hertzmann, J. Zhang, and T. Xu. Layoutgan: Generating graphic layouts with wireframe discriminators. In *International Conference on Learning Representations*, 2019. 3

[19] K. Li, C. Wigington, C. Tensmeyer, H. Zhao, N. Barmpalios, V. I. Morariu, V. Manjunatha, T. Sun, and Y. Fu. Cross-domain document object detection: Benchmark suite and method. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12915–12924, 2020. 2

[20] Y. Li, Y. Zou, and J. Ma. Deeplayout: A semantic segmentation approach to page layout analysis. In *International Conference on Intelligent Computing (ICIC)*, pages 266–277, 2018. 2

[21] C. Lin, S. Zhuang, S. You, X. Liu, and Z. Zhu. Real-time foreground object segmentation networks using long and short skip connections. *Information Sciences*, 2021. 4

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 2, 5

[23] H. Liu, R. Xiong, Q. Song, F. Wu, and W. Gao. Image super-resolution based on adaptive joint distribution modeling. In *IEEE Visual Communications and Image Processing*, 2017. 3

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1, 2

[25] Y. Lu and C. L. Tan. Constructing area voronoi diagram in document images. In *IAPR International Conference on Document Analysis and Recognition*, pages 342–346, 2005. 2

[26] G. Mandal and D. Bhattacharjee. Learning-based single image super-resolution with improved edge information. *Pattern Recognition and Image Analysis*, 30(3):391–400, 2020. 3

[27] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot. Texture feature benchmarking and evaluation for historical document image analysis. *International Journal on Document Analysis and Recognition*, 20(1):1–35, 2017. 2

[28] P. A. Praczyk and J. Nogueras-Iso. Automatic extraction of figures from scientific publications in high-energy physics. *Information Technology and Libraries*, 32(4):25–52, 2013. 7

[29] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 4

[30] F. Shafait and T. M. Breuel. The effect of border noise on the performance of projection-based page segmentation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):846–851, 2010. 2

[31] F. Shafait, J. Van Beusekom, D. Keysers, and T. M. Breuel. Background variability modeling for statistical layout analysis. In *International Conference on Pattern Recognition*, pages 1–4, 2008. 2

[32] N. Siegel, N. Lourie, R. Power, and W. Ammar. Extracting scientific figures with distantly supervised neural networks. In *ACM/IEEE on Joint Conference on Digital Libraries*, pages 223–232, 2018. 3

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[34] Y. Soullard, P. Tranouez, C. Chatelain, S. Nicolas, and T. Paquet. Multi-scale gated fully convolutional densenets for semantic labeling of historical newspaper images. *Pattern Recognition Letters*, 131:435–441, 2020. 2

[35] W. Swaileh, K. A. Mohand, and T. Paquet. Multi-script iterative steerable directional filtering for handwritten text line extraction. In *IAPR International Conference on Document Analysis and Recognition*, 2015. 2

[36] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *International Conference on Computer Vision*, pages 5229–5238, 2019. 3

[37] T. A. Tran, I.-S. Na, and S.-H. Kim. Hybrid page segmentation using multilevel homogeneity structure. In *International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2015. 2

[38] N. Vasilopoulos and E. Kavallieratou. Complex layout analysis based on contour classification and morphological operations. *Engineering Applications of Artificial Intelligence*, 65:220–229, 2017. 2

[39] K. Vyas and F. Frasincar. Determining the most representative image on a web page. *Information Sciences*, 512:1234–1248, 2020. 1

[40] X. Wang. Laplacian operator-based edge detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):886–890, 2007. 1, 3

[41] C. Wick and F. Puppe. Fully convolutional neural networks for page segmentation of historical document images. In *IAPR International Workshop on Document Analysis Systems*, pages 287–292, 2018. 2

[42] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2129–2138, 2018. 3

[43] X. Wu, Z. Hu, X. Du, J. Yang, and L. He. Document layout analysis via dynamic residual feature fusion. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2021. 2

[44] X. Wu, Y. Zheng, H. Ye, W. Hu, T. Ma, J. Yang, and L. He. Counting crowds with varying densities via adaptive scenario discovery framework. *Neurocomputing*, 397:127–138, 2020. 2

[45] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. 2

[46] Y. Xu, F. Yin, Z. Zhang, and C.-L. Liu. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1057–1063, 2018. 2

[47] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2017. 2, 4, 5, 6, 7, 8

[48] C. Yuan, H. Huang, C. Feng, G. Shi, and X. Wei. Document-level relation extraction with entity-selection attention. *Information Sciences*, 568:163–174, 2021. 1

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 6

[50] W. Zhao, J. Zhang, J. Yang, T. He, H. Ma, and Z. Li. A novel joint biomedical event extraction framework via two-level modeling of documents. *Information Sciences*, 550:27–40, 2021. 1

[51] X. Zheng, X. Qiao, Y. Cao, and R. W. Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[52] Y. Zheng, S. Kong, W. Zhu, and H. Ye. Scalable document image information extraction with application to domain-specific analysis. In *IEEE International Conference on Big Data*, 2019. 2