



SelfDocSeg: A Self-Supervised vision-based Approach towards Document Segmentation

Subhajit Maity*¹ **Sanket Biswas***² Siladitya Manna³ Ayan Banerjee²
Josep Lladós² Saumik Bhattacharya⁴ Umapada Pal³

¹Technology Innovation Hub, Indian Statistical Institute, Kolkata, India

²Computer Vision Center, Universitat Autònoma de Barcelona, Spain

³CVPR Unit, Indian Statistical Institute, Kolkata, India

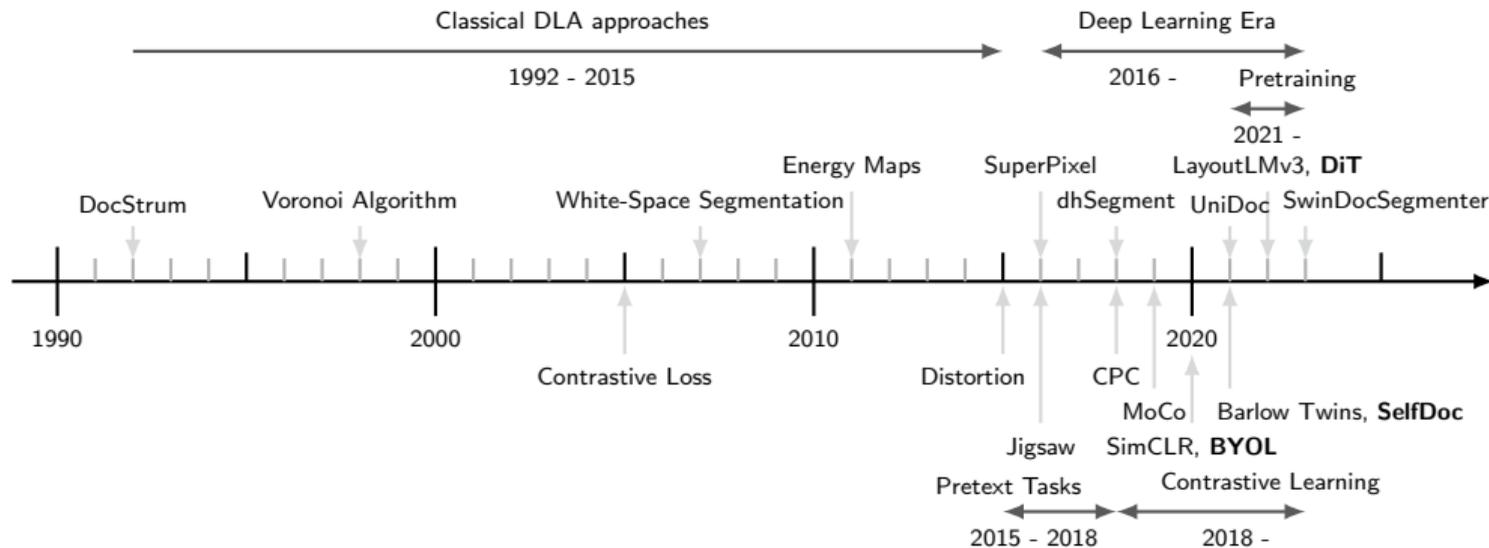
⁴Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India



- 1 Background
- 2 SelfDocSeg
- 3 Experiments
- 4 Conclusion

Going Back in Time

Document Layout Analysis (DLA)



Self-Supervised Learning

Where DLA and Self-supervised Learning stand today?

- The classical image processing techniques for DLA had long been lost.
- Modern self-supervision is bringing the classical algorithms, like Felzenszwalb, Normalized Cuts *etc.* back for guidance.
- The current document pre-training techniques for DLA use *multi-modal* approaches and *large-scaled* datasets.

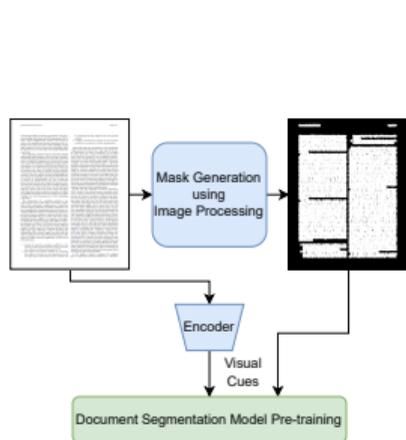
What do we need?

- A strategy to use the document images without annotations for DLA.
- A *data-efficient* pre-training strategy with *unlabelled* document images to alleviate the usual requirement of computational resources.

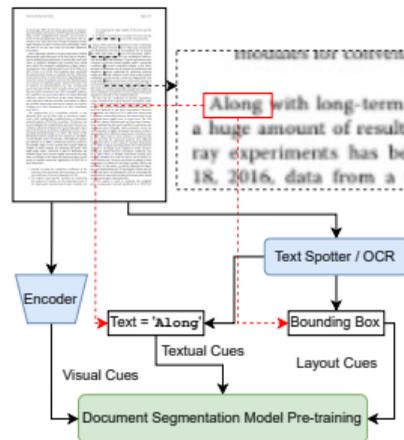
Motivation behind SelfDocSeg

What's different from prior state-of-the-arts (SOTAs)?

- No textual cue or layout cue from trained OCR. Only visual cues are enough to capture global and local context.
- Superior data efficiency



(a) SelfDocSeg



(b) Existing Self-supervised SOTAs

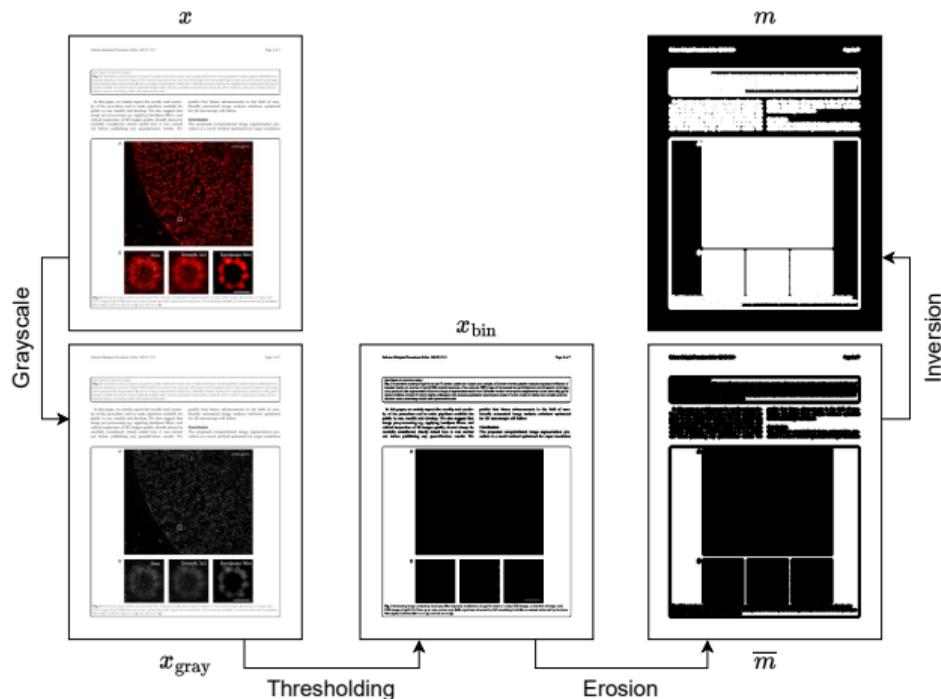
- A novel *vision-based self-supervised* framework, specifically designed to pre-train an image encoder for DLA task.
- A *pseudo physical layout guided strategy* for self-supervision in the region of interest localization for document segmentation.
- A *data-efficient pre-training strategy* to learn multiple document object representations simultaneously in the self-supervised setting

Overall Algorithmic Overview of SelfDocSeg

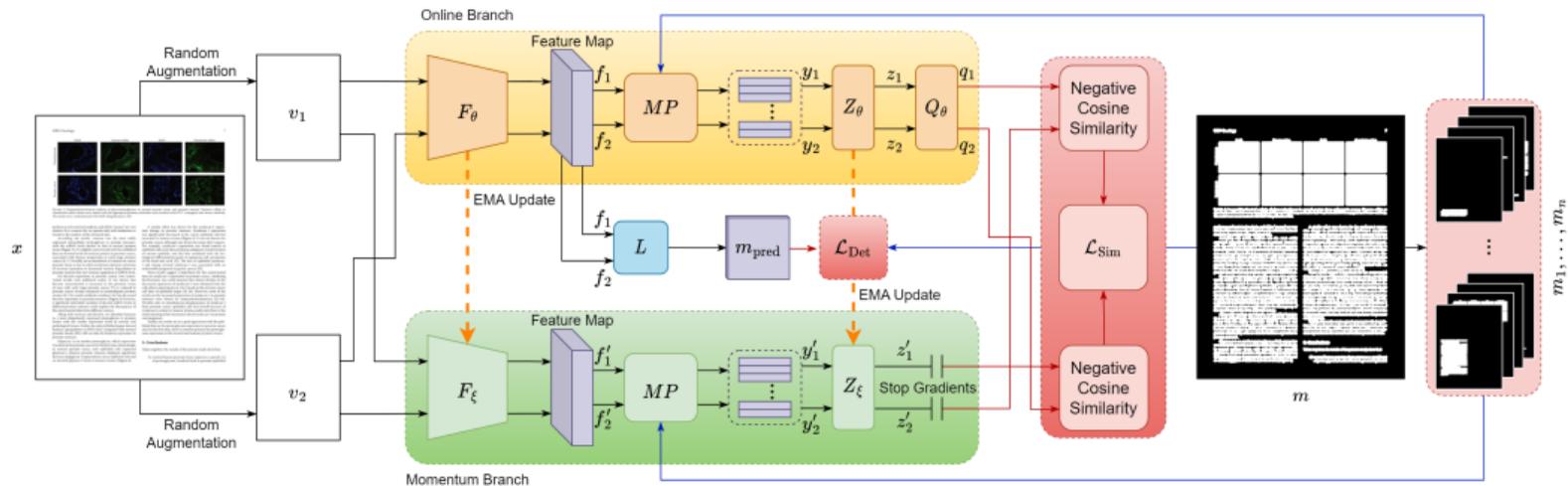
- 1 First, an approximate layout mask m is generated.
- 2 The image encoder is pre-trained with the help of m as visual cues in a non-weight-shared two-branch network similar to BYOL [28].
- 3 The encoder pre-training is done with two objectives - (a) *localization* and (b) *representation* of layout objects.
- 4 Localization objective is optimized using a detection loss \mathcal{L}_{Det} .
- 5 Layout object representations are extracted by *mask pooling* operation on encoded feature-maps for each object.
- 6 Representation objective is learnt via a similarity loss \mathcal{L}_{Sim} .

Mask Generation

- 1 Grayscale Conversion:**
CIE grayscale conversion on the document image x
- 2 Thresholding:**
Global thresholding with value 239 for 8-bit integer pixels
- 3 Erosion:**
 5×5 rectangular kernel
- 4 Inversion:**
pixel values subtracted from 255 to get the layout mask m



Architecture



Architecture Details

Online Branch

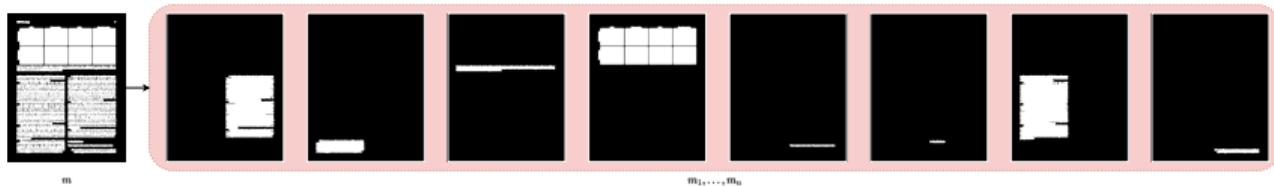
The online branch has an encoder F_θ , a projector Z_θ and a predictor Q_θ . It is updated using backpropagation.

Momentum Branch

The momentum branch has an encoder F_ξ and a projector Z_ξ . It is updated using an exponentially moving average (EMA).

Layout Objects for Mask Pooling

The layout object masks m_1, \dots, m_n are separated from mask m as separate contours.



Mask Pooling

Mask pooling [35] is used for extracting layout object representation from encoded feature maps. It is just an average pooling inside each contour masked by m .

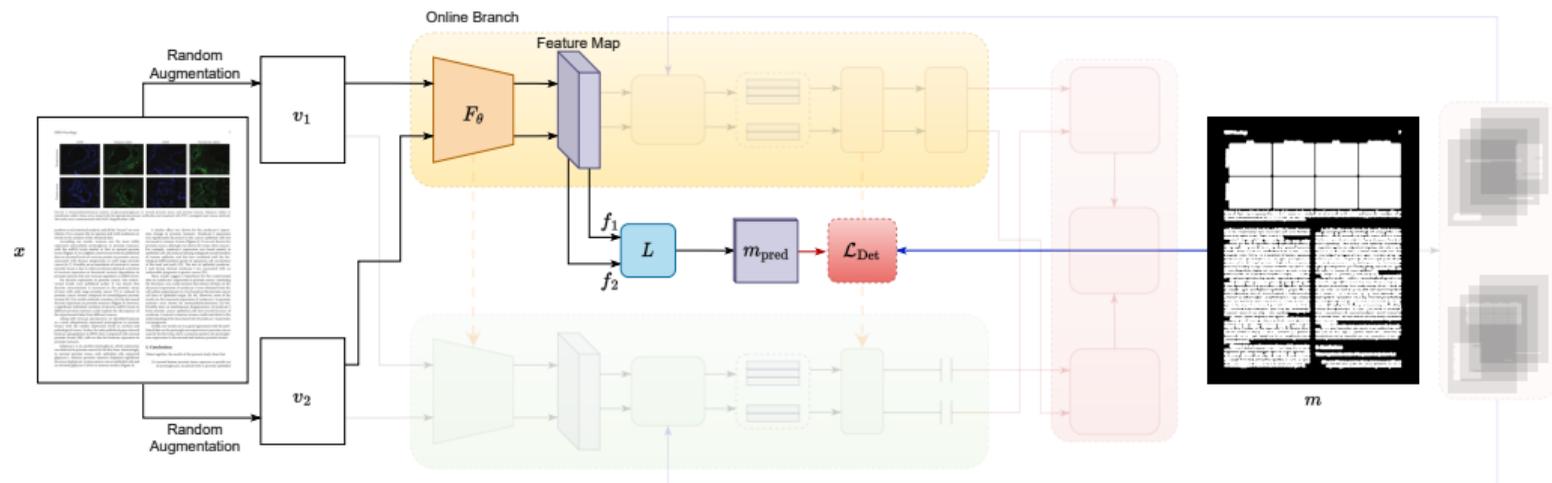
$$y^{(k)} = \frac{1}{\sum_{i,j} m_k[i,j]} \sum_{i,j} m_k[i,j] f[i,j]$$

Weight Updates

Online Branch: $\theta \leftarrow \text{optimizer}(\theta, \nabla_{\theta} \mathcal{L}_{\text{total}}, \eta)$

Momentum Branch: $\xi \leftarrow \tau \xi + (1 - \tau) \theta$

Learning to Localize Layout Objects



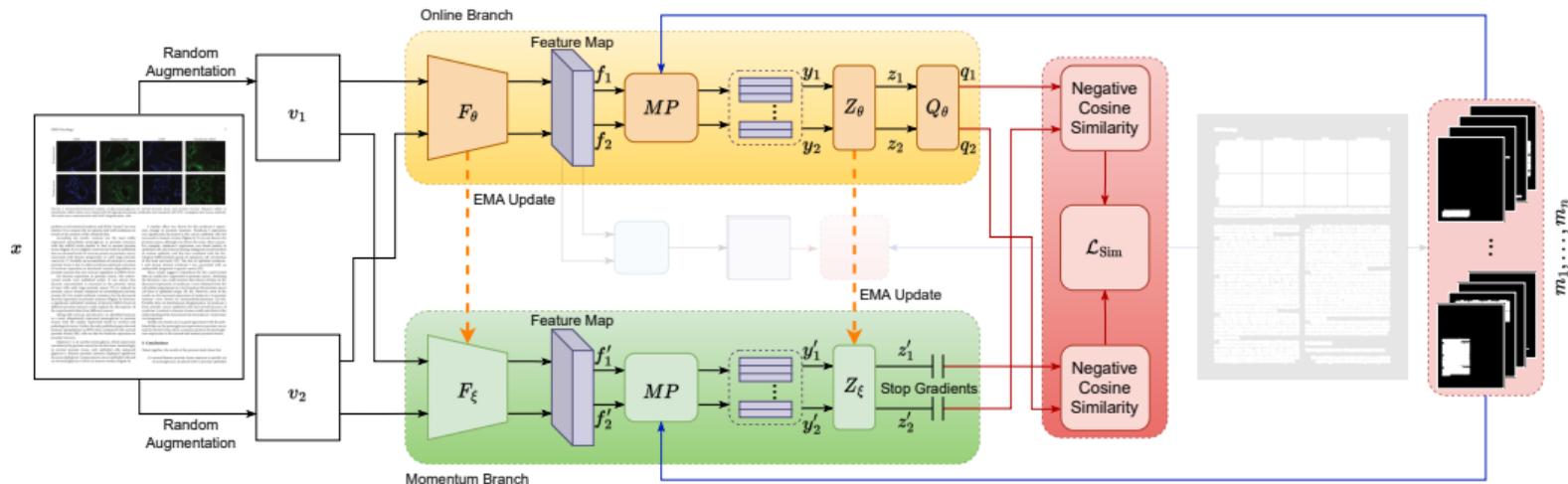
Localization Pre-training Objective

- The mask predictor module L is used to predict the layout mask m_{pred} .
- m_{pred} translates to the pixel-level probability of a layout object being present.
- We formulate this prediction objective as an imbalanced classification task and use Focal Loss [43] \mathcal{L}_{Det} for the same.

Detection Loss

$$\mathcal{L}_{\text{Det}} = -\frac{\alpha}{\sum_{i,j} m[i,j]} \cdot \sum_{i,j} (m[i,j](1 - m_{\text{pred}}[i,j])^\gamma \log m_{\text{pred}}[i,j] \\ + (1 - m[i,j])m_{\text{pred}}[i,j]^\gamma \log(1 - m_{\text{pred}}[i,j]))$$

Learning to Recognize Layout Objects



Representation Pre-training Objective

- Representation of each layout object is extracted from feature maps of encoders F_θ and F_ξ in the online and the momentum branches using masked pooling.
- The encoder is trained using cosine similarity loss [28] \mathcal{L}_{Sim} between online predictor Q_θ and momentum projector Z_ξ .

Similarity Loss

$$\mathcal{L}_{\text{Sim}} = 4 - 2 \left(\frac{\langle q_1, z'_2 \rangle}{\|q_1\|_2 \cdot \|z'_2\|_2} + \frac{\langle q_2, z'_1 \rangle}{\|q_2\|_2 \cdot \|z'_1\|_2} \right)$$

Pre-training Dataset:

- DocLayNet [51]

Fine-tuning Dataset:

- DocLayNet [51]
- PubLayNet [63]
- PRImA [2]
- Historic Japanese [54]

Our Backbone:

- ResNet 50 [34]

Competitors:

- DocSegTr [8]
- LayoutParser [55]
- Biswas *et al.* [9]
- Mask RCNN [33] (Fully supervised)
- LayoutLMv3_{Base} [36]
- UDoc [29]
- DiT_{Base} [40]

Our Detector:

- Mask RCNN [33]

Frameworks: PyTorch, OpenCV, PyTorch Lightning, Lightly, Detectron2

Source code: <https://github.com/MaitySubhajit/SelfDocSeg>

Results - Quantitative

Methods	Cues			# Data	DocLayNet mAP	PubLayNet mAP	PRImA mAP	HJ mAP
	V	L	T					
DocSegTr [8]	✓	✗	✗	-	-	90.4	42.5	83.1
LayoutParser [55]	✓	✓	✓	-	-	86.7	64.7	81.6
Biswas <i>et al.</i> [9]	✓	✗	✗	-	-	89.3	56.2	82.0
Mask RCNN [33]	✓	✗	✗	-	72.4	88.6	56.3	80.1
LayoutLMv3 _{Base} [36]	✓	✓	✓	11M	-	95.1	40.3	82.7
UDoc[29]	✓	✓	✓	1M	-	93.9	-	-
DiT _{Base} [40]	✓	✗	✗	42M	-	93.5	-	-
BYOL [28]	✓	✗	✗	81k	63.5	79.0	28.7	59.8
SelfDocSeg	✓	✗	✗	81k	74.3	89.2	52.1	78.8

Results - Visual Comparison with Ground-truth

3

CONFIDENTIAL INFORMATION
Subject to Confidentiality and/or Proprietary Rights

4.28 SUBSEQUENT EVENTS

5. FINANCIAL RISK MANAGEMENT OBJECTIVES AND POLICY

5.1 EXCHANGE AND INTEREST RATE RISK

5.1.1 Analysis of sensitivity to interest rates

	Note	August 31, 2013	August 31, 2012
Financial liabilities excluding derivative financial instruments	4.15	2,607	2,686
Fixed rate liabilities		2,324	2,403
Variable rate liabilities		283	283
Impact of interest rate and cross-currency swaps	4.16	(67)	(2)
On fixed rate liabilities		177	216
On variable rate liabilities		(244)	(218)
Financial liabilities after impact of interest rate and cross-currency swaps		2,540	2,684
Fixed rate liabilities		2,361	2,619
Variable rate liabilities		179	65

4.29 If August 31, 2013 and 2012, a 0.5% increase or decrease in interest rates would have had no material impact on net income before tax or on shareholders' equity as a substantial amount of those notes are at a fixed rate of interest.

5.1.1 If August 31, 2013 and 2012, a 0.5% increase or decrease in interest rates would have had no material impact on net income before tax or on shareholders' equity as a substantial amount of those notes are at a fixed rate of interest.

Source - Registration Document Fiscal 2013 - 2014

Left: Prediction

3

CONFIDENTIAL INFORMATION
Subject to Confidentiality and/or Proprietary Rights

4.28 SUBSEQUENT EVENTS

5. FINANCIAL RISK MANAGEMENT OBJECTIVES AND POLICY

5.1 EXCHANGE AND INTEREST RATE RISK

5.1.1 Analysis of sensitivity to interest rates

	Note	August 31, 2013	August 31, 2012
Financial liabilities excluding derivative financial instruments	4.15	2,607	2,686
Fixed rate liabilities		2,324	2,403
Variable rate liabilities		283	283
Impact of interest rate and cross-currency swaps	4.16	(67)	(2)
On fixed rate liabilities		177	216
On variable rate liabilities		(244)	(218)
Financial liabilities after impact of interest rate and cross-currency swaps		2,540	2,684
Fixed rate liabilities		2,361	2,619
Variable rate liabilities		179	65

4.29 If August 31, 2013 and 2012, a 0.5% increase or decrease in interest rates would have had no material impact on net income before tax or on shareholders' equity as a substantial amount of those notes are at a fixed rate of interest.

5.1.1 If August 31, 2013 and 2012, a 0.5% increase or decrease in interest rates would have had no material impact on net income before tax or on shareholders' equity as a substantial amount of those notes are at a fixed rate of interest.

Source - Registration Document Fiscal 2013 - 2014

Right: Ground-truth

Results - Visual Comparison with Ground-truth



Left: Prediction

Right: Ground-truth

Results - Visual Comparison with Ground-truth

Table 98A

2003	Software Rm	Patents and trademarks Rm	Capitalised exploration costs Rm	Capital work in progress Rm	Other intangible assets Rm	Total Rm
Cost						
Balance at 30 June 2002	736	543	888	397	13	2,577
Acquisition of businesses (refer note 4E)	—	2	—	—	—	2
Additions	293	95	63	242	3	696
to enhance existing operations	289	5	—	82	1	377
to expand operations	4	90	63	160	2	319
Removing costs capitalised	8	—	—	—	—	8
Transfer (to)/from property, plant and equipment	70	107	—	(211)	—	(34)
Discontinuation of foreign entities (refer note 3E)	(23)	(95)	(93)	(4)	—	(195)
Disposals and scrapping	(1)	(4)	(235)	—	—	(244)
Balance at 30 June 2003	1,083	644	633	424	16	2,820
Amortisation and amounts written off						
Balance at 30 June 2002	142	372	278	—	—	792
Acquisition of businesses (refer note 4E)	—	2	—	—	—	2
Current year charge	276	60	31	—	3	370
Impairment of assets (refer note 34) (refer note 3E)	(20)	(5)	(13)	—	—	(38)
Transfer from property, plant and equipment	24	—	—	—	—	24
Disposals and scrapping	(1)	(1)	(200)	—	—	(202)
Balance at 30 June 2003	361	376	29	—	3	769
Carrying value at 30 June 2003	722	268	624	424	13	2,051
Carrying value at 30 June 2002	594	171	679	397	13	1,864
Cost price of assets not yet in use	29	—	—	404	—	433
Cost of fully amortised assets in use	74	43	—	—	—	117

Notes: Expenses on intangible assets amongst the operations of a purchaser comprise research, training, salaries of staff in the group as well as ongoing exploration and development activities in the upstream oil and gas industry, mainly in Azerbaijan.

Table 98B

	2003 Rm	2002 Rm	2001 Rm
Estimated future aggregate amortisation expense per annum			
Within one year	436	235	167
One to two years	388	326	54
Two to three years	264	295	72
Three to four years	70	112	67
Four to five years	80	77	34
Over five years	871	609	685
	2,051	1,804	1,014

Left: Prediction

Table

2003	Software Rm	Patents and trademarks Rm	Capitalised exploration costs Rm	Capital work in progress Rm	Other intangible assets Rm	Total Rm
Cost						
Balance at 30 June 2002	736	543	888	397	13	2,577
Acquisition of businesses (refer note 4E)	—	2	—	—	—	2
Additions	293	95	63	242	3	696
to enhance existing operations	289	5	—	82	1	377
to expand operations	4	90	63	160	2	319
Removing costs capitalised	8	—	—	—	—	8
Transfer (to)/from property, plant and equipment	70	107	—	(211)	—	(34)
Discontinuation of foreign entities (refer note 3E)	(23)	(95)	(93)	(4)	—	(195)
Disposals and scrapping	(1)	(4)	(235)	—	—	(244)
Balance at 30 June 2003	1,083	644	633	424	16	2,820
Amortisation and amounts written off						
Balance at 30 June 2002	142	372	278	—	—	792
Acquisition of businesses (refer note 4E)	—	2	—	—	—	2
Current year charge	276	60	31	—	3	370
Impairment of assets (refer note 34) (refer note 3E)	(20)	(5)	(13)	—	—	(38)
Transfer from property, plant and equipment	24	—	—	—	—	24
Disposals and scrapping	(1)	(1)	(200)	—	—	(202)
Balance at 30 June 2003	361	376	29	—	3	769
Carrying value at 30 June 2003	722	268	624	424	13	2,051
Carrying value at 30 June 2002	594	171	679	397	13	1,864
Cost price of assets not yet in use	29	—	—	404	—	433
Cost of fully amortised assets in use	74	43	—	—	—	117

Notes: Expenses on intangible assets amongst the operations of a purchaser comprise research, training, salaries of staff in the group as well as ongoing exploration and development activities in the upstream oil and gas industry, mainly in Azerbaijan.

Table

	2003 Rm	2002 Rm	2001 Rm
Estimated future aggregate amortisation expense per annum			
Within one year	436	235	167
One to two years	388	326	54
Two to three years	264	295	72
Three to four years	70	112	67
Four to five years	80	77	34
Over five years	871	609	685
	2,051	1,804	1,014

Right: Ground-truth

Generalizability & Effectiveness of Pre-training

Experimented with fewer annotations in a semi-supervised setting for fine-tuning.

% Annotations	mAP
10%	41.3
50%	65.1
100%	74.3

Contribution of Detection & Representation Objectives

Experimentation with \mathcal{L}_{Det} and \mathcal{L}_{Sim} in strip-down style.

Loss	mAP
w/o \mathcal{L}_{Sim}	39.1
w/o \mathcal{L}_{Det}	69.7
Combined ($\mathcal{L}_{\text{total}}$)	74.3

- **SelfDocSeg** being a *OCR-free pertaining strategy* is computationally less expensive, as large-scale OCR systems are difficult and time-consuming to train.
- Ideally **SelfDocSeg** provides a *wide range of flexibility with backbones and detectors* and can achieve better performance with a stronger backbone.
- **SelfDocSeg** is *data-efficient* compared to state-of-the-art pre-training methods for DLA and can *generalize quickly over different domains* in the fine-tuning stage.

What's Next?

- Although being a superior pre-training strategy in terms of data efficiency, **SelfDocSeg** is not yet tested for data-hungry models like LayoutLMv3 or DiT.
- Following the recent trends to promote performance improvement, the scope of exploration remains in multi-domain and multi-lingual pre-training for introducing data variety.
- **SelfDocSeg** is particularly designed to cater to DLA. Thus, pre-training and fine-tuning for multi-task settings in downstream with vision-based self-supervision can be explored.

- **SelfDocSeg** is *self-supervised* and *vision based* and is motivated by recent self-supervised works in computer vision.
- Self-supervision for documents is possible in a unimodal setting without dependency on OCR or textual knowledge.
- The complete visual representation approach of **SelfDocSeg** facilitates superior understanding and encoding of the visual modality in the layout-guided document understanding paradigm.



arXiv: <https://arxiv.org/abs/2305.00795>

Code: <https://github.com/MaitySubhajit/SelfDocSeg>

Project Page: <https://maitysubhajit.github.io/SelfDocSeg>