A Comprehensive Study of ImageNet Pre-Training for Historical Document Image Analysis

Linda Studer[*†], **Michele Alberti**[*†], **Vinaychandran Pondenkandath**[*†], **Pinar Goktepe**[*†], **Thomas Kolonko**[*†], **Andreas Fischer**[†‡], **Marcus Liwicki**[†§], **Rolf Ingold**[†]

[†]*Document Image and Voice Analysis Group (DIVA)*
University of Fribourg, Switzerland
{firstname}.{lastname}@unifr.ch

[‡]*Institute of Complex Systems (iCoSys)*
University of Applied Sciences and Arts Western Switzerland
andreas.fischer@hefr.ch

[§]*Machine Learning Group*
Luleå University of Technology, Sweden
marcus.liwicki@ltu.se

***Abstract*—Automatic analysis of scanned historical documents comprises a wide range of image analysis tasks, which are often challenging for machine learning due to a lack of human-annotated learning samples. With the advent of deep neural networks, a promising way to cope with the lack of training data is to pre-train models on images from a different domain and then fine-tune them on historical documents. In the current research, a typical example of such cross-domain transfer learning is the use of neural networks that have been pre-trained on the ImageNet database for object recognition. It remains a mostly open question whether or not this pre-training helps to analyse historical documents, which have fundamentally different image properties when compared with ImageNet. In this paper, we present a comprehensive empirical survey on the effect of ImageNet pre-training for diverse historical document analysis tasks, including character recognition, style classification, manuscript dating, semantic segmentation, and content-based retrieval. While we obtain mixed results for semantic segmentation at pixel-level, we observe a clear trend across different network architectures that ImageNet pre-training has a positive effect on classification as well as content-based retrieval.**

I. Introduction

Historical documents span centuries of different writing supports (including stone, palm leaf, papyrus, parchment, and paper in different states of decay), writing instruments, languages, scripts, fonts, ornaments, illustrations, and so on. Furthermore, the image acquisition methods may vary substantially depending on the type of document. When performing automatic image analysis for a specific type of document using machine learning, one of the main challenges is to collect a sufficient amount of representative learning samples. In the case of ancient languages and scripts, such annotations often can only be provided by experts in the respective field and are thus costly to obtain.

[*] These authors contributed equally to this work.

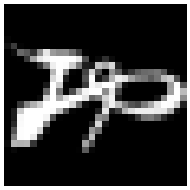
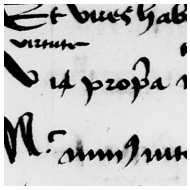
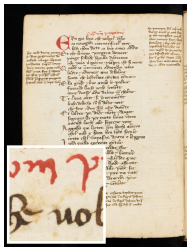
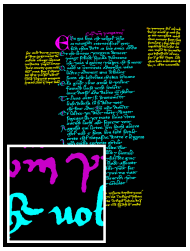
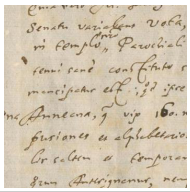
In recent years, the use of deep neural networks has strongly influenced the state of the art for historical document analysis. However, deep neural network models have millions of parameters to fine-tune and a random initialization [1] may not be the best option when facing a lack of annotated training data. Several promising alternatives have been suggested including layer-wise pre-training [2], [3] and transfer learning [4]. The latter is the main focus of the present paper. Transfer learning aims to fine-tune network parameters with respect to another image analysis task – which features a large amount of annotated training data – and then use these parameters as an initialisation for the image analysis task at hand.

Transfer learning is a widespread technique in computer vision [5], [6]. Since the publication of large datasets such as ImageNet [7], CIFAR-10 [8], PASCAL [9], and COCO [10], many architectures have been trained on them and their weights made publicly available to be used for transfer learning. Although transfer learning has been around for the last two decades [11], it has only become popular in the last years with the breakthrough of Convolutional Neural Networks (CNNs) architectures consistently winning the Large Scale Visual Recognition Challenge (ILSVRC) [7] competition since 2012 [12]. Pre-training on ImageNet and successive fine-tuning on another dataset has become a widely used practice [13], [14]. It is generally believed that this approach helps to learn good and general features.

Contribution

Previous studies mostly focus on fine-tuning on datasets similar to the dataset used for pre-training. Moreover, they only explore a small set of tasks and neural network architectures. Historical documents have very different image properties when compared to the natural images found in the ImageNet dataset. It is therefore not immediately intuitive that pre-training a model on ImageNet for historical document analysis will have the same benefits.

TABLE I: This table gives an overview of the different tasks. The Kuzushiji-MNIST (KMNIST) dataset contains different Hiragana (cursive Japanese) characters, here depicted is the character for "o". The expected output is the character label of the image. The Classification of Medieval Handwritings in Latin Scripts (CLaMM) dataset is annotated for style classification and manuscript dating. The expected output is the style or date label for a given image. The Historical Manuscript Database DIVA-HisDB is annotated for semantic segmentation at pixel level. The example shown here is from manuscript CB55. The output is a segmentation label for each pixel, here shown with different colours. The Historical Writer Identification (Historical-WI) dataset consists of images from different writers. Here a section from one of the pages from writer 100 is depicted. The output of the network is a ranking of the most similar writers, based on the input image.

Dataset	Input	Task	Output
KMNIST		Hiragana Classification	お (o)
CLaMM		Style Classification Manuscript Dating	Semihybrida 1451-1475 C.E.
DIVA-HisDB		Semantic Segmentation at Pixel Level	
Historical WI		Writer Identification	Writer 100

In this paper, we provide a comprehensive empirical study of the impact of transfer learning from ImageNet pre-trained models to historical document analysis. A variety of different applications, datasets and network architectures are taken into account. The applications can be grouped into three categories: classification, semantic segmentation at pixel level and content-based image retrieval. Most of the datasets we use were published as part of previous competitions at the International Conference on Document Analysis and Recognition (ICDAR). Note that the main aim of our study is to investigate the effect of pre-training on relevant and high-quality datasets, and not to outperform the winners of the competitions by optimizing task-specific pre- and post-processing methods.

II. Related Work

ImageNet is the most widely used dataset for pre-training and transfer learning. Popular beliefs as to why ImageNet is particularly suited for this task are its large size, the high number of distinct classes and the close similarity of many of the classes, e.g. a number of different dog breeds.

Huh et al. [13] examined the impact of various aspects of ImageNet pre-training and successive fine-tuning on the PASCAL [9] dataset, such as dataset size, number of classes, using fine-grained versus coarser class labels and the ratio of images per class. Additionally, they showed that the aforementioned commonly held beliefs are not accurate, and that transfer learning still works well with restrictions, such as only using half of the dataset.

He et al. [15] showed that pre-training on ImageNet speeds up convergence early in training, but that training from scratch will eventually catch up and sometimes even surpass the pre-trained and fine-tuned accuracy. They further argue that ImageNet pre-training does not automatically give better regularisation and that it shows no benefit when the target tasks or metrics are more sensitive to spatially localised predictions. Training from scratch requires different normalisation and regularisation methods as compared to transfer learning. This can skew results, benefiting the pre-trained paradigm over learning from scratch.

Similarly, Kornblith et al. [16] showed that although ImageNet pre-training accelerates convergence, it does not necessarily lead to a better performance if run long enough. They also investigated how transfer learning relates to the architecture used in the context of image classification. Their findings suggest that the accuracy increase from ImageNet pre-training fades quickly as the size of the dataset for the task at hand grows larger. They conclude that pre-training is and will remain an essential tool in the near future but also highlight clearly that it has limitations.

When it comes to cross-domain transfer learning, its usefulness - especially from natural images such as ImageNet - is subject of an open discussion. There are cases where transfer learning across domains has been proven to be successful [17], [18]. In contrast, there is literature suggesting that this technique is harmful to the final performance of the networks. Tensmeyer et al. [19] specifically question the usefulness of transfer learning from ImageNet (natural images of 3D objects) on document analysis, which are 2D entities. They argue that feature mappings for natural images fundamentally differ from feature mappings for documents. They also question whether architectures optimized for natural image classification are a good fit for historical document analysis. In their work, they do not provide conclusive results, focus only on the AlexNet [12] architecture and lack a thorough comparative study. The impact of transfer learning from ImageNet was also a topic of discussion at ICDAR 2017. The organizers of the Competition on Historical Document Writer Identification [20] speculated that the competition participant who used a deep learning-based method, may have performed relatively poorly due to their network being initialized with ImageNet pre-trained weights.

Therefore, a conclusive answer on what the real reasons behind the benefit of transfer learning are and whether these can be harnessed in a cross-domain scenario is yet to be found.

III. Study Design

In this section, we present the details of our empirical study, namely the tasks we choose to perform, the datasets and network architectures we use and finally the training procedure for each specific task.

A. Tasks

In this work, we choose three tasks as representatives of some of the most common challenges in the field of historical document analysis. Specifically, our use cases include image classification, semantic segmentation at pixel level and content-based image retrieval. We believe that their radically different natures will give a robust estimation of the generality of our findings. Table I gives an overview of the input and output of the different tasks.

1) Classification: This task is well known in the computer vision community and consists of producing one or more descriptive labels for a given input image. In the context of historical image document analysis this task can be, for example, formulated as character recognition [21], [22], style/script classification [23], [24] or manuscript dating [24], [25]. We train the networks to minimize the cross-entropy loss function shown below:

$$L(\vec{x}, y) = -log\left(\frac{e^{\vec{x}_y}}{\sum_{i=0}^{n} e^{\vec{x}_i}}\right) \tag{1}$$

where n is the number of classes, x is a vector of size n representing the output of the network, and $y = \{1..n\}$ is a scalar representing the class label, e.g. style of the document.

2) Semantic Segmentation at Pixel Level: Semantic segmentation at pixel level is a specific case of a classification task. In this task, each pixel of an input image has to be assigned a label. This is often performed to analyse the layout of historical documents [26], [27], or as a form of pre-processing for further tasks, e.g. line segmentation [28]. Neural networks for semantic segmentation are trained similarly to the ones for classification, but the architectures employed are different (see section III-C).

3) Content-based Image Retrieval: Image similarity (or content-based image retrieval) is another typical scenario found in computer vision. In the context of historical document image analysis, this can be seen in the form of writer identification [20], signature verification [29] or watermark recognition [18]. To train the networks for this task, we use the triplet loss approach [30], [31]. The triplet loss operates on a tuple of three images $\{a, p, n\}$ where a is the anchor (reference), p is the positive sample (an image of the same class as the reference), and n is the negative sample (an image of another class). The loss function is then defined as:

$$L(\delta_+, \delta_-) = max(\delta_+ - \delta_- + \mu, 0) \tag{2}$$

where δ_+ and δ_- are the Euclidean distances between the anchor-positive and anchor-negative pairs in the high dimensional output space of the network and μ is the margin used.

B. Datasets

The datasets are available for download through DeepDIVA[1]. The image input size depends on the architecture and is described in section III-D. Table I shows an example image for each dataset.

1) Kuzushiji-MNIST: The KMNIST dataset [21] contains grayscale images of ten different Hiragama characters written in *Kuzushiji* (cursive Japanese). It is a curated subset of the full Kuzushiji dataset, which was created during the digitisation of around 300'000 old Japanese books. The images in KMNIST are from 35 classic books printed in the 18th century. The training set contains 7'000 and the test set 1'000 images per class, each of size 28×28 pixels.

2) ICDAR2017 Classification of Medieval Handwritings in Latin Scripts: This dataset was published for the Classification of Medieval Handwritings in Latin Scripts (CLaMM) competition [24] at the ICDAR 2017 conference and includes 3'540 images annotated for style classification and manuscript dating. The test set contains 2'000 images. The dataset is divided into 12 classes for style classification (Caroline, Cursiva, Half-Uncial, Humanistic, Humanistic Cursive, Hybrida, Praegothica, Semihybrida, Semitextualis, Southern Textualis, Textualis, Uncial). Each of the 15 classes provided for manuscript dating corresponds to a specific time interval, ranging from 500 C.E. to 1600 C.E. The competition provided two variations of the dataset, we use the subset that contains images of mixed resolutions and colour representations.

3) ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts: The DIVA-HisDB dataset [32] consists of three different medieval manuscripts (CB55, CSG18, CSG863), each containing 50 pixel-wise annotated pages with a size of approximately $4k \times 5.5k$ pixels. The manuscripts have a challenging layout with four different classes (main text body, decoration, comment and background). There is also an additional label for boundary pixels. These pixels originate from the labelling process and are background pixels along the border of the text, which are labelled as text. For our purposes, we relabelled these pixels as background. The same training and test dataset split is used as in the ICDAR 2017 Competition on Layout Analysis for Challenging Medieval Manuscripts [33].

4) ICDAR2017 Historical Writer Identification: The Historical-WI dataset [20] consists of colored and binarized versions of handwritten historical documents. The training dataset consists of 394 writers with three pages per writer, which gives a total of 1'182 pages. The dataset for the competition contains 720 writers and five pages per writer, which makes 3'600 instances in total. We use the coloured version of the dataset for our experiments.

C. Model Architectures

For the classification and content-based image retrieval experiments, we investigate four well-known architectures: VGG19 with batch normalisation (VGG19 BN), Inception V3, ResNet152 and DenseNet121. A simple architecture with only

[1] https://bit.ly/DeepDIVA

three layers is also trained for each task to give a baseline for the performance. VGG19 uses batch normalisation [34] and consists of alternating blocks of convolutional and max-pooling layers. The Inception architecture [35] introduced Inception blocks, which combine different convolutional filters and layers into one block by concatenation. A second classification head further back in the architecture is added to counteract the vanishing gradient. Another way to combat the vanishing gradient was introduced with the ResNet architecture [36]. The layers in these type of networks contain direct, additive connections from one layer to a next one, so-called skip connections. We use the ResNet152 architecture. The idea of skip connections has also been extended to not only connect one layer to the next one, but to have blocks of densely-connected layers. Unlike for the skip connections, the output is not added but concatenated. These dense blocks are alternated with 1×1 convolutions and max-pooling layers in order to reduce the number of parameters in the model. Here, we use the DenseNet121 [37], which features such dense blocks.

For the segmentation experiments, we use two different architectures: SegNet and DeepLabV3. SegNet [38] can use any VGG architecture as an encoder, we use VGG19 BN. For each encoder layer, there is a decoder layer which uses the max-pooling indices from the respective encoder layer to perform non-linear up-sampling. The DeepLabV3 architecture [39] uses a ResNet as the encoder, ResNet18 in our case, and an Atrous Spatial Pyramid Pooling (ASPP) as the decoder. A simple architecture with five layers is also trained on the dataset to give a baseline for the performance.

Deep learning is a fast-moving field, and in recent years, the models we use marked milestones in advances in terms of network architecture. All the architectures are available in DeepDIVA[2].

D. Training Procedure

All experiments are run using the DeepDIVA framework [40]. The models are trained long enough to reach convergence on the training data. Each architecture is trained from scratch as well as with ImageNet [7] pre-training to evaluate the effect of pre-training. All hyper-parameters can be found in our fork of DeepDIVA[3].

1) Classification: The architectures described in section III-C are trained with data balancing. Three different classification tasks are performed. For the character recognition task on the KMNIST dataset, the models are trained for 35 epochs. The input images are resized to match the required input size of the respective network. On the CLaMM dataset, the models are trained for 50 epochs for manuscript dating and style classification. 10 random sections of the required input size of the respective network are sampled from each input image, evaluated and their output is averaged.

2) Semantic Segmentation at Pixel Level: The architectures described in section III-C are trained for 50 epochs. Since the images are very large, using the whole image as an input for the network is not feasible. Instead, a total of $60'000$ sections of size 256×256 are randomly sampled per training epoch. In the test phase, crops of size 256×256 are sampled as a sliding window (with 50% overlap) to segment the full image. Both architectures used for this task feature encoders, for which ImageNet pre-trained weights are available. For the experiments that use pre-training, we initialise only the encoder using these weights, the weights of the decoder are initialised randomly.

3) Content-based Image Retrieval: The architectures (see Section III-C) are trained for ten epochs with 1.5 million triplets. The 1.5 million triplets are generated every epoch from the training set (see Section III-B4) with 1'182 pages from 394 unique writers. . The evaluation is performed on the test set with 3'600 pages from 720 writers, with each page used as a query in turn. All the networks are designed to embed the input images in a 128-dimensional space. The images are randomly cropped to match the required input size of the respective network. During training one random section per page is fed to the network. During the test phase, ten random sections of the input image are evaluated, and their output is averaged.

IV. Results

In the following, results are presented for each of the three chosen tasks individually, i.e. classification, semantic segmentation at pixel level and content-based image retrieval.

A. Classification

Table II reports the accuracies achieved by the different architectures trained from scratch and with ImageNet pre-training on the KMNIST and CLaMM datasets. In general, the network architectures clearly profit from pre-training.. ResNet152 shows the largest increase in performance while DenseNet121 benefits the least.

1) Optical Character Recognition: For this task, we report the mean accuracy along with the standard deviation computed over five runs of each experiment. Comparing the mean performance of each model using the t-test, the improvement from trained from scratch to pre-trained is statistically significant for VGG19 BN, InceptionV3 and ResNet152. DenseNet121 shows a small decrease of 0.08% with pre-training. The ResNet152 benefits the most from pre-training with a delta of 1.42%, which also makes it the best performing model.

2) Style Classification: Pre-training leads to a higher accuracy for all architectures with an average increase of 8.1%. VGG19 BN, Inception V3 and ResNet152 profit the most from pre-training with an increase in accuracy of around 9.5%.

3) Manuscript Dating: Pre-training improves the performance of all the architectures with an average increase of 11.4%. ResNet152 shows the largest increase in accuracy with 17.3%, which also makes it the best performing model.

B. Semantic Segmentation at Pixel Level: DIVA-HisDB

The performance of the segmentation models are evaluated with the layout analysis tool [41] used in the ICDAR 2017 competition [33]. The tool computes the mean Intersection

[2]https://bit.ly/2R8pBqx

[3]https://bit.ly/2I8c3dX

TABLE II: Accuracy (%) on the test set for the different architectures trained for the different classification tasks. Character recognition is performed on the KMNIST dataset. The reported accuracy is the average along with the standard deviation from five runs. Style classification and manuscript dating are preformed on the CLaMM dataset.

	Character Recognition			Style Classification			Manuscript Dating		
	scratch	pre-trained	Δ	scratch	pre-trained	Δ	scratch	pre-trained	Δ
3-layer CNN	92.98±0.22	n/a	-	12.4	n/a	-	11.7	n/a	-
VGG19 bn	98.17±0.18	98.35±0.15	+0.18	42.5	52.1	+9.6	24.0	36.1	+12.1
Inception v3	97.82±0.11	98.51±0.11	+0.69	46.5	**55.5**	+9.0	24.8	35.4	+10.6
ResNet152	97.27±0.26	**98.69**±0.10	+1.42	39.1	49.3	+10.2	20.6	**37.9**	+17.3
DenseNet121	98.64±0.06	98.56±0.06	-0.08	47.3	50.9	+3.6	30.7	36.4	+5.7

TABLE III: Mean IU (%) on the test set of the different architectures trained on the three manuscripts of the DIVA-HisDB dataset. For pre-training, only the encoder is initialized with the pre-trained weights from ImageNet.

	Manuscript CB55			Manuscript CSG18			Manuscript CSG863		
	scratch	pre-trained	Δ	scratch	pre-trained	Δ	scratch	pre-trained	Δ
5-layer CNN	55.7	n/a	-	56.8	n/a	-	45.6	n/a	-
SegNet	86.9	72.9	-14.0	73.0	**75.3**	+2.3	81.6	61.9	-19.7
DeepLabv3	**92.9**	91.4	-1.5	69.8	73.1	+3.3	85.5	**86.7**	+1.2

TABLE IV: Mean average precision (%) achieved on the test set by the different architectures trained on the Historical-WI dataset for writer identification.

	Writer Identification		
	scratch	pre-trained	Δ
3-layer CNN	11.4	n/a	-
VGG19 bn	14.6	24.0	+9.4
Inception v3	9.1	26.1	+17.0
ResNet152	24.7	22.1	-2.6
DenseNet121	27.2	**34.6**	+8.2

over Union (mean IU) between the predicted results and the ground truth. Table III shows the mean IU achieved by the models on the test set of the three different manuscripts. The results regarding the effect of transfer-learning from ImageNet are mixed for both architectures. On some manuscripts pre-training on ImageNet increases the performance, but on others, the pre-trained network performs much worse. In terms of dataset size, semantic segmentation outnumbers the classification and content-based image retrieval by far, as every pixel is a data point. Kornblith et al. [16] have found that the impact of ImageNet pre-training on the performance of the model becomes smaller the larger the dataset gets. This could explain why pre-training is not beneficial for this task.

C. Content-based Image Retrieval: Writer Identification

Table IV reports the mean average precision achieved by the different architectures trained from scratch and with pre-training on the Historical-WI dataset. Pre-training improves the performance of all models except ResNet152. Inception V3 profits the most from pre-training with an increase in performance of +17.0%.

V. Conclusion

For the classification and content-based image retrieval tasks we find a clear trend that cross-domain transfer learning from ImageNet pre-training leads to an improved performance. For semantic segmentation at pixel level we obtain mixed results. In some instances pre-training is beneficial but harmed the performance in others. We speculate that this could be possibly attributed to the larger amount of training data available for semantic segmentation, as each pixel can be considered an individual data point.

Overall, in historical document image analysis, the lack of annotated training data is often one of the most limiting factors for machine learning. Facing this restriction, ImageNet pre-training can significantly help to improve the performance of deep learning models. In this paper we only investigate the effect of transfer learning from ImageNet. It would also be interesting to see how pre-training on other type of datasets, especially domain-specific ones.

Acknowledgment

The work presented here has been partially supported by the HisDoc III project funded by the Swiss National Science Foundation with the grant number 205120_169618 and by the Rising Tide foundation with the grant number CCR-18-130.

References

[1] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[2] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] M. Alberti, M. Seuret, V. Pondenkandath, R. Ingold, and M. Liwicki, "Historical Document Image Segmentation with LDA-Initialized Deep Neural Networks," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing - HIP2017*, Kyoto, Japan, nov 2017, pp. 95–100.

[4] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[5] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[6] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[8] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," 2014.

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[10] 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*. Springer, 2014, pp. 740–755.

[11] L. Pratt and S. Thrun, *Machine Learning - Special Issue on Inductive Transfer*. Springer, 1997.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *arXiv preprint arXiv:1608.08614*, 2016.

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[15] K. He, R. B. Girshick, and P. Dollár, "Rethinking imagenet pre-training," *CoRR*, vol. abs/1811.08883, 2018.

[16] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" *arXiv preprint arXiv:1805.08974*, 2018.

[17] M. S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwickit, "Transforming sensor data to the image domain for deep learning—an application to footstep detection," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2665–2672.

[18] V. Pondenkandath, M. Alberti, N. Eichenberger, R. Ingold, and M. Liwicki, "Identifying cross-depicted historical motifs," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 333–338.

[19] C. Tensmeyer, D. Saunders, and T. Martinez, "Convolutional neural networks for font classification," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 985–990.

[20] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, S. Nikos, and B. Gatos, "Icdar2017 competition on historical document writer identification (historical-wi)," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1377–1382.

[21] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[22] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A complete optical character recognition methodology for historical documents," in *2008 The Eighth IAPR International Workshop on Document Analysis Systems*. IEEE, 2008, pp. 525–532.

[23] A. M. A. Al-Aziz, M. Gheith, and A. F. Sayed, "Recognition for old arabic manuscripts using spatial gray level dependence (sgld)," *Egyptian Informatics Journal*, vol. 12, no. 1, pp. 37–43, 2011.

[24] F. Cloppet, V. Eglin, M. Helias-Baron, C. Kieu, N. Vincent, and D. Stutzmann, "Icdar2017 competition on the classification of medieval handwritings in latin script," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1371–1376.

[25] F. Wahlberg, T. Wilkinson, and A. Brun, "Historical manuscript production date estimation using deep convolutional neural networks," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 205–210.

[26] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "Icdar 2009 page segmentation competition," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 1370–1374.

[27] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Icdar 2013 competition on historical newspaper layout analysis (hnla 2013)," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1454–1458.

[28] M. Alberti, L. Vögtlin, V. Pondenkandath, M. Seuret, R. Ingold, and M. Liwicki, "Labeling, Cutting, Grouping: an Efficient Text Line Segmentation Method for Medieval Manuscripts," in *2019 15th International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, Sep 2019.

[29] P. Maergner, V. Pondenkandath, M. Alberti, M. Liwicki, K. Riesen, R. Ingold, and A. Fischer, "Offline Signature Verification by Combining Graph Edit Distance and Triplet Networks." Springer, Cham, aug 2018, pp. 470–480.

[30] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds. Cham: Springer International Publishing, 2015, pp. 84–92.

[31] V. Balntas, "Learning local feature descriptors with triplets and shallow convolutional neural networks," *Bmvc*, vol. 33, no. 1, pp. 119.1–119.11, 2016.

[32] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 471–476.

[33] F. Simistira, M. Bouillon, M. Seuret, M. Wursch, M. Alberti, R. Ingold, and M. Liwicki, "Icdar2017 competition on layout analysis for challenging medieval manuscripts," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 1361–1370.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[37] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[38] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *CoRR*, vol. abs/1505.07293, 2015.

[39] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[40] M. Alberti, V. Pondenkandath, M. Würsch, R. Ingold, and M. Liwicki, "Deepdiva: a highly-functional python framework for reproducible experiments," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 423–428.

[41] M. Alberti, M. Bouillon, R. Ingold, and M. Liwicki, "Open evaluation tool for layout analysis of document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4. IEEE, 2017, pp. 43–47.