

# Supplementary Materials: “DIG: Complex Layout Document Image Generation with Authentic-looking Text for Enhancing Layout Analysis”

Anonymous Author(s)

## CCS CONCEPTS

• Applied computing → Multi / mixed media creation; • Computing methodologies → Object detection.

## KEYWORDS

Complex layout document, Image generation, Document layout analysis, Authentic-looking text, Multimodal pre-training

## 1 MORE IMPLEMENTATION DETAILS

Since the inaccurate annotation of DSSE-200, we use labelme<sup>1</sup> to manually annotate again, still following the rule of dividing document components into six categories: Text, Title, Figure, Table, Caption and List.

The PRImA dataset divides the document components into ten categories: Text, Separator, Noise, Image, Graphic, Chart, LineDrawing, Table, Maths, and Frame. However, as Separator, Noise, LineDrawing, and Frame do not carry substantial information, we exclude them from our detection targets. According to the definition of the PRImA dataset, "Image" represents natural images and "Graphic" represents geometric shapes. Due to the lack of a strict distinction between "Image" and "Graphic," we merge "Graphic" into the "Image" category. Therefore, the PRImA dataset we used in the experiment divides document components into five categories: Text, Image, Chart, Table, and Maths. After cleaning the data to address issues such as mislabeling and inaccuracies, we retained 368 images. The statistics of document components in the DSSE-200 and PRImA datasets are depicted in detail in Table 1 and Table 2, respectively.

Through both pre-training and fine-tuning of ControlNet, all images are resized to 768\*768 pixels. During the generation of images, we set the resolution to 768\*768, with a guidance scale of 9.0, control strength of 2.0, and sampling step of 50. Apart from the text prompts generated based on layout, we do not set any additional positive or negative prompts.

## 2 MODEL PERFORMANCE WITHOUT USING THE SCORING FUNCTION

We test the performance of models trained only on generated data, joint training, and pre-training without using the scoring function, along with the changes in mAP compared to using the scoring function. According to the results shown in Table 3, the mAP score significant decrease across all experiments. When trained solely with *syn* or *syn\_layout*, the improvement given by the scoring function is similar, as there is no real data for referencing during scoring. However, when trained jointly, the improvement given by the scoring function on *syn\_layout* evidently higher than on *syn*.

<sup>1</sup><https://github.com/labelmeai/labelme>

Table 1: Statistics of document components in the DSSE-200 dataset.

Text	Title	Figure	Table	Caption	List	Total
1481	696	281	86	110	223	2877

Table 2: Statistics of document components in the PRImA dataset.

Text	Image	Chart	Table	Math	total
6522	452	35	44	37	7090

Table 3: The layout analysis performance of models trained without scoring function.

Dataset	Training data	mAP↑		$\Delta$
		w	w/o	
DSSE-200	Syn(10)	45.72	44.56	↓1.16
	Syn_layout(10)	36.72	35.19	↓1.53
	Real+Syn(10)	55.23	54.14	↓1.09
	Real+Syn_layout(10)	53.82	50.15	↓3.67
	Syn_layout(10)→Real+Syn(10)	56.07	54.67	↓1.40
PRImA	Syn(10)	55.32	54.09	↓1.23
	Syn_layout(10)	32.91	31.33	↓1.58
	Real+Syn(10)	59.50	57.81	↓1.69
	Real+Syn_layout(10)	56.25	54.01	↓2.25
	Syn_layout(10)→Real+Syn(10)	62.26	59.86	↓2.40

This is because *syn\_layout* contains more noisy layouts compared to *syn*. And with real data as a reference, the scoring function is able to filter out more low-quality regions from *syn\_layout*, thereby benefiting more from the filtering mechanism.

## 3 VISUALIZATION OF IMAGES GENERATED FROM ORIGINAL LAYOUTS

We present the images generated by DIG using the original layouts from the DSSE-200 and PRImA datasets in Figure 1 and Figure 2, respectively.

## 4 VISUALIZATION OF IMAGES GENERATED FROM NEW LAYOUTS

We present the images generated by DIG using the new layouts generated from the DSSE-200 and PRImA datasets in Figure 3 and Figure 4, respectively.

## Real Images

## Layouts

## Generated Images

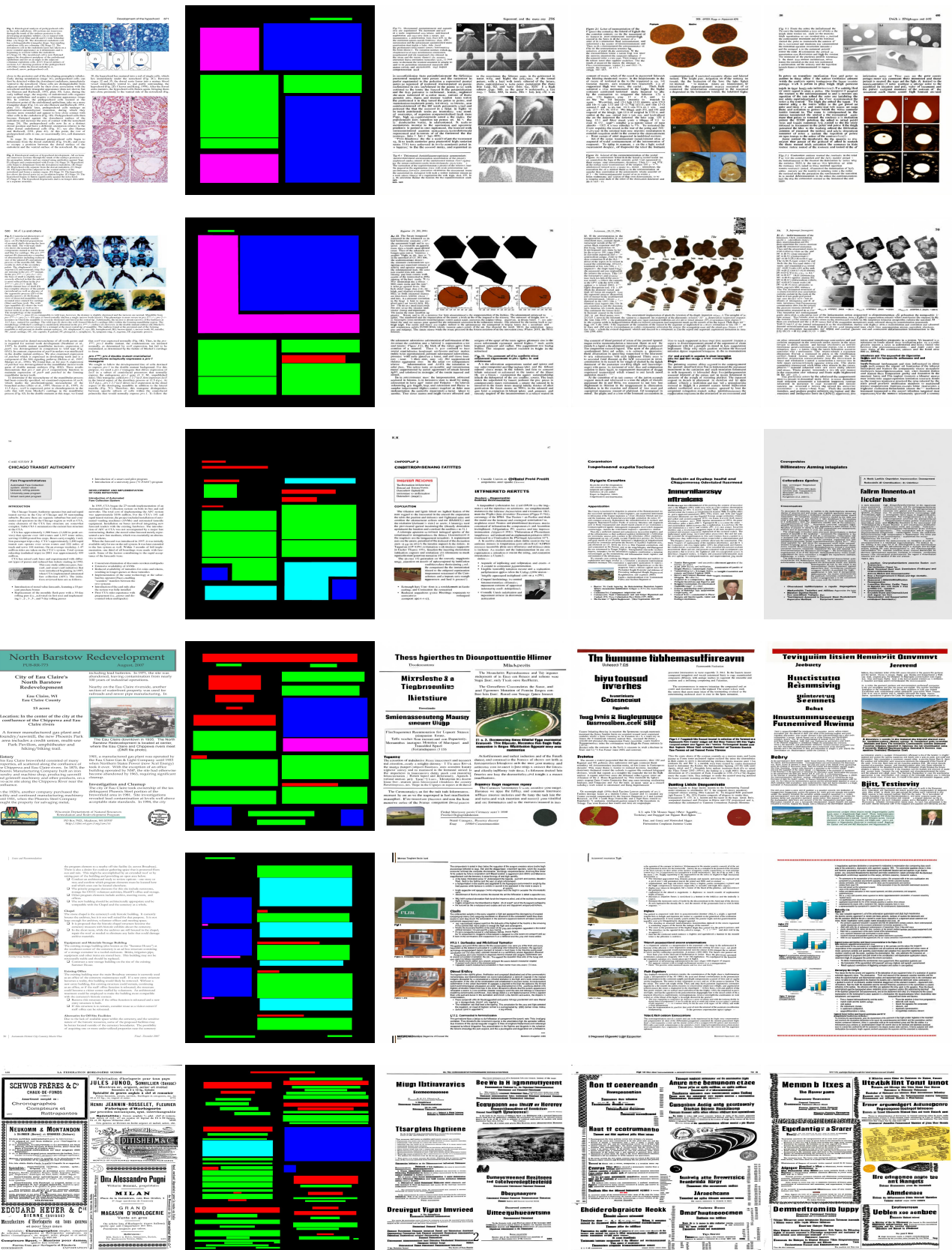


Figure 1: The Generated document images based on the original layouts from DSSE-200.

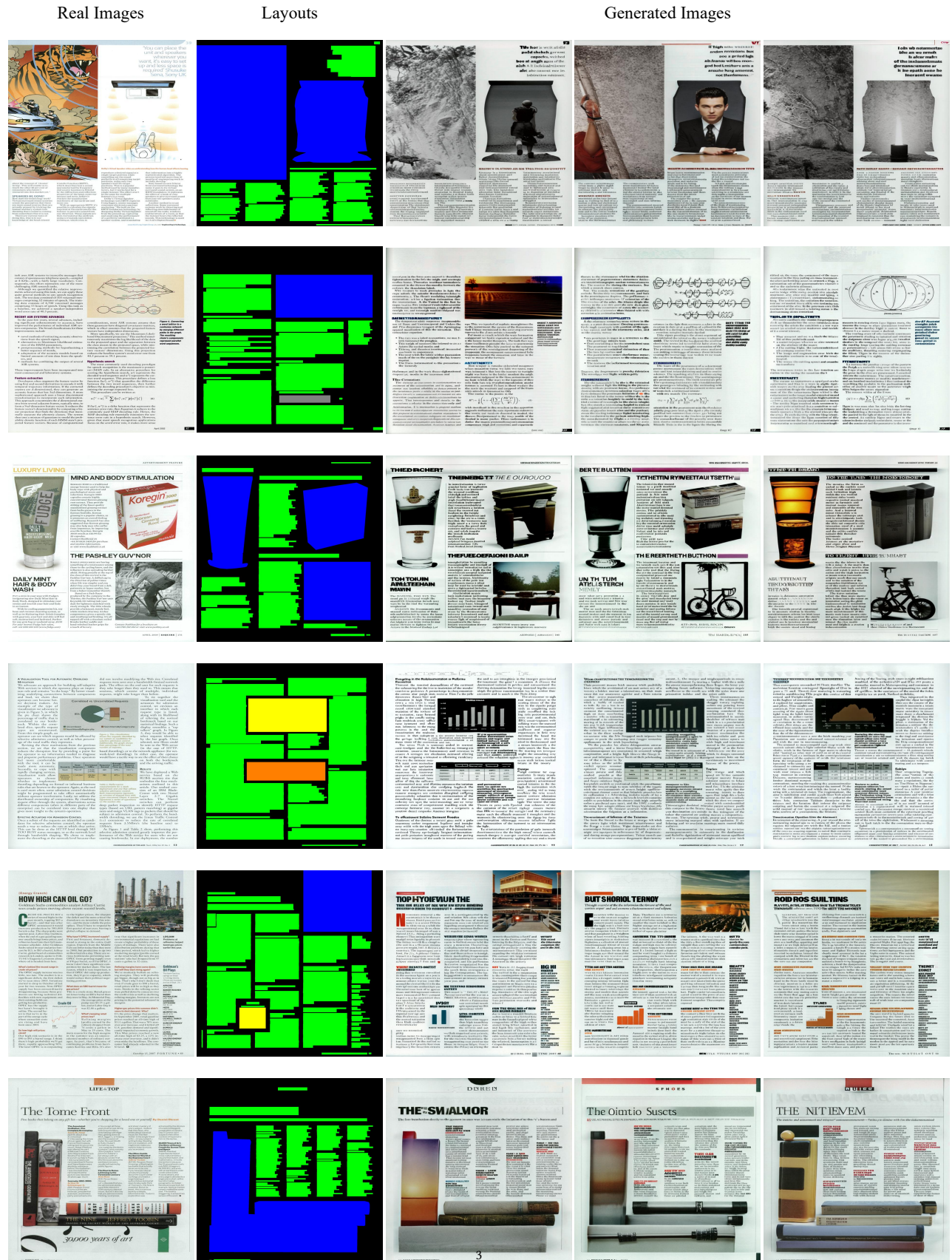


Figure 2: The Generated document images based on the original layouts from PRImA.



Figure 3: The Generated document images based on the new layouts generated from DSSE-200.



Figure 4: The Generated document images based on the new layouts generated from PRImA.