Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks

SUPPLEMENTAL MATERIAL

Xiao Yang[‡], Ersin Yumer[†], Paul Asente[†], Mike Kraley[†], Daniel Kifer[‡], C. Lee Giles[‡]
[‡]The Pennsylvania State University [†]Adobe Research
xuy111@psu.edu {yumer, asente, mkraley}@adobe.com dkifer@cse.psu.edu giles@ist.psu.edu

1. Synthetic Document Data

We introduced two methods to generate documents. In the first method, we generate LaTeX source files in which elements like paragraphs, figures, tables, captions, section headings and lists are randomly arranged using the "textblock" environment from the "textpos" package. Compiling these LaTeX files gives single, double, or triple-column PDFs. The generation process is summarized in Algorithm 1.

Algorithm 1 Synthetic Document Generation

```
$s \leftarrow$ a string containing preamble and necessary packages of a LaTeX source file
Select a LaTeX source file type $T \in \{$single-column, double-column, triple-column$\}$
while space remains on the page do
    Select an element type $E \in \{$figure, table, caption, section heading, list, paragraph$\}$
    Select an example $e$ of type $E$
    $s_e \leftarrow$ a string of LaTeX code that generates $e$ using the "textblock" environment
    $s \leftarrow s + s_e$
end while
Output: $s$
Output: A PDF document after compiling $s$
```

Elements in a document are carefully selected following the guidelines below. Figure 1 shows several examples of the figures and tables used in the synthetic data generation.

- Candidate figures include natural images from MS COCO [3], academic-style figures and graphic drawings downloaded using web image search.
- Candidates tables include table images downloaded using web image search. Various queries are used to increase the diversity of downloaded tables.
- For paragraphs, we randomly sample sentences from a 2016 English Wikipedia dump [2].
- For section headings, we sample sentences and phrases that are section or subsection headings in the "Contents" block in a Wikipedia page.
- For lists, we sample list items from Wikipedia pages, ensuring that all items in a list come from the same Wikipedia page.
- For captions, we either use the associated caption (for images from MS COCO) or the title of the image in web image search, which can be found in the span with class name "irc_pt".

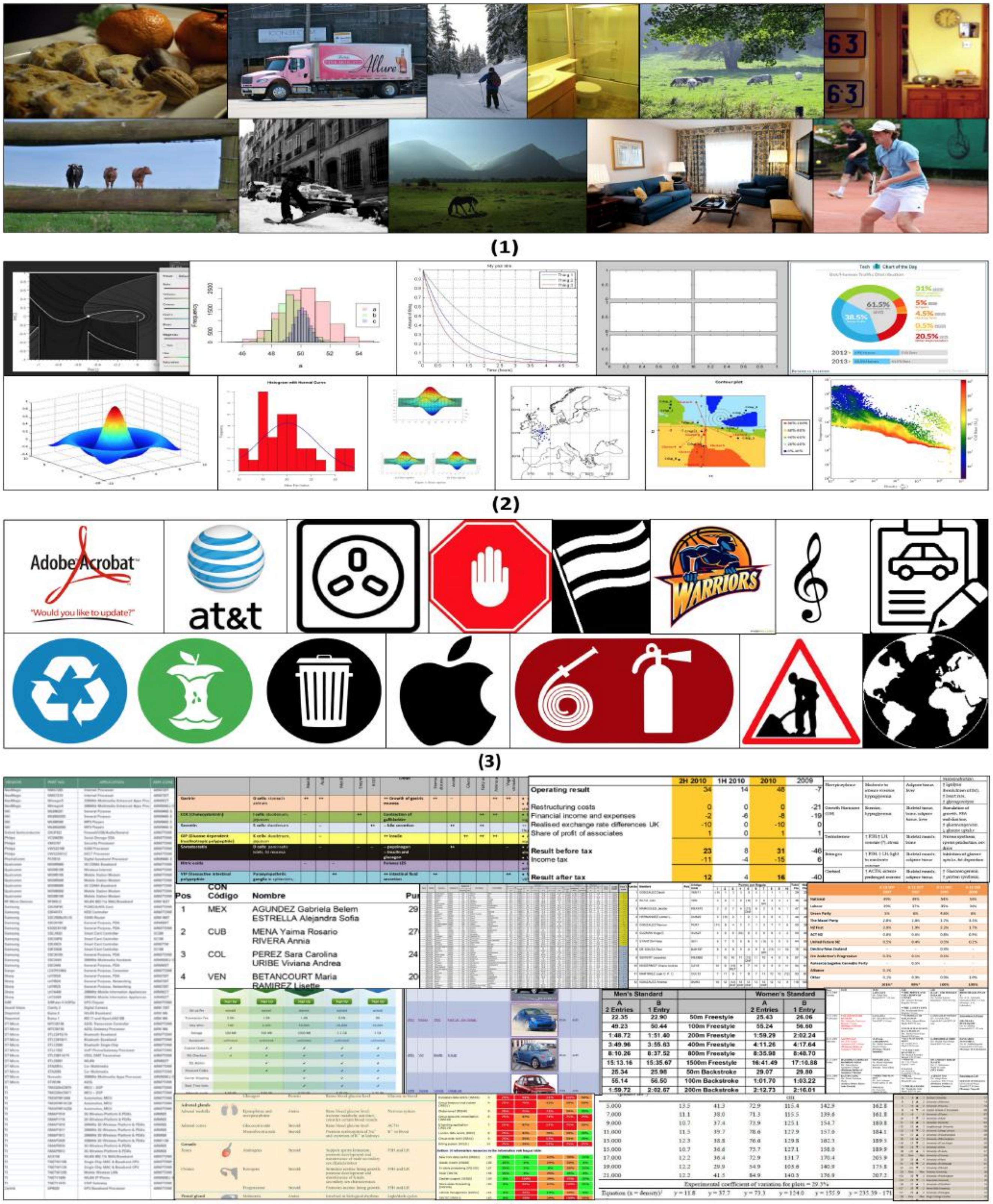
In the second document generation method, we collected and labeled 271 documents with varied, complicated layouts. We then randomly replaced each element with a standalone paragraph, figure, table, caption, section heading or list generated as stated above. Figure 2 shows several examples from the 271 documents.

2. Visualizing the Segmentation Results

Each pixel p in the model's output layer is assigned the color of the most likely class label l. The RGB value of that color is then weighted by the probability $\mathrm{P}(l)$.

3. Post-processing

We apply an optional post-processing step to clean up segment masks for documents in PDF format. First, we obtain candidate bounding boxes by using the auto-tagging capabilities of Adobe Acrobat [1] and parsing the results. Boxes are stored in a tree structure, and each node's box can be a TextRun (a sequence of characters), TextLine (potentially a text line), Paragraph (potentially a paragraph) or Container (potentially figures or tables). Note that we



	2H 2010	1H 2010	2010	2009
Operating result	34	14	48	-7
Restructuring costs	0	0	0	-21
Financial income and expenses	-2	-6	-8	-19
Realised exchange rate differences UK	-10	0	-10	0
Share of profit of associates	1	0	1	1
Result before tax	23	8	31	-46
Income tax	-11	-4	-15	6
Result after tax	12	4	16	-40

Pos	CON Código	Nombre	Pu
1	MEX	AGUNDEZ Gabriela Belem ESTRELLA Alejandra Sofia	29
2	CUB	MENA Yaima Rosario RIVERA Annia	27
3	COL	PEREZ Sara Carolina URIBE Viviana Andrea	24
4	VEN	BETANCOURT Maria RAMIREZ Lisette	20

Men's Standard			Women's Standard	
A 2 Entries	B 1 Entry		A 2 Entries	B 1 Entry
22.35	22.90	50m Freestyle	25.43	26.06
49.23	50.44	100m Freestyle	55.24	56.60
1:48.72	1:51.40	200m Freestyle	1:59.29	2:02.24
3:49.96	3:55.63	400m Freestyle	4:11.26	4:17.64
8:10.26	8:37.52	800m Freestyle	8:35.98	8:48.70
15:13.16	15:35.67	1500m Freestyle	16:41.49	17:10.88
25.34	25.98	50m Backstroke	29.07	29.80
56.14	56.50	100m Backstroke	1:01.70	1:03.22
1:59.72	2:02.67	200m Backstroke	2:12.73	2:16.01

[illegible]			cm			
5,000	13.5	41.3	72.9	115.4	142.9	162.8
7,000	11.1	38.0	71.3	115.5	139.6	161.8
9,000	10.7	37.4	73.9	125.1	154.7	180.8
11,000	11.5	39.7	78.6	127.9	157.0	184.1
13,000	12.3	38.8	76.6	129.8	162.3	189.5
15,000	10.7	36.6	73.7	127.1	158.0	189.9
17,000	12.2	36.4	72.9	131.7	170.4	205.9
19,000	12.2	29.9	54.9	105.6	140.9	175.8
21,000	12.2	41.5	84.9	140.3	176.9	207.2
Experimental coefficient of variation for plots = 29.3%						
Equation (x = density)[1]	y = 11.8	y = 37.7	y = 73.3	y = 124.0	y = 155.9	y = 235.39 - 171

Figure 1: Sample figures and tables used in synthetic documents generation. (1) Natural images from MS COCO dataset. (2) Academic-style figures from web image search. (3) Symbols and graphic drawings from web image search. (4) Tables from web image search.

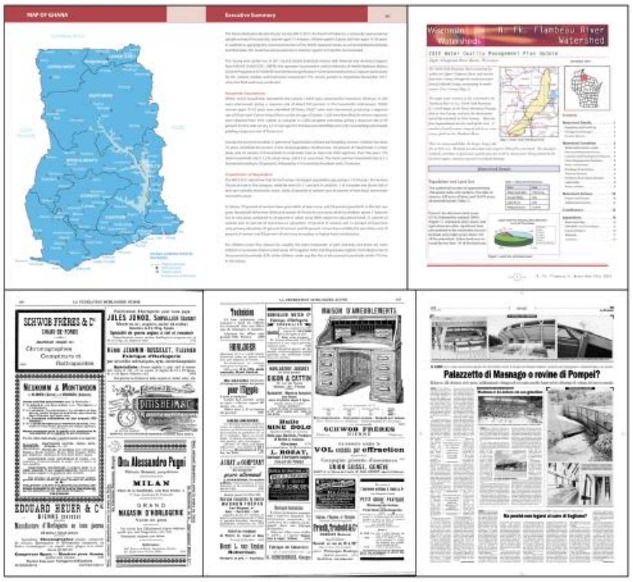


Figure 2: Examples of documents with complicated layout. We labeled regions in each document and then randomly replaced them with a standalone paragraph, figure, table, caption, section heading or list, as described in Sec. 1

ignore the semantic meanings associated with these boxes and only use the boxes as candidate bounding boxes in post-processing. Figure 3 (2) and 4 (2) illustrate candidate bounding boxes for each document.

Algorithm 2 Segmentation Post-processing

```
Input: $P \leftarrow$ probability map, $P(u,v) \in R^{|C|}$ is a vector
    containing the probability of each class $c \in C$ at location $(u,v)$
Input: $Boxes \leftarrow$ candidate bounding boxes
  $S \leftarrow$ segmentation to be generated
  for each location $(x,y) \in S$ do
      $S(x,y) \leftarrow$ background
  end for
  for each $b \in Boxes$ do          ▷ parent box comes before child boxes
      $\bar{p} \leftarrow \sum_{(u,v)\in b} P(u,v)$
      $l \leftarrow$ argmax $\bar{p}$
      for each location $(u,v) \in b$ do
          if S(u, v) is background then
              $S(u,v) \leftarrow l$
          end if
      end for
  end for
Output: $S$
```

Using these bounding box candidates, we refine the segmentation masks by first calculating the average class probability for pixels belonging to the same box, followed by assigning the most likely label to these pixels. The process is summarized in Algorithm 2.

4. Additional Visualization Results

Figures 3 and 4 show additional visualization examples of synthetic documents, and Figure 5 shows additional examples of real documents.

References

[1] Adobe Acrobat. http://www.adobe.com/accessibity/products/acrobat.html. 1

[2] Wikipedia. https://dumps.wikimedia.org/. 1

[3] 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. Springer, 2014. 1

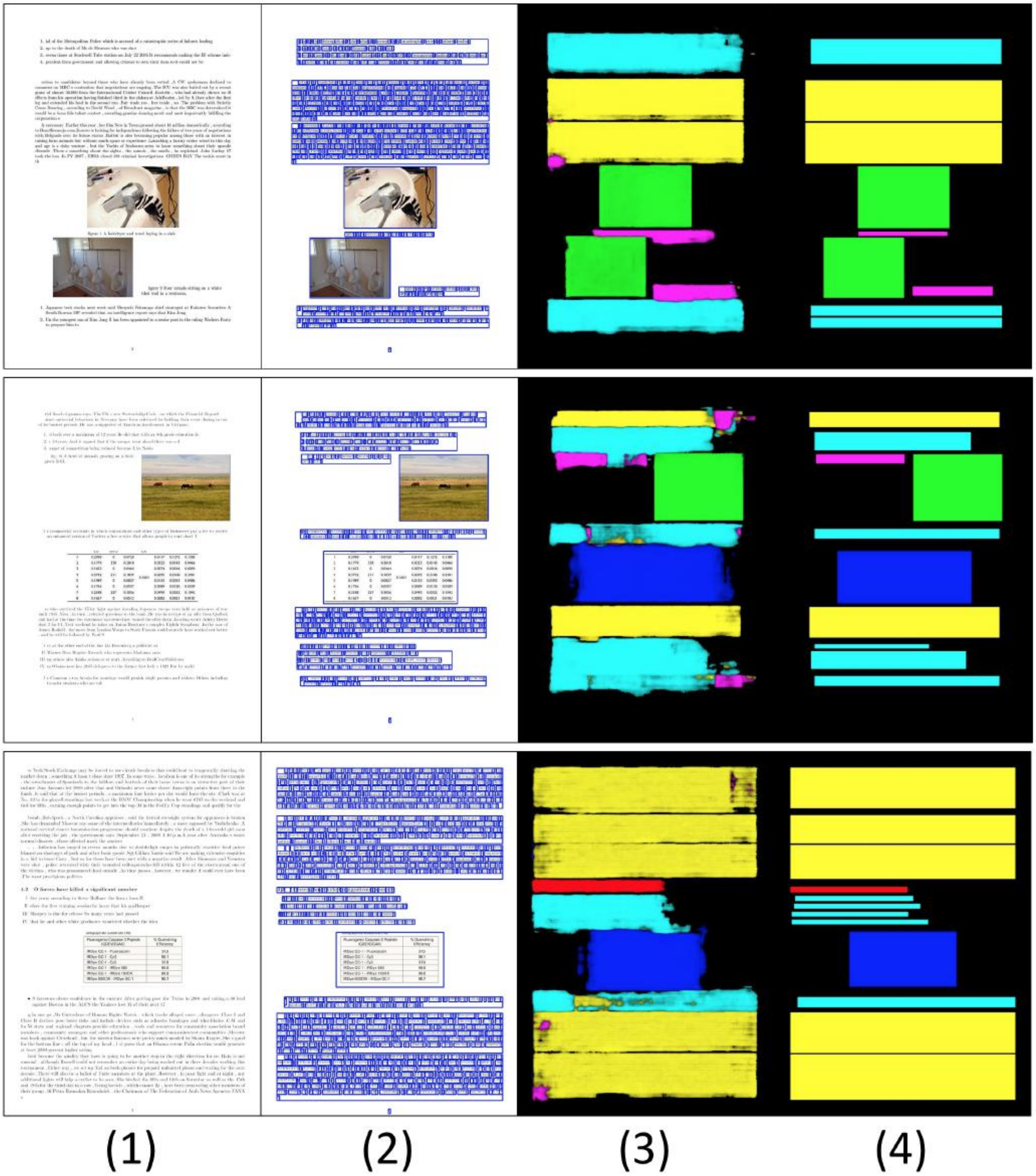


Figure 3: Synthetic documents and the corresponding segmentations. (1) Input synthetic documents. (2) Candidate bounding boxes obtained by parsing the PDF rendering commands. (3) Raw segmentation outputs. (4) Segmentations after post-processing. Segmentation label colors are: **figure** , **table** , **section heading** , **caption** , **list** and **paragraph** .

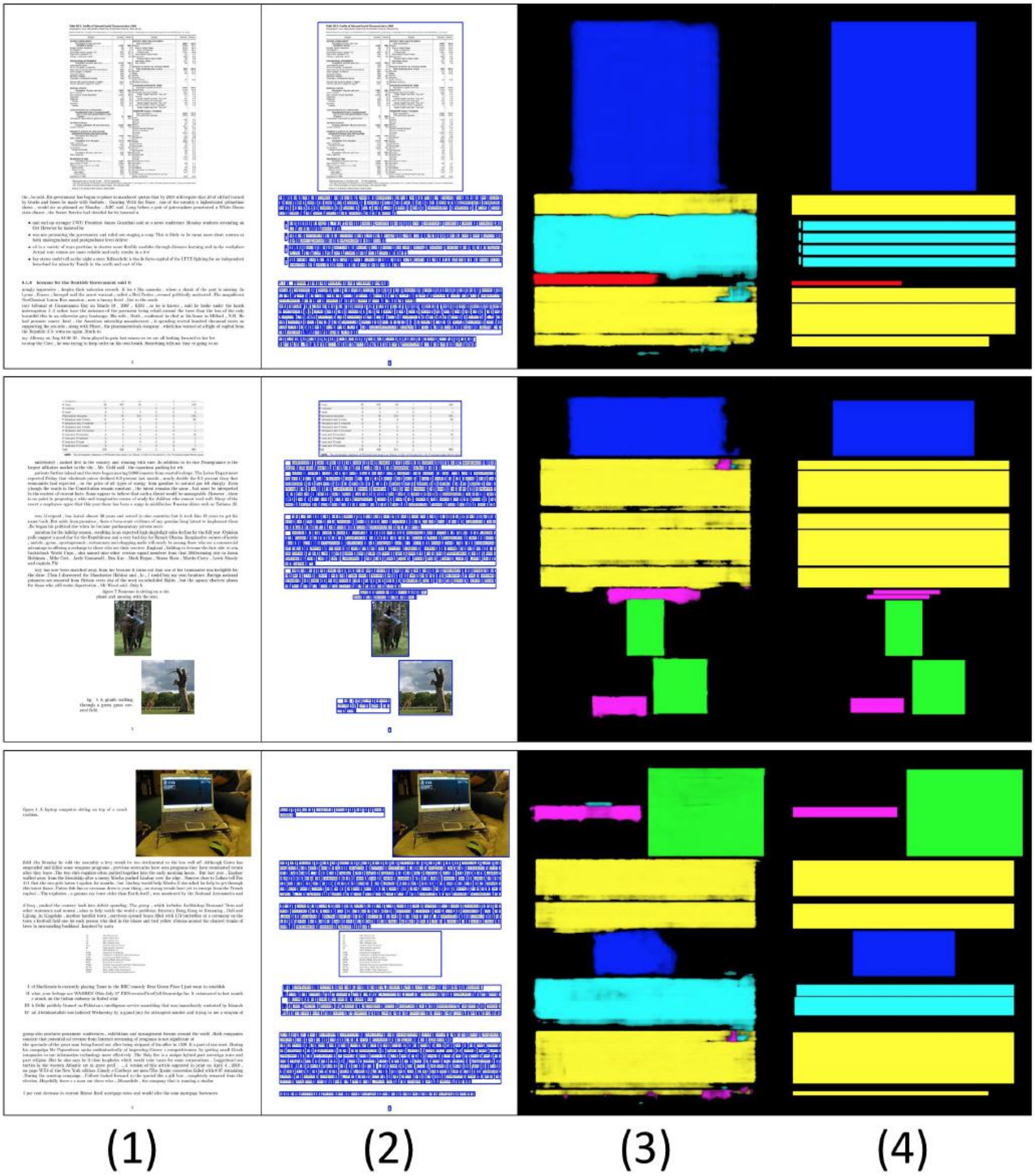


Figure 4: Synthetic documents and the corresponding segmentations. (1) Input synthetic documents. (2) Candidate bounding boxes obtained by parsing the PDF rendering commands. (3) Raw segmentation outputs. (4) Segmentations after post-processing. Segmentation label colors are: **figure** , **table** , **section heading** , **caption** , **list** and **paragraph** .

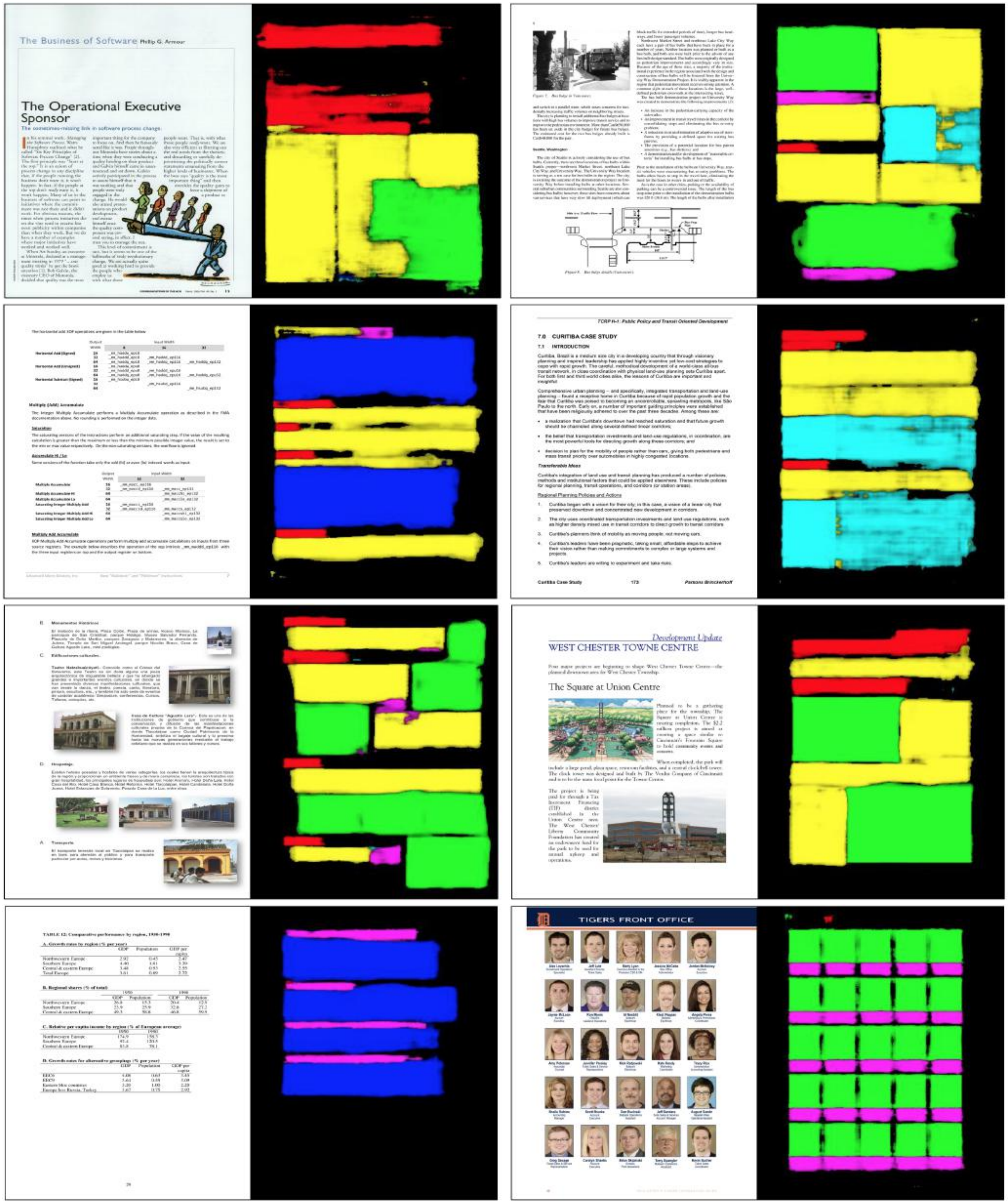

The Business of Software
The Operational Executive Sponsor
7.0 CURITIBA CASE STUDY
Development Update
WEST CHESTER TOWNE CENTRE
The Square at Union Centre
TIGERS FRONT OFFICE

Figure 5: Real documents and the corresponding segmentations. Segmentation label colors are: **figure**, **table**, **section heading**, **caption**, **list** and **paragraph**.