A APPENDIX

A.0.1 Implementation details. The first stage of *DocLangID* is trained for 100 epochs and the second stage is trained for an additional 50 epochs. We use a batch size of 16 and a learning rate of 0.001 for the first stage and 0.3 for the second. For the optimization of the network parameters we use Adam optimizer. We apply random augmentations on the training images, including rotations up to 45 degrees, brightness, sharpness and contrast manipulations, and noise added to the image. Also, we apply center-cropping on all images with a target size of $(1024, 1024)$ and use constant aspect ratios of $(256, 256)$ for a patch, which results in a maximum of 16 possible patches. We implemented the deep-learning algorithms using the PyTorch-Lightning framework [3] and conducted all the experiments with a NVIDIA RTX3090 GPU.

Table A.1: Average inference times in seconds for each language class, with and without input preprocessing time.

Dataset	Language	Preprocessing Time		No Preprocessing Time	
		Tesseract	DocLangID	Tesseract	DocLangID
IMPACT	Dutch	14.4	1.40	13.6	0.89
	Polish	13.1	1.44	12.5	0.96
	Slovenian	36.9	4.52	40.3	2.37
	Spanish	12.5	1.37	12.2	0.85
	Czech	14.4	1.34	9.9	0.89
	Bulgarian	36.2	1.76	31.1	1.13
Average		**21.2**	**2.95**	**19.9**	**1.18**
WPI	Dutch	7.3	0.46	7.1	0.003
	German	9.7	0.83	7.2	0.004
	French	6.2	0.51	6.3	0.003
	English	7.1	0.74	5.5	0.003
Average		**7.5**	**0.63**	**6.5**	**0.003**

Table A.2: DocLangID classification metric results for each WPI language (50 few-shot samples)

Language	Precision	Recall	F1
Dutch	0.88	0.53	0.66
German	0.72	0.87	0.79
French	0.84	0.66	0.74
English	0.65	0.91	0.76
Avg.	**0.77**	**0.74**	**0.74**

Table A.3: Classification accuracies depending on the amount of few-shot samples used.

Few-shot samples	ResNet-FewShot	ResNet-Meta	DocLangID (Ours)
5	0.42	0.57	**0.63**
10	0.51	0.61	**0.63**
20	0.55	0.64	**0.66**
50	0.61	0.67	**0.72**

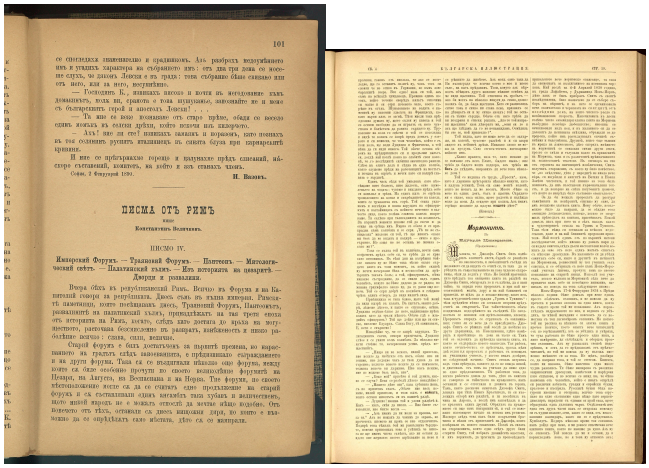

Figure 6: Example images from the IMPACT [14] dataset belonging to the Bulgarian language.

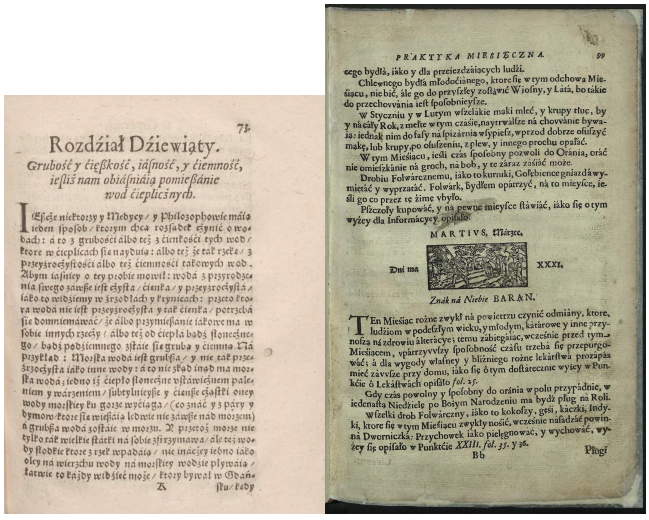

Figure 7: Example images from the IMPACT [14] dataset belonging to the Polish language.

Table A.4: Influence of different numbers of patches on the recognition performance (accuracy scores using 50 few-shot samples)

Patches	ResNet-FewShot	ResNet-Meta	DocLangID (Ours)
2	0.33	0.42	**0.63**
4	0.47	0.59	**0.69**
6	0.44	0.66	**0.70**
8	0.43	0.58	**0.70**
10	0.49	0.65	**0.73**
16 (max)	0.61	0.67	**0.72**

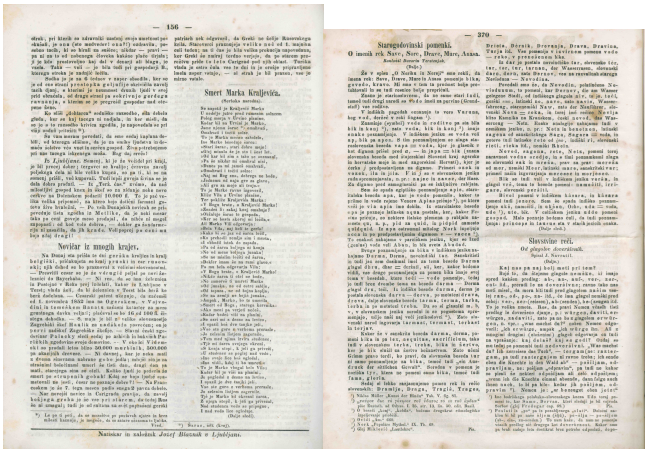

Figure 8: Example images from the IMPACT [14] dataset belonging to the Slovenian language.

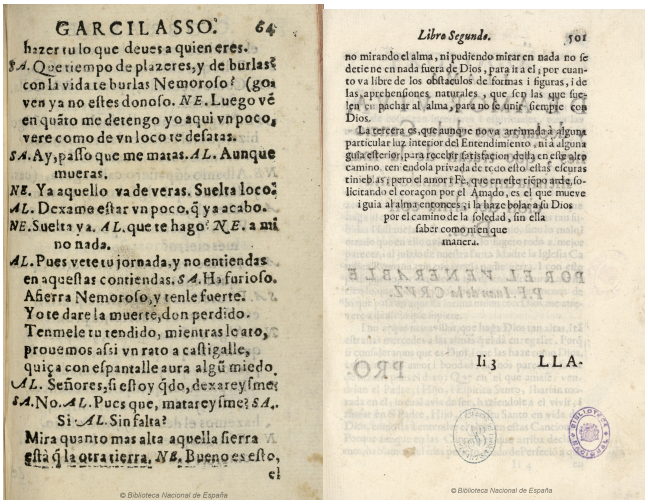

Figure 9: Example images from the IMPACT [14] dataset belonging to the Spanish language.

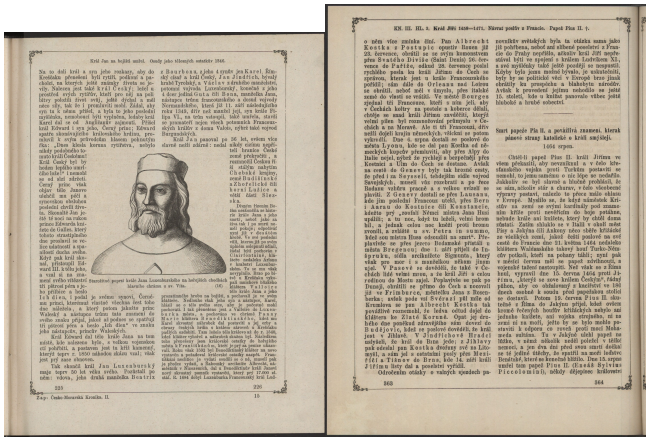

Figure 10: Example images from the IMPACT [14] dataset belonging to the Czech language.

Figure 11: Example images from the IMPACT [14] dataset belonging to the Dutch language.

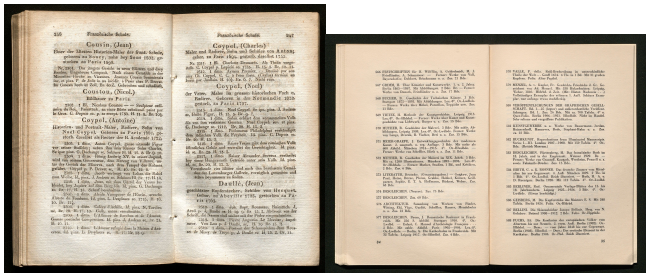

Figure 12: Example images from the WPI dataset belonging to the German language.

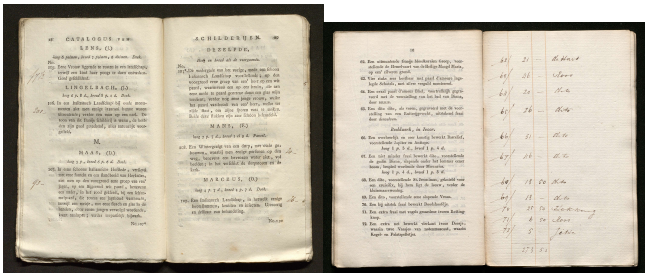

Figure 13: Example images from the WPI dataset belonging to the Dutch language.

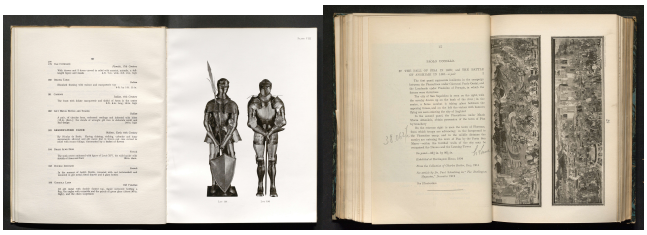

Figure 14: Example images from the WPI dataset belonging to the English language.

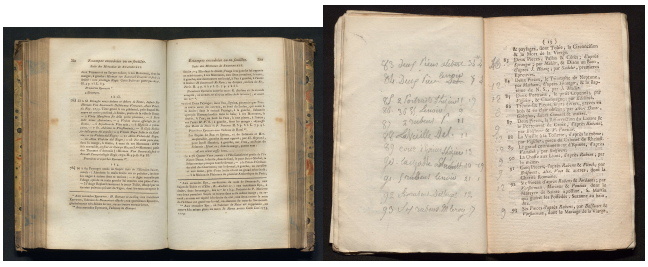

Figure 15: Example images from the WPI dataset belonging to the French language.