Supplementary Material for: Handwritten and Printed Text Segmentation: A Signature Case Study [1]

Sina Gholamian
Thomson Reuters AI Labs
Toronto, Canada
sina.gholamian@thomsonreuters.com

Ali Vahdat
Thomson Reuters AI Labs
Toronto, Canada
ali.vahdat@thomsonreuters.com

Abstract

In this supplementary document, we provide further background and additional details on the experiments, architecture, and results for our work on handwritten and printed text segmentation.

1. Additional Background and Related Work

1.1. Scope

Converting hard-copy documents, including microfilms for archival and historical documents [2], into digital format has been the focus of the computer vision research community, and several approaches have been proposed for this purpose [3, 4]. For a variety of reasons, such as ease of access and document understanding, paper documents in different domains, such as books, archival documents, medical records, legal documents, survey forms, and more, are converted to their digital format through optical character recognition (OCR). Due to the high demand for document digitization, several commercial and open-source OCR engines, such as Amazon's Textract [5], Tesseract [6], and ABBYY FineReader [7] have been available for some time. These tools significantly speed up the automatic digitization of documents at scale. However, given the diversity of document types, their structures, overlapping handwritten and printed text, and also the poor quality of original documents and their scanned version in the case of historical documents, the quality of OCR tools can degrade in the digitization process [2, 8]. Consequently, our approach seeks to tackle and improve the quality of text segmentation to benefit the downstream tasks.

1.2. Models

Parts of this section were presented in the main text, and in the following, we offer a more comprehensive review. Prior work has employed various methods for handwritten and printed text segmentation. Early approaches [9, 10] treated the problem as a binary classification task using classifiers like KNN and SVM. They utilized connected components (CCs) (i.e., group of pixels), and various sets of features to determine whether a CC represents handwritten (HT) or printed (PT) text. These features include geometric features of CCs such as their heights, widths, and the spread between CCs [9], as well as geometric-invariant features such as invariant moments [10]. More recently, Li et al. [11] applied conditional random fields (CRFs), with formulating both unary and pair-wise potentials for adjacent connected components by leveraging convolutional neural networks (CNNs) architecture for the separation of CCs. The primary limitation of CC-based approaches is that they assign a single class for the entire component, rather than assigning classes at the pixel-level. Additionally, they fail to detect overlapping regions since the entire connected component is categorized as either HT or PT.

Due to the drawbacks of connected components, pixel-level segmentation methods were introduced that leverage Markov random fields (MRFs) [12]. The authors of [12] applied both patch-level and pixel-level classification for PT and HT segmentation. These classifications were initially identified using a G-means based approach (a modified version of k-means [13]), followed by a relabeling step based on MRFs. Seuret et al. [14] classified the foreground pixels as either printed or handwritten text using an MLP architecture with two fully-connected hidden units. This MLP was followed by a post-processing step designed to correct probable mistakes based on adjacent pixels.

As encoder-decoder architectures have proven to perform well in object segmentation [15], recent works [2, 16, 17, 18] have predominantly applied a U-Net based architecture [15] for HT and PT segmentation. This architecture consists of an encoder-decoder design, similar to SSP shown in Figure 2. Jo et al [16] utilized a U-Net architecture to perform binary classification of handwritten text. Similarly, authors in several studies [2, 17, 18] leveraged a fully convolutional network (FCN) to classify three classes: handwritten (HT), printed (PT), and background

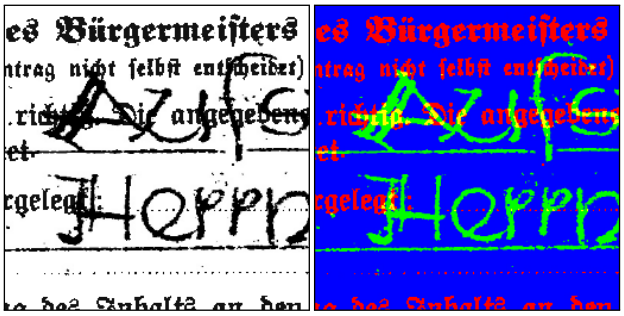
	Group	Layer type	Filter	Input(s)	Output(s)	Output size
Input	_	Input	_	img_input	i_o	$256 \times 256 \times 3$
Fine Feature Path	G 1	FFP	_	i_o	FFP_o	$256 \times 256 \times 4$
		BatchNorm	_	FFP_o	g1_b_o	$256 \times 256 \times 4$
		ReLu	_	g1_b_o	g1_r_o	$256 \times 256 \times 4$
Semantic Segmentation Path	G 2	SSP	_	i_o	SSP_o	$256 \times 256 \times 4$
		BatchNorm	_	SSP_o	g2_b_o	$256 \times 256 \times 4$
		ReLu	_	g2_b_o	g2_r_o	$256 \times 256 \times 4$
Concatenation	G 3	Concat	_	g1_r_o, g2_r_o	g3_cc_o	$256 \times 256 \times 8$
		Conv	$1 \times 1/4$	g3_cc_o	g3_cv_o	$256 \times 256 \times 4$
		Softmax	_	g3_cv_o	g3_s_o	$256 \times 256 \times 4$
Output	_	Output	_	g3_s_o	MFM_{output}	$256 \times 256 \times 4$

Table 1: The architecture and connections of the high-level model, Mixed Feature Model (MFM).

(BG). The approaches are adapted for different applications such as web-based services [17] and understanding of historical documents [2, 18]. They also incorporate a conditional random field (CRF) post-processing step to re-label pixels based on their adjacent majority pixels. These approaches adhere to a three-class formulation of text segmentation, which assigns overlapping pixels to either the HT or PT classes. Moreover, the CRF post-processing often performs aggressive re-labeling that degrades the segmentation performance [2, 18].

1.2.1 WGM-SYN Dataset

Besides the SignaTR6K dataset, we also performed evaluations on the WGM-SYN dataset [2]. This dataset contains a subset of historical and archival records and documents from the "Pilotprojekt zur Wiedergutmachung" archive [19]. The dataset is comprised of forms, typewritten certificates, declarations, and testimonies with different layouts, both in color and grayscale. Then, the documents were manually annotated for different text types with VOTT3 [20]. This process resulted in 319 images of handwritten and 767 images of machine-written text, both from microfilm and document scans. Finally, after some preprocessing, noise removal, and binarization steps, data synthesis techniques adopted from [16] are applied. This process yields final training, validation, and testing set of sizes 3335, 430, and 430, respectively. Each data sample consists of a grayscale image crop of size 256*256 pixels containing both handwritten and printed text, along with a color-coded label: handwritten in green, printed in red, and background in blue. Figure 1 illustrates an example from the WGM-SYN dataset.



(a) Gray-scale (b) Ground truth

Figure 1: An example from the WGM-SYN dataset [2]. Red: class PT, printed, Green: class HT, handwritten, and Blue: class BG, background.

2. Architecture

In this section, we provide additional details on the overall architecture of MFM in Table 1 outlines the stages and connections for MFM, and Figure 2 offers a detailed depiction of the SSP and FFP blocks.

3. Additional Experiments and Results

3.1. Experiment Configuration

Table 2 provides additional details on the experimented configurations. For learning rate (LR), we start with the initial value of 0.001 and divide by ten if there is no improvement in validation loss values after four epochs.

3.2. Additional Results

Tables 3 and 4 show the IoU values for WMG-SYN and SignaTR6K datasets, respectively. Overall trends of the results show that going from the three-class formulation to the four-class formulation improves the IoU values. Additionally, using larger model backbones generally improves the segmentation performance. Among all the model architectures, the ResNet34 and InceptionV3 backbones achieve the highest performance, which we attribute to their residual connections and varied-size convolutions as they can better incorporate the finer features from the image.

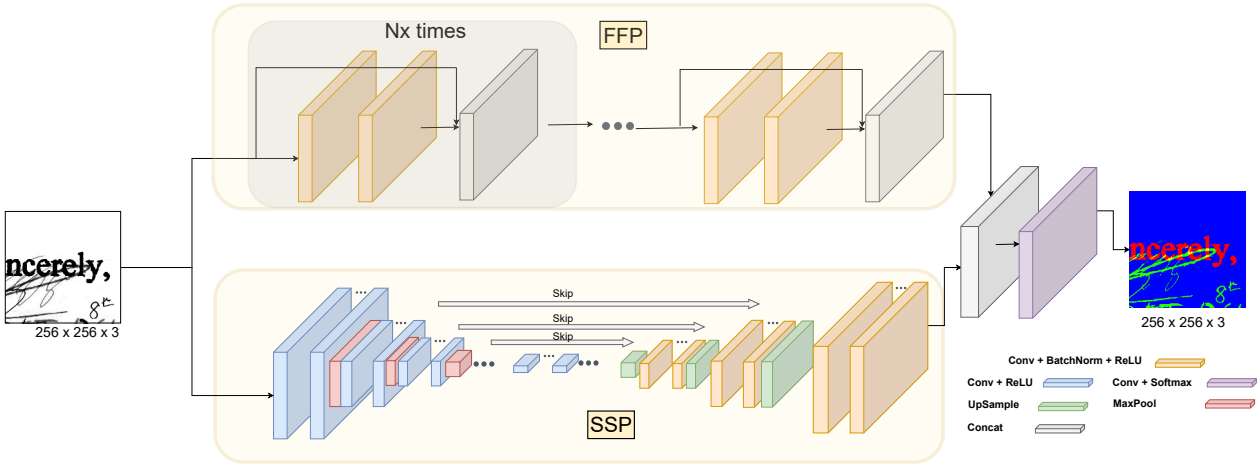


Figure 2: Our proposed architecture uses the fine feature path (FFP) in parallel with the U-Net architecture (SSP) to capture low-level image features, while the U-Net captures high-level features through a condensing and expanding pipeline. In SSP, the three dots in the middle indicate that different sets of encoders and decoders can be applied. As such, we explored various backbones, including FCN-light, VGG16, InceptionV3, and ResNet34.

Parameter	Details
Training epochs	50
Batch size	8
Loss functions	Cross-entropy, WCE, Focal, WF, Dice, WD, and Fusion
Weighted Loss functions weights	3-class: [PT, HT, BG] = [0.4, 0.5, 0.1], 4-class: [PT, HT, BG, OV]=[0.3, 0.3, 0.1, 0.3]
Initial learning rate (LR)	0.001
LR schedule	LR = LR/10 if no reduction in validation loss for 4 epochs.
Initial weights	FCN: None; SSP: None; MFM: SSP initialized with prior training weights
Optimizer	Adam
SignaTR6K	Training: 5169; Validation: 530; Test: 558
WGM-SYN	Training: 3335; Validation: 430; Test: 430
Problem formulation	three-class and four-class
Architecture variations	WGM-MOD & FCN-light (3-class); FCN-light (4-class); SSP (VGG16, InceptionV3, and Resnet34); MFM (VGG16, InceptionV3, and ResNet34)

Table 2: Experimentation parameters.

3.3. Visual Comparisons

Figures 3 and 4 provide additional visual comparisons for WGM-SYN and SignaTR6K datasets. Figures 3b and 4b show the ground truth, and we can visually observe a trend that the performance on the HT and PT overlapping regions improves from (c) to (p). It is also visually noticeable that CRF post-processing aggressively relabels pixels, and CRFH generally improves the results.

3.4. Dataset Availability

The SignaTR6K dataset is available for download through this link: `https://forms.office.com/r/2a5RDg7cAY`.

References

[1] Sina Gholamian and Ali Vahdat. Handwritten and Printed Text Segmentation: A Signature Case Study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[2] Mahsa Vafaie, Oleksandra Bruns, Nastasja Pilz, Jörg Waitelonis, and Harald Sack. Handwritten and printed text identification in historical archival documents. In *Archiving Conference*, volume 19, pages 15–20. Society for Imaging Science and Technology, 2022.

[3] Noman Islam, Zeeshan Islam, and Nazia Noor. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*, 2017.

[4] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020.

[5] Amazon Textract. `https://aws.amazon.com/textract/`. Accessed: 2023-01-21.

[6] Tesseract. `https://github.com/tesseract-ocr/tesseract`. Accessed: 2023-01-21.

[7] ABBYY FineReader PDF. `https://pdf.abbyy.com/`. Accessed: 2023-01-21.

Formulation	Backbone	# Parameters	Loss function	IoU %				With CRF (IoU %)				With CRFH (IoU %)			
				PT	HT	BG	Mean	PT	HT	BG	Mean	PT	HT	BG	Mean
3-Class	**FCN-light** [17]	∼295K	Weighted CE	24.00	26.00	72.00	41.00	11.00	23.00	72.00	36.00	-	-	-	-
	WGM-MOD [2]	∼295K	Weighted CE	42.00	36.00	74.00	50.00*	41.00	32.00	74.00	49.00	-	-	-	-
4-Class (Ours)	**FCN-light**	∼295K	CE	46.49	41.98	73.77	54.08	38.95	29.35	73.92	47.41	46.57	41.59	73.87	54.01
			Focal	46.02	42.01	73.80	53.95	32.96	24.50	74.00	43.82	46.11	41.23	73.98	53.77
			Dice	48.01	47.28	71.22	55.55	48.59	48.48	71.22	56.10	48.07	47.79	71.29	55.72
			Weighted CE	46.07	42.18	73.82	54.02	40.54	30.07	73.98	48.20	46.31	41.68	73.95	53.98
			Weighted Focal	43.71	41.77	73.95	53.14	32.67	24.40	74.07	43.71	44.06	41.12	74.11	53.10
			Weighted Dice	47.97	47.22	71.02	55.40	48.69	48.34	71.02	56.02	48.04	47.70	71.10	55.61
			Fusion	48.12	47.25	71.93	55.77	46.34	44.25	71.86	54.15	48.19	47.89	72.38	56.15
	SSP - VGG16	∼24M	CE	35.14	32.40	73.96	47.17	30.97	19.14	73.88	41.33	35.21	32.08	74.13	47.14
			Focal	34.56	32.22	74.02	46.93	29.11	19.42	73.91	40.81	34.67	31.79	74.22	46.89
			Dice	40.64	39.64	71.13	50.47	43.30	40.16	71.21	51.55	40.81	40.25	71.25	50.77
			Weighted CE	35.73	33.44	74.43	47.87	32.00	20.70	74.12	42.28	35.93	32.95	74.65	47.84
			Weighted Focal	34.90	34.35	74.36	47.87	29.18	20.72	74.06	41.32	35.11	33.80	74.56	47.82
			Weighted Dice	41.80	40.64	70.67	51.03	44.14	41.39	70.70	52.07	42.52	41.19	70.97	51.56
			Fusion	39.99	39.56	72.31	50.62	35.39	27.93	72.35	45.22	40.10	39.94	72.63	50.89
	MFM (FFP + SSP) - VGG16	∼24M	CE	42.41	40.41	73.76	52.19	33.84	25.74	73.97	44.52	42.44	39.13	74.09	51.89
			Focal	41.61	39.11	73.77	51.50	30.85	23.13	74.04	42.67	41.64	37.74	74.13	51.17
			Dice	21.33	19.96	80.50	40.59	23.39	21.96	80.00	41.78	21.35	20.00	80.51	40.62
			Weighted CE	42.60	40.83	73.85	52.43	34.88	25.90	74.03	44.94	42.63	39.61	74.16	52.13
			Weighted Focal	42.22	40.83	73.85	52.30	30.79	23.74	74.10	42.88	42.25	39.51	74.18	51.98
			Weighted Dice	28.51	27.57	76.39	44.16	30.89	29.32	75.67	45.29	28.70	27.87	76.37	44.31
			Fusion	44.80	45.64	72.10	54.18	40.77	36.92	72.00	49.90	44.88	44.67	72.86	54.14
	SSP - InceptionV3	∼30M	CE	49.54	42.13	74.10	55.26	42.80	33.23	74.17	50.07	49.56	41.24	74.34	55.05
			Focal	47.49	42.03	73.94	54.49	35.24	26.26	74.11	45.20	47.54	40.95	74.22	54.24
			Dice	51.22	49.71	71.39	57.44	51.87	**52.01**	71.38	58.42	51.32	50.42	71.56	57.77
			Weighted CE	48.82	41.55	74.34	54.90	43.84	32.55	74.26	50.22	48.92	40.63	74.59	54.71
			Weighted Focal	45.06	40.39	74.19	53.21	34.27	24.89	74.29	44.48	45.27	39.25	74.52	53.01
			Weighted Dice	51.35	49.58	71.00	57.31	52.04	51.12	71.03	58.06	51.45	50.32	71.15	57.64
			Fusion	51.29	49.51	71.71	57.50	48.82	46.27	71.78	55.62	51.37	50.46	72.15	58.00
	MFM (FFP + SSP) - InceptionV3	∼30M	CE	52.51	44.09	73.91	56.84	46.03	35.94	74.21	52.06	52.55	42.17	74.35	56.36
			Focal	51.56	43.74	73.89	56.40	35.89	27.15	74.41	45.82	51.63	41.74	74.43	55.93
			Dice	20.91	19.28	81.52	40.57	25.51	24.61	79.94	43.35	20.92	19.29	81.54	40.58
			Weighted CE	51.58	43.72	73.95	56.42	44.78	34.04	74.32	51.05	51.64	41.85	74.40	55.96
			Weighted Focal	51.01	43.62	73.89	56.18	35.76	26.60	74.22	45.53	51.08	41.60	74.35	55.68
			Weighted Dice	22.30	21.25	80.73	41.42	29.11	29.48	77.97	45.52	22.32	21.26	80.74	41.44
			Fusion	53.22	51.25	71.84	58.77	50.05	48.83	72.03	56.97	**53.32**	49.83	72.72	58.62
	SSP - ResNet34	∼24M	CE	51.94	43.74	73.94	56.54	45.76	35.32	74.18	51.75	51.98	43.13	74.08	56.39
			Focal	50.99	43.45	73.93	56.12	35.76	27.78	74.33	45.96	50.99	42.34	74.19	55.84
			Dice	20.88	19.54	**82.08**	40.83	22.00	20.83	81.89	41.57	20.88	19.55	**82.08**	40.83
			Weighted CE	51.45	43.27	74.03	56.25	45.76	35.01	74.13	51.63	51.43	42.55	74.20	56.06
			Weighted Focal	50.16	43.02	73.97	55.72	35.64	27.54	74.25	45.81	50.18	41.89	74.25	55.44
			Weighted Dice	52.04	50.75	70.97	57.92	52.35	51.94	71.00	58.43	52.09	51.25	71.07	58.14
			Fusion	52.58	50.69	71.77	58.35	50.19	48.91	71.94	57.01	52.60	51.86	72.19	58.88
	MFM (FFP + SSP) - ReNet34	∼24M	CE	52.40	44.02	73.91	56.78	47.52	36.40	74.12	52.68	52.43	42.12	74.34	56.30
			Focal	51.95	43.98	73.89	56.61	36.73	27.17	74.31	46.07	51.90	41.99	74.33	56.08
			Dice	29.68	29.43	77.06	45.39	31.62	25.39	75.15	44.05	29.84	29.63	77.07	45.52
			Weighted CE	52.63	44.12	73.94	56.90	47.60	37.25	74.12	52.99	52.56	42.22	74.36	56.38
			Weighted Focal	51.68	43.95	73.90	56.51	35.73	27.22	74.30	45.75	51.72	41.98	74.34	56.01
			Weighted Dice	51.44	50.60	71.96	58.00	52.97	51.42	71.90	58.76	51.54	49.37	72.75	57.89
			Fusion	52.99	51.72	71.69	58.80	51.15	50.79	71.83	57.92	53.01	51.29	72.49	**58.93**

Table 3: IoU performance (%) on the WGM dataset [2]. The maximum value of a column (i.e., class) is underlined. The overall maximum of a class with different post-processing is marked in bold and underlined. For example, the best mean IoU for the WGM-SYN dataset is for Fusion loss, with CRFH, and MFM-ResNet34 architecture at **58.93**. The best performing configuration of prior work, marked with (*), is 50.00.

[8] Bala Mallikarjunarao Garlapati and Srinivasa Rao Chalamala. A system for handwritten and printed text classification. In *2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)*, pages 50–54. IEEE, 2017.

[9] Jürgen Franke and Matthias Oberlander. Writing style detection by statistical combination of classifiers in form reader applications. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 581–584. IEEE, 1993.

[10] R Kandan, Nirup Kumar Reddy, KR Arvind, and AG Ramakrishnan. A robust two level classification algorithm for text localization in documents. In *Advances in Visual Computing: Third International Symposium, ISVC 2007, Lake Tahoe, NV, USA, November 26-28, 2007, Proceedings, Part II 3*, pages 96–105. Springer, 2007.

[11] Xiao-Hui Li, Fei Yin, and Cheng-Lin Liu. Printed/handwritten texts and graphics separation in complex documents using conditional random fields. In *2018 13th IAPR International Workshop on*

Formulation	Backbone	# Parameters	Loss function	IoU %				With CRF (IoU %)				With CRFH (IoU %)			
				PT	HT	BG	Mean	PT	HT	BG	Mean	PT	HT	BG	Mean
3-Class	**FCN-based [17, 2]**	∼295K	CE	62.56	88.09	98.40	83.02*	52.68	89.68	99.26	80.46	62.72	90.58	99.05	84.11
			Focal	62.34	88.00	98.45	82.93	44.60	84.86	99.28	76.25	62.57	90.83	99.21	84.21
			Dice	60.45	87.29	97.85	81.86	61.24	88.83	98.17	82.74	60.52	87.88	97.97	82.12
			Weighted CE	60.58	84.89	97.78	81.09	53.45	90.73	99.52	81.23	60.84	86.33	98.24	81.80
			Weighted Focal	61.25	85.74	98.02	81.67	44.27	85.77	99.50	76.51	61.55	87.42	98.58	82.52
			Weighted Dice	60.21	87.00	97.74	81.65	60.72	88.17	97.98	82.29	60.27	87.46	97.83	81.86
			Fusion	61.52	88.39	98.38	82.76	58.94	92.55	99.37	83.62	61.67	90.45	98.91	83.68
4-Class (Ours)	**FCN-based**	∼295K	CE	64.55	89.21	98.39	84.05	54.60	89.68	99.23	81.17	64.87	91.81	99.06	85.25
			Focal	64.10	88.86	98.32	83.76	46.64	86.01	99.26	77.30	64.34	91.65	99.11	85.03
			Dice	64.37	88.68	98.37	83.81	65.17	90.13	98.59	84.63	64.57	89.14	98.46	84.06
			Weighted CE	63.78	87.52	98.16	83.15	54.77	90.50	99.42	81.56	64.12	89.43	98.71	84.09
			Weighted Focal	63.68	87.71	98.19	83.20	48.18	87.10	99.45	78.24	64.05	89.72	98.80	84.19
			Weighted Dice	64.21	88.57	98.30	83.69	65.08	90.10	98.53	84.57	64.44	89.06	98.39	83.96
			Fusion	64.68	88.48	98.33	83.83	59.83	91.77	99.28	83.63	65.00	90.72	98.91	84.87
	SSP - VGG16	∼24M	CE	47.36	81.02	98.23	75.54	32.20	81.80	99.52	71.18	47.44	84.20	99.06	76.90
			Focal	47.17	81.46	98.18	75.60	28.51	79.66	99.53	69.23	47.21	8.458	99.05	76.95
			Dice	52.83	80.12	97.17	76.71	54.13	84.14	97.95	78.74	52.99	81.45	97.49	77.31
			Weighted CE	52.80	83.54	97.95	78.10	37.51	85.04	99.61	74.05	52.99	81.45	97.49	77.31
			Weighted Focal	47.24	81.03	97.92	75.40	28.07	80.15	99.61	69.28	47.39	82.85	98.42	76.22
			Weighted Dice	43.54	74.56	96.45	71.52	42.80	79.25	97.48	73.17	43.63	76.54	96.93	72.37
			Fusion	49.36	80.54	97.82	75.91	38.66	84.94	99.59	74.39	49.52	83.31	98.57	77.14
	MFM (FFP + SSP) - VGG16	∼24M	CE	61.52	88.90	98.67	83.03	45.72	87.01	99.54	77.42	61.69	91.27	99.34	84.10
			Focal	60.99	88.45	98.65	82.69	37.41	82.76	99.55	73.24	61.08	90.85	99.34	83.75
			Dice	59.73	88.43	98.65	82.27	61.66	90.50	98.94	83.70	60.42	90.75	99.30	83.49
			Weighted CE	57.15	87.28	98.52	80.98	41.46	86.51	99.61	75.86	57.54	88.94	98.99	81.83
			Weighted Focal	57.90	87.35	98.49	81.24	36.60	82.71	99.60	72.97	58.35	89.05	98.97	82.12
			Weighted Dice	58.14	87.67	98.62	81.48	60.19	90.31	98.98	83.16	58.82	90.36	99.37	82.85
			Fusion	59.80	88.33	98.62	82.25	51.14	90.13	99.59	80.28	60.46	90.25	99.18	83.30
	SSP - InceptionV3	∼30M	CE	72.82	92.44	98.73	87.99	63.32	92.91	99.55	85.26	72.77	94.90	99.29	88.99
			Focal	72.20	92.14	98.69	87.68	56.95	90.34	99.54	82.28	72.18	94.86	99.32	88.79
			Dice	71.11	91.11	98.55	86.92	71.54	93.53	99.07	88.05	71.11	92.22	98.79	87.37
			Weighted CE	70.32	90.53	98.34	86.39	61.39	92.52	**99.63**	84.52	70.35	92.15	98.72	87.07
			Weighted Focal	71.64	91.06	98.39	87.03	56.10	90.48	99.60	82.06	71.62	92.68	98.79	87.69
			Weighted Dice	71.96	91.60	98.59	87.39	72.15	93.29	98.96	88.14	71.92	92.35	98.76	87.68
			Fusion	71.19	91.10	98.51	86.93	65.80	93.97	99.59	86.45	71.18	93.32	99.01	87.84
	MFM (FFP + SSP) - InceptionV3	∼30M	CE	73.10	92.66	98.77	88.18	63.48	92.89	99.55	85.31	73.05	94.89	99.36	**89.10**
			Focal	72.80	92.50	98.75	88.01	57.47	90.58	99.54	82.53	72.77	94.74	99.35	88.95
			Dice	72.56	92.20	98.70	87.82	72.38	93.60	99.00	88.33	72.52	94.73	99.37	88.87
			Weighted CE	72.66	91.90	98.54	87.70	62.65	93.04	99.62	85.10	72.63	93.31	98.94	88.29
			Weighted Focal	72.55	92.13	98.62	87.77	59.35	91.64	99.59	83.53	72.50	93.77	99.07	88.45
			Weighted Dice	72.59	92.31	98.70	87.87	72.60	94.00	99.07	88.56	72.54	94.60	99.32	88.82
			Fusion	72.55	92.25	98.71	87.83	65.57	93.49	99.53	86.19	72.49	94.62	99.35	88.82
	SSP - ResNet34	∼24M	CE	73.02	92.27	98.71	88.00	63.68	92.57	99.55	85.27	72.97	94.71	99.27	88.98
			Focal	72.74	92.26	98.71	87.90	58.54	90.95	99.55	83.01	72.71	94.88	99.32	88.97
			Dice	72.40	91.91	98.65	87.65	72.62	93.38	98.98	88.33	72.35	92.55	98.79	87.90
			Weighted CE	72.43	91.34	98.45	87.40	62.94	92.97	**99.63**	85.18	72.43	92.82	98.80	88.02
			Weighted Focal	72.39	91.38	98.43	87.40	58.85	91.38	99.61	83.28	72.35	92.94	98.81	88.03
			Weighted Dice	71.86	91.89	98.65	87.47	71.99	93.32	98.96	88.09	71.79	92.56	98.80	87.72
			Fusion	72.96	91.88	98.60	87.81	68.64	94.86	99.58	87.69	72.92	94.08	99.10	88.70
	MFM (FFP + SSP) - ResNet34	∼24M	CE	72.81	92.56	98.78	88.05	63.04	92.94	99.55	85.17	72.75	**94.93**	99.39	89.02
			Focal	73.04	92.46	98.75	88.08	53.02	89.04	99.55	80.54	73.00	94.75	99.35	89.03
			Dice	72.96	92.38	98.72	88.02	73.16	93.35	98.93	88.48	72.93	94.79	99.36	89.03
			Weighted CE	72.96	91.96	98.69	87.87	64.49	92.47	99.56	85.51	72.90	94.19	99.27	88.79
			Weighted Focal	73.18	92.10	98.63	87.97	54.96	89.71	99.60	81.42	73.16	93.72	99.06	88.65
			Weighted Dice	72.78	92.32	98.71	87.94	72.95	93.66	99.00	88.53	72.75	94.67	99.34	88.92
			Fusion	**73.26**	92.45	98.73	88.15	68.38	94.85	99.56	87.60	73.21	94.59	99.31	89.04

Table 4: IoU performance (%) on the SignaTR6K dataset. Partial results for this dataset were presented in the main paper [1] and this table provides the full results. The maximum value of a column (i.e., class) is underlined, and the overall maximum of a class with different post-processing is denoted in bold and underlined. For example, the best mean IoU for the SignaTR6K dataset is for CE loss, with CRFH, and MFM-InceptionV3 architecture at **89.10**. The best performing configuration of prior work, i.e., excluding results for Fusion loss and CRFH, marked with (*), is 83.02.

Document Analysis Systems (DAS), pages 145–150. IEEE, 2018.

[12] Xujun Peng, Srirangaraj Setlur, Venu Govindaraju, and Ramachandrula Sitaram. Handwritten text separation from annotated machine printed documents using markov random fields. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(1):1–16, 2013.

[13] Greg Hamerly and Charles Elkan. Learning the k in k-means. *Advances in neural information processing systems*, 16, 2003.

[14] Mathias Seuret, Marcus Liwicki, and Rolf Ingold. Pixel level handwritten and printed content discrimination in scanned documents. In *2014 14th Interna-*

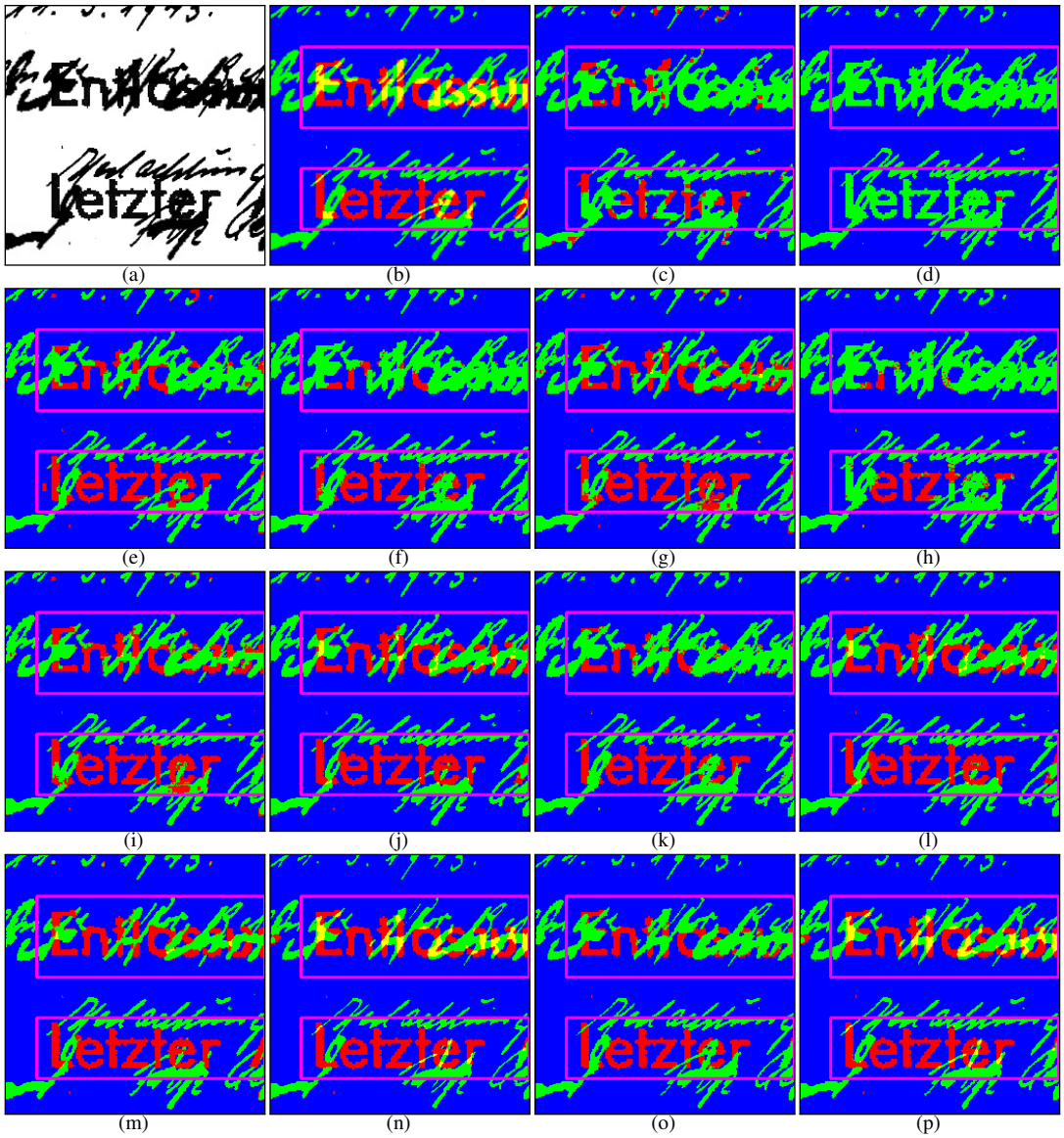


Figure 3: Example visual comparisons on the test set of the WGM-SYN dataset for our approach compared to the ground truth and prior works. (a) Input image; (b) Ground truth; (c) & (d) 3-class FCN-based [17] with CE loss without (c) and with (d) CRF post-processing; (e) & (f) 3-class FCN-based [2] with CE loss without (e) and with (f) CRF post-processing; (g), (h), & (i) Our FCN-based 4-class formulation with CE loss without CRF (g), with CRF (h), and with CRFH (i); (j), (k), & (l) SSP-ResNet34 with CE loss without CRF (j), with CRF (k), and with CRFH (l); (m) MFM-ResNet34 with CE loss without CRF; (n), (o), & (p) MFM-ResNet34 with Fusion loss without CRF (n), with CRF (o), and with CRFH (p).

tional Conference on Frontiers in Handwriting Recognition, pages 423–428. IEEE, 2014.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[16] Junho Jo, Hyung Il Koo, Jae Woong Soh, and Nam Ik Cho. Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications*, 79(43):32137–32150, 2020.

[17] Nicolas Dutly, Fouad Slimane, and Rolf Ingold. PHTI-WS: a printed and handwritten text identification web service based on FCN and CRF post-processing. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 20–25. IEEE, 2019.

[18] Anastasia Prikhodina. Handwritten and printed text separation for historical documents. 2021.

Figure 4: Additional visual results on the test set of the SignaTR200 dataset for our approach compared to the ground truth and prior works. (a) Input image; (b) Ground truth; (c) & (d) 3-class FCN-based [17, 2] with CE loss without (c) and with (d) CRF post-processing; (e), (f), & (g) Our FCN-based 4-class formulation with CE loss without CRF (e), with CRF (f), and with CRFH (g); (h), (i), & (j) SSP-ResNet34 with CE loss without CRF (h), with CRF (i), and with CRFH (j); (k), (l), & (m) MFM-ResNet34 with CE loss without CRF (k), with CRF (l), and with CRFH (m); (n), (o), & (p) MFM-ResNet34 with Fusion loss without CRF (n), with CRF (o), and with CRFH (p).

[19] Wiedergutmachung. https://www.fiz-karlsruhe.de/en/forschung/wiedergutmachung. Accessed: 2023-01-29.

[20] VOTT. https://github.com/microsoft/VoTT. Accessed: 2023-01-29.