HANDWRITTEN TEXT RECOGNITION ADAPTATION FOR LOW-RESOURCE LANGUAGES: A CASE STUDY ON HISTORICAL LATIN MANUSCRIPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Handwritten Text Recognition (HTR) remains a challenging task in document digitization, particularly for historical manuscripts written in low-resource languages such as Latin. In this paper, we focus on recognizing Latin texts from 16th–18th century manuscripts, which exhibit a wide range of handwriting styles. To address this, we propose AdapterTrOCR, a modular extension of the TrOCR model that incorporates two adapter modules: one for historical language adaptation and another for handwriting style adaptation. This architecture enables a robust transition from a modern English HTR model to one specialized in historical Latin. Given the limited availability of annotated data, we also explore Handwritten Text Generation (HTG) as a data augmentation strategy. Our results show the effectiveness of modular adaptation and synthetic data in improving HTR performance, achieving reductions in character error rate (CER) by 13.33% to 35.65% and word error rate (WER) by 8.56% to 27.72%.

1 INTRODUCTION

Text recognition is an ongoing research challenge in the context of document digitization, aiming to extract textual content from real-world and visually complex scanned documents. It encompasses a variety of tasks, ranging from printed text recognition to scene and handwritten text recognition. While printed text recognition typically involves clean layouts and consistent fonts in scanned documents, the other subtasks present greater challenges. Scene text recognition must handle complex backgrounds, varying lighting conditions, distortions, and font diversity (Xu et al., 2024; Du et al., 2025; Zhao et al., 2024), whereas the variability in individual handwriting styles, slant, and spacing complicates handwritten text recognition (HTR) (Li et al., 2023; 2025; Gu et al., 2025). Additionally, alternative approaches for text recognition address specialized content such as mathematical expressions and tables (Zheng et al., 2021; Kishor et al., 2023; Wan et al., 2024; Loitongbam & Middleton, 2025).

In this paper, we focus on HTR as a means of digitizing historical manuscripts. These manuscripts, belonging to a well-known European library, are written in Latin and date from the 16th to 18th centuries. They exhibit a wide range of handwriting styles and cover diverse domains, from logic to physics[1].

HTR is usually implemented as a two-step process: first, detecting individual text lines, hereafter referred to as line images, and then recognizing the text within each line (Figure 3 in the Appendix shows an example of how scanned manuscript pages are segmented into line images). An alternative is provided by end-to-end methods that unify line detection and text recognition (Wigington et al., 2018; Mao et al., 2024; Hamdi et al., 2025). However, our preliminary experiments showed that, in our case, these methods tend to be more error-prone and more difficult to interpret and debug than the standard two-step pipeline. Therefore, in this work, we adopt the standard two-step approach. While the first step can be effectively addressed by fine-tuning an object recognition model to detect line objects, the second step proves more challenging due to the scarcity of training data for low-resource Latin scripts and the highly variable handwriting styles found in manuscripts from past

[1]To preserve anonymity, we will provide more details about the collection of manuscripts upon acceptance.

centuries. For this reason, the present work focuses exclusively on the second step of the standard HTR pipeline, hereinafter called line-level HTR.

To recognize the text within line images, we fine-tune TrOCR (Li et al., 2023), a well-established encoder-decoder model pretrained for HTR in English, on a Latin dataset. This transfer learning approach is commonly used for languages with fewer HTR resources than English (Ströbel et al., 2022; Lauar & Laurent, 2024), as it facilitates the learning process by transferring the HTR knowledge encoded in the English-trained TrOCR model.

Since the simple fine-tuning is not sufficient, we imagine the transition from a modern English TrOCR to a historical Latin TrOCR[2] as a linear equation. This equation is realized by integrating two adapter modules into the TrOCR architecture, resulting in the proposed AdapterTrOCR model. The first module performs historical language adaptation, transforming the representations learned from English into a space suitable for historical Latin handwritten text. The second module focuses on style adaptation, allowing the model to adjust to the specific handwriting styles found in the manuscripts. Both modules are trained on dedicated datasets and then integrated into the AdapterTrOCR model, which is subsequently fine-tuned on the Latin corpus. Although AdapterTrOCR is designed for Latin, its modular architecture makes it easily adaptable to other languages.

Due to limited data availability, we also explore handwritten text generation (HTG) as a form of data augmentation for specific handwriting styles. While HTG typically performs well on handwriting styles that are well represented in the training data and where data augmentation is less critical, we propose a solution that also benefits underrepresented handwritten styles.

The contributions of our work are summarized as follows:

1. We develop AdapterTrOCR, a new model for historical Latin HTR by decomposing the components needed to transition from a modern English HTR model to one tailored for historical Latin manuscripts handwritten in specific handwriting styles.
2. We propose DiffLine, a new HTG-based data augmentation method, and demonstrate its effectiveness for HTR, particularly in the case of underrepresented handwriting styles.

The remainder of the paper is organized as follows. Section 2 reviews related work on HTG and HTR. Section 3 introduces our methodology, including AdapterTrOCR and DiffLine, which are evaluated in Section 4. Finally, Section 5 presents our conclusions, limitations, and directions for future research.

2 Related work

Handwritten Text Generation. As in many image generation tasks, adversarial training has made a significant contribution to the generation of photorealistic images containing handwritten text. Generative Adversarial Network (GAN)-based methods range from using only textual input for conditioning (Alonso et al., 2019; Fogel et al., 2020; Zdenek & Nakayama, 2021) to incorporating both text and handwriting style as conditioning signals (Kang et al., 2020; 2021; Mattick et al., 2021; Bhunia et al., 2021; Pippi et al., 2023; Gan et al., 2022; Wang et al., 2025; Hoai Nam et al., 2025). More recently, diffusion models have become dominant in this domain due to their ability to produce higher-quality image samples than GANs (Dhariwal & Nichol, 2021).

While diffusion-based models for HTG typically rely on a U-Net architecture (Ronneberger et al., 2015) for denoising, they mainly differ in how the text and style conditions are encoded and integrated into the diffusion process. Zhu et al. (2023), Mayr et al. (2024) and Dai et al. (2024) use a transformer decoder to merge the style and text embeddings into a single representation, which is then provided to the U-Net as a unified conditioning signal. Gui et al. (2023) propose an HTG model that follows the InstructPix2Pix framework (Brooks et al., 2023), where the text condition is represented as a glyph image displayed in a standard font and concatenated with the input image, while the style embedding guides the U-Net's noise prediction. This approach is further extended

[2]In the context of this paper, Historical Latin TrOCR refers to a TrOCR model trained to recognize Latin texts in historical manuscripts dating from the 16th to 18th centuries.

by Ding et al. (2023) by introducing a filtering module that discards synthetic text images with low HTR scores.

Considering that style is a global property affecting the entire image and text is a sequential and spatial signal, WordStylist (Nikolaidou et al., 2023) injects the style embedding into the U-Net by summing it with the timestep embedding, while the text condition is incorporated via cross-attention. The method is further refined in DiffusionPen (Nikolaidou et al., 2024), which employs a CANINE-C text encoder (Clark et al., 2022) and a MobileNetV2 (Sandler et al., 2018) for style encoding.

A recent alternative to adversarial and diffusion-based models is presented by Pippi et al. (2025), where the authors propose an autoregressive transformer-based approach. The method reconstructs input text images without background, aiming to enhance the clarity and quality of text rendering in the generated outputs.

Handwritten Text Recognition. Text recognizers for line images typically rely on convolutional neural networks (CNNs) to learn spatial patterns (Puigcerver, 2017; Shi et al., 2017; Puigcerver, 2017; Wigington et al., 2018; Ahlawat et al., 2020; Yousef & Bishop, 2020; Chaudhary & Bali, 2022; Coquenet et al., 2021), or incorporate attention mechanisms such as transformer-based blocks (Wang et al., 2020; Kang et al., 2022; Li et al., 2022).

More recently, Li et al. (2023) proposed TrOCR, an encoder-decoder architecture in which the encoder is based on the BEiT model (Bao et al., 2021), and the decoder is initialized with the weights of a RoBERTa model (Liu et al., 2019). Rather than using the full Transformer-based encoder-decoder structure, Li et al. (2025) employ only the Transformer encoder, initialized with the weights of a Vision Transformer (ViT) (Dosovitskiy et al., 2021), for text recognition within the line images. This approach includes a convolutional-based feature extractor and utilizes the Sharpness-Aware Minimization (SAM) optimizer (Foret et al., 2021). Alternatively, Fujitake (2024) propose an HTR model based on a Transformer decoder initialized with a GPT model (Radford et al., 2019), while image patches are represented using the patch embedding technique described by Dosovitskiy et al. (2021).

Unlike the aforementioned models, which adopt writer-independent approaches, Wang & Du (2022) embed handwriting style into a vector representation and integrate it into a CNN model to enhance text recognition performance. Another writer-specific personalization method is presented by Gu et al. (2025), where learnable writer-specific vectors are combined with input line images through spatial concatenation or padding. While this approach is similar to our proposed HTR model, MetaWriter (Gu et al., 2025) applies personalization only at the level of the convolutional layer due to the limitations of the padding-based implementation. In contrast, our style adapters can be seamlessly integrated throughout the entire network, allowing for a deeper and more holistic influence of the handwriting style on the HTR model.

3 Proposed methodology

We begin by introducing the proposed AdapterTrOCR model, followed by a description of the HTG-based data augmentation strategy used to expand the training data for AdapterTrOCR.

3.1 AdapterTrOCR for Handwritten Text Recognition

The HTR model we propose is based on the TrOCR architecture (Li et al., 2023), which employs an encoder-decoder design. As TrOCR is pretrained for English handwriting recognition on modern datasets, we adapt it to handle historical Latin manuscripts written in distinct handwriting styles. To this end, we introduce AdapterTrOCR, which incorporates two modules that independently adapt TrOCR to historical Latin HTR and to the recognition of text in a specific handwriting style.

The adaptation is carried out in two steps (Fig. 1a). First, we train the two adapters on specific datasets and tasks, which will be described in detail below. These adapters utilize only the decoder component of TrOCR, as some of the training tasks involve only the language modality, which is handled exclusively by the decoder. In the second step, we integrate the trained adapter weights into the full TrOCR architecture and fine-tune the entire model on a Latin-based dataset.

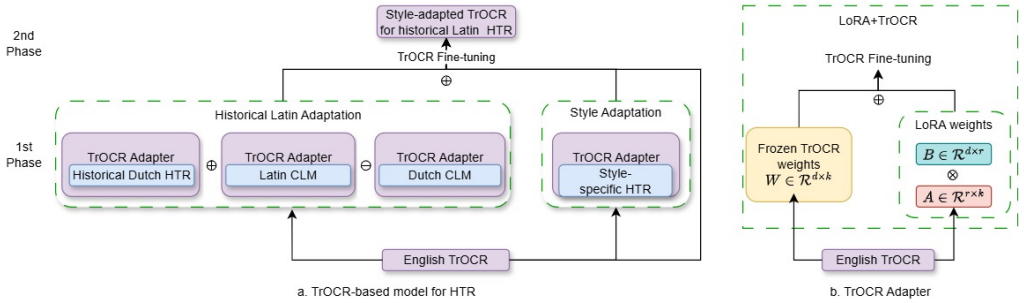

Figure 1: The historical Latin and style adaptations of the TrOCR model for HTR (a) and the TrOCR adapter implemented with LoRA (b).

Given a TrOCR layer $\boldsymbol{h}_1 = \boldsymbol{W}\boldsymbol{h}_0$, where $\boldsymbol{h}_0 \in \mathbb{R}^k$ is the input, $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ is the initial TrOCR weight, and $\boldsymbol{h}_1 \in \mathbb{R}^d$ is the output of the layer, the adapters are trained to learn new weights $\boldsymbol{W}' \in \mathbb{R}^{d\times k}$ that are integrated into the layer as follows: $\boldsymbol{h}_1 = (\boldsymbol{W} \oplus \boldsymbol{W}')\boldsymbol{h}_0$, where $\oplus$ represents the element-wise addition. To reduce the training overhead of the adapters, we obtain the matrix $\boldsymbol{W}'$ using a low-rank decomposition implemented via LoRA (Hu et al., 2022). The layer is adapted as follows:

$$\boldsymbol{h}_1 = (\boldsymbol{W} \oplus \boldsymbol{W}')\boldsymbol{h}_0 = (\boldsymbol{W} \oplus \boldsymbol{B}\boldsymbol{A})\boldsymbol{h}_0 \tag{1}$$

where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ are the low-rank decomposition matrices of $\boldsymbol{W}'$ and $r \ll min(d, k)$. During the training of the adapters, all TrOCR parameters are frozen except for the $\boldsymbol{B}$ and $\boldsymbol{A}$ weights (Fig. 1b). After the adapters' training, the LoRA weights of the historical Latin adapter $\boldsymbol{W}'_{hl}$ and the LoRA weights of the style adapter $\boldsymbol{W}'_{s}$ are integrated into a TrOCR layer as follows:

$$\boldsymbol{h}_1 = (\boldsymbol{W} \oplus \lambda_{hl}\boldsymbol{W}'_{hl} \oplus \lambda_s\boldsymbol{W}'_s)\boldsymbol{h}_0 \tag{2}$$

This formulation relies on the compositionality rules proposed by Zhang et al. (2023), which enable a model to transition from a task to another one by performing arithmetic operations on the parameters of adapters trained for individual tasks. All TrOCR parameters, including the weights of the adapters, are subsequently fine-tuned on a Latin-specific dataset.

Historical Latin adaptation To define the historical Latin adaptation, we require the TrOCR model to distinguish between "task ability" and "language ability". "Task ability" refers to the model's capacity to perform historical HTR using a proxy language, while "language ability" denotes the adaptation from the proxy language to Latin. This approach involves first preparing the model to handle historical handwriting in the proxy language ("task ability"), and then removing the difference between this language and the target language, which in our case is Latin.

To define the "task ability", we build an adapter that redirects TrOCR from modern HTR—obtained through pretraining on modern English handwriting datasets, as discussed by Li et al. (2023)—to the task of historical HTR. This adapter is a TrOCR-based decoder trained as a parameter-efficient module (PEM) using LoRA on the *VOC and notarial deeds dataset* (Keijser, 2024). The choice of this dataset is motivated by its inclusion of manuscripts from a similar time period as those in our collection of manuscripts (from the 16th to 18th century). The proxy language is Dutch, as it is the language used in the VOC and notarial deeds manuscripts.

To extract the "language ability", we rely on the assumption that the difference between two languages can be captured by the difference between the weights of a model trained on a proxy task in one language and the weights of a model trained on the same task in the second language (Zhao et al., 2025; Ansell et al., 2022; Zhang et al., 2025). Based on this assumption, we define two adapters—both TrOCR-based decoders—and train them separately as PEMs using LoRA on Dutch and Latin. Since there is no restriction on the choice of the proxy task (Zhao et al., 2025; Ansell et al., 2022; Zhang et al., 2025), we employ a self-supervised task, such as causal language modeling (CLM), to train the adapters.

Once the adapters are trained, we rely again on the compositionality rules discussed by Zhang et al. (2023) and define the LoRA weights $\boldsymbol{W}'_{hl}$ as:

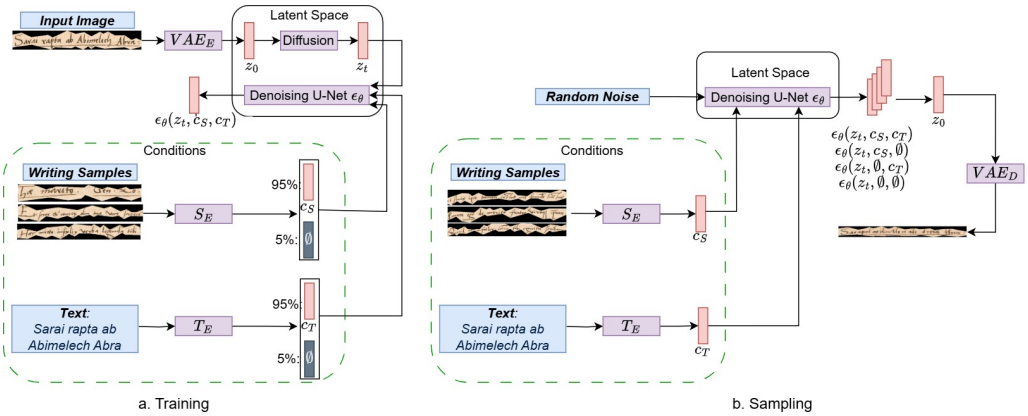

Figure 2: Training (a) and sampling (b) procedures of the proposed model DiffLine for HTG.

$$\boldsymbol{W}'_{hl} = \boldsymbol{W}'_h \oplus \lambda_l(\boldsymbol{W}'_l \ominus \boldsymbol{W}'_d) \tag{3}$$

where $\boldsymbol{W}'_h$ denotes the LoRA weight learned by the task adapter for historical HTR, $\boldsymbol{W}'_d$ and $\boldsymbol{W}'_l$ represent the LoRA weights associated with the Dutch and Latin language adapters, respectively, and $\ominus$ is the element-wise subtraction.

Style adaptation Similar to historical Latin adaptation, style adaptation is achieved by training a TrOCR-based decoder as a PEM on a subset of images containing text written in a specific handwriting style (extracted from the same manuscript). In this work, we assume that all line images extracted from a given manuscript are written by a single author and therefore exhibit a consistent handwriting style. To obtain sufficient data for training the style adapter, we augment the real annotated line images from a manuscript with synthetic line images generated in the corresponding handwriting style. This step is particularly important for underrepresented manuscripts that do not contain sufficient annotated line images in the training data. After training the style adapter, the LoRA weight $\boldsymbol{W}'_s$ is integrated into the TrOCR architecture, as indicated in Eq. 2.

3.2 Handwritten Text Generation for Data Augmentation

Since the annotated data is limited, we propose an HTG module to augment our dataset. The goal of integrating HTG into our framework is to enrich small subsets of line images handwritten in a specific style by generating synthetic lines that preserve the same handwriting characteristics. These synthetic line images are used to augment the training data for both AdapterTrOCR and the style adapter.

To achieve HTG-based data augmentation, we define a diffusion-based model capable of generating synthetic data conditioned on both style and text. The proposed model, which we call DiffLine, is based on DiffusionPen (Nikolaidou et al., 2024), the current state-of-the-art in HTG. As previously noted, DiffusionPen, built on top of WordStylist (Nikolaidou et al., 2023), has the advantage of treating handwriting style as a global condition that influences the entire line image, while modeling text as a spatial condition. Building upon DiffusionPen, we integrate dual classifier-free guidance for both conditions and enhance the style encoder by employing a more robust training regime than the one originally used in DiffusionPen.

Diffusion Models with Dual Classifier-free Guidance for Style and Text Conditions Diffusion models are a class of generative models that produce data by reversing a gradual noising process. At each denoising step, a U-Net model ϵ_θ is employed to predict the added noise. This prediction is then used to iteratively reconstruct a clean sample from pure noise, conditioned on both style and text, represented by the embeddings $\boldsymbol{c}_S$ and $\boldsymbol{c}_T$, respectively.

Our U-Net backbone architecture follows that of DiffusionPen and WordStylist, on top of which we integrate classifier-free guidance for two conditions (Brooks et al., 2023). The text condition is

fed into the cross-attention layers of the U-Net, while the style embedding is concatenated with the timestep embedding, which informs the model of the noise level in the input. By incorporating the style embedding in this way, every ResNet block in the U-Net has direct access to it, allowing the style embedding to influence the entire image generation process.

Given a line image encoded by a Variational Autoencoder (VAE) (Rombach et al., 2021) as $\boldsymbol{z}_t$, along with the two conditions $\boldsymbol{c}_S$ and $\boldsymbol{c}_T$, the diffusion model is trained to predict the Gaussian noise ϵ added at each timestep during the forward diffusion process:

$$L = ||\epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_S, \boldsymbol{c}_T)|| \tag{4}$$

To implement the classifier-free guidance, our model should support both conditional and unconditional denoising with respect to the two conditions. To implement this, we separately cancel text and style conditions for 5% of the training instances. Once the model is trained, we sample new synthetic images using the following adaptation of the noise predicted by the U-Net model ϵ_θ:

$$\epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_S, \boldsymbol{c}_T) = \epsilon_\theta(\boldsymbol{z}_t, \emptyset, \emptyset) + s_T \cdot (\epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_T, \emptyset) - \epsilon_\theta(\boldsymbol{z}_t, \emptyset, \emptyset)) + s_S \cdot (\epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_T, \boldsymbol{c}_S) - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_T, \emptyset)) \tag{5}$$

where s_T and s_S are guidance scales for the text and style conditions. Sampling process illustrated in Eq. 5 assumes that the synthetic images should first be generated using the text condition and then adapted to the handwriting style indicated by the style condition. In this way, we give higher priority to the text condition than to the style constraint. The hyperparameters s_T and s_S are selected by maximizing the cosine similarity between the style embeddings of the generated images and the style embeddings of 20 randomly selected images displaying handwriting from a specific writer. The training and sampling with DiffLine is illustrated in Fig. 2.

Style encoder To implement the style encoder, we use the MobileNetV2 model (Sandler et al., 2018). Following a similar approach to (Nikolaidou et al., 2024), we train MobileNetV2 to capture the stylistic characteristics of handwriting by contrastively learning the differences between handwriting styles. The style encoder of DiffusionPen is trained using a triplet loss, which is sensitive to the selection of negative samples (i.e., line images written by different authors with a different handwriting style). Since negative samples can vary significantly in their similarity to the anchor line image, the triplet loss may lead to an inconsistent training process, making it difficult to establish a standardized contrastive learning framework.

To address the above issue, we adopt a softmax-based contrastive loss (Chen et al., 2020), which encourages higher similarity between an anchor line image and a randomly selected positive line image, i.e., one written in the same handwriting style and included in the same batch. The similarity score is normalized over the similarity scores between the anchor and all other line images in the batch. By normalizing across the entire batch, this approach eliminates the need for explicit hard negative mining, as a sufficiently large batch is expected to contain negative samples of varying difficulty.

Knowing that MobileNetV2 encodes the anchor line image x into f_x, f_{pos} is a positive image for the anchor image x, the batch size is N and $sim(*)$ stands for cosine similarity, we define the softmax-based contrastive loss as follows:

$$L_{contrastive}(f_x, f_{pos}) = \frac{exp(sim(f_x, f_{pos}))}{\sum_{k=1}^{N} exp(sim(f_x, f_k))} \tag{6}$$

While the writing particularities are important to allocate a line image to a manuscript, we also need to preserve a certain level of generalization that is required to recognize letters regardless of the writer. For example, while different people write the letter "a" differently, any reader should still be able to recognize the letter. As the contrastive loss might be inclined to group positive embeddings into tight clusters, which might affect this generalization ability of the style encoder, a Sinkhorn-based loss (Sepanj & Fiegth, 2025) is defined to regularize the loss $L_{contrastive}$.

Given $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ as the similarity matrix between all line images of a certain batch, $\boldsymbol{T}$ as a matrix that indicates the transport plan obtained by applying the Sinkhorn–Knopp algorithm (Cuturi, 2013) on $\exp(\boldsymbol{S})$ and $\boldsymbol{U} \in \mathbb{R}^{N \times N}$ as a uniform matrix where $U_{i,j} = 1/N$, we compute the Sinkhorn-based loss using the Kullback–Leibler divergence between $\boldsymbol{T}$ and $\boldsymbol{U}$ scaled by the weight $\lambda_{Sinkhorn}$. The final contrastive loss $L_{contrastive}$ is defined as follows:

$$L_{contrastive}(f_x, f_{pos}, T, U) = \frac{exp(sim(f_x, f_{pos}))}{\sum_{k=1}^{N} exp(sim(f_x, f_k))} + \lambda_{Sinkhorn} D_{\mathrm{KL}}(\boldsymbol{T}||\boldsymbol{U}) \tag{7}$$

4 Experiments

4.1 Experimental setup

Data To evaluate AdapterTrOCR, we use a collection of five Latin manuscripts written between the 16th and 18th centuries by five different writers. Four of these manuscripts come from our internal collection, with transcriptions prepared by colleagues skilled in Latin and paleography. The fifth manuscript is the *Lateinische Gedichte* volume by Rudolf Gwalther (Stotz & Ströbel, 2021). The number of line transcriptions per manuscript ranges from 866 to 4037. The complete dataset contains about 10K line images, with each transcription averaging 8 tokens and 58.14 characters per line. The full dataset will be released upon acceptance. For the DiffLine training, we exclude the well-known Bullinger dataset (Hodel et al., 2023) because it does not include information about the writers. For the training of AdapterTrOCR, we also exclude the Bullinger dataset as additional training data since it does not provide significant improvements, probably due to the difference in layout and style when compared with our data. More details can be found in Appendix A.1.

Metrics To evaluate the HTR task we rely on the HTR-specific metrics like character error rate (CER) and word error rate (WER). Additionally, we include accuracy (Acc) to measure the ability of the models to generate a transcription identical to the ground truth.

Models for data augmentation with handwritten text generation We compare our model, DiffLine for HTG with the following diffusion-based baselines: One-DM (Dai et al., 2024), DiffusionPen (Nikolaidou et al., 2024) and WordStylist (Nikolaidou et al., 2023)[3] Additional baseline models that we consider are VATr (Pippi et al., 2023) and HWT (Bhunia et al., 2021).

Models for line-level HTR As our model is built on top of TrOCR, which was a state-of-the-art line-level HTR model in 2023, we consider as baselines only models proposed after that year with publicly available code. Therefore, besides TrOCR, we include HTR-VT (Li et al., 2025) and ViTLP (Mao et al., 2024) as additional baselines. We also report results for PyLaia (Puigcerver, 2017), which is the underlying model used for line-level HTR in Transkribus[4], a widely used platform in the digital humanities community. While HTR-VT and PyLaia are line-level HTR models, ViTLP is an end-to-end model. To enable a fair comparison with ViTLP, we run our proposed AdapterTrOCR only on the lines detected by a fine-tuned YOLO model, considering only those with a confidence score above 70%. More details on the line detection process are provided in Appendix A.2.

Implementation details To evaluate the effect of the number of annotated line images at the manuscript level, we work with two scenarios. The first scenario evaluates HTR for the manuscript with the smallest number of transcriptions (866), while the second scenario evaluates HTR for the manuscript with the largest number of transcriptions (4037).

We begin by training DiffLine and the HTG baselines on our data collection to enable manuscript-specific data augmentation. Since DiffLine outperforms the baselines (see Section 4.2), we use it to generate 2000 synthetic line images reflecting the handwriting style of the manuscript of each scenario. In total, we generate two sets of 2000 line images, one for each scenario. We use the same arbitrary Latin text to generate the synthetic images of the two scenarios. Further details on the choice to use 2000 synthetic line images per scenario are provided in Appendix A.3.

For each scenario, we define the test set by randomly selecting 300 ground-truth instances from the corresponding manuscript. The training data for AdapterTrOCR in a given scenario consists

[3] Another recently proposed HTG model is Emuru (Pippi et al., 2025). We do not use this model as a baseline due to the lack of the complete code for training/inference.

[4] https://www.transkribus.org/

of: the remaining instances from that manuscript, 2000 synthetic instances generated to capture its handwriting style, and the ground-truth instances from the other four manuscripts. As mentioned above, the 2000 synthetic images are also used to augment the manuscript-specific instances when training the style adapters. Further details on the training procedure and hyperparameter selection for DiffLine and AdapterTrOCR are provided in Appendix A.4.

Table 1: Comparison between DiffLine and the baseline methods for HTG-based data augmentation (DA) in terms of HTR performance. The HTR results were generated using TrOCR, fine-tuned on the Latin training data specified in the Setup column. GT and Syn- indicate the inclusion of ground-truth and synthetic training data, respectively, for the manuscript associated with each scenario. The DA Method refers to the method used to generate the synthetic data. While CER and WER are standard metrics for evaluating HTR performance, Acc represents the accuracy, measured as the percentage of transcriptions that are counted correct only when the entire line exactly matches the ground-truth text.

		1st Scenario - underrepresented manuscripts			2nd Scenario - well-represented manuscript		
Setup	DA Method	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$
No GT - No Syn	-	**2.33**	46.86	91.17	0.33	17.54	54.05
No GT - Yes Syn	HWT	1.23	63.23	98.86	2.12	19.19	51.55
No GT - Yes Syn	VATr	1.42	61.87	97.43	3.32	19.48	50.45
No GT - Yes Syn	One-DM	1.87	55.43	93.98	6.35	17.97	47.75
No GT - Yes Syn	WordStylist	1.66	58.09	96.32	2.66	18.06	47.68
No GT - Yes Syn	DiffusionPen	2.00	49.46	92.91	8.67	16.03	49.87
No GT - Yes Syn	DiffLine	1.67	**46.57**	**89.27**	**12.04**	**11.44**	**31.86**
Yes GT - No Syn	-	**5.00**	23.33	60.73	34.11	4.74	14.50
Yes GT - Yes Syn	HWT	4.12	25.45	62.45	34.52	5.12	14.98
Yes GT - Yes Syn	VATr	4.23	25.04	62.47	34.45	4.65	14.98
Yes GT - Yes Syn	One-DM	4.89	23.82	61.34	36.78	4.47	14.65
Yes GT - Yes Syn	WordStylist	**5.00**	24.06	61.58	36.12	**4.07**	14.14
Yes GT - Yes Syn	DiffusionPen	4.00	23.72	61.27	35.45	4.34	14.19
Yes GT - Yes Syn	DiffLine	4.33	**23.32**	**60.03**	**37.79**	4.47	**13.59**

4.2 RESULTS

Data augmentation with Handwritten Text Generation We begin by comparing DiffLine with the baseline data augmentation methods in the context of historical Latin HTR. In addition to the scenario where synthetic data is used to extend the training set, we also evaluate a setting in which, for a given manuscript, only synthetic data is used, without any ground-truth annotations. This latter evaluation is important for assessing the standalone quality of the synthetic data. As for this phase, we only want to select the best method for data augmentation, the evaluation is done using only TrOCR fine-tuned on our Latin collection without any task, language and style adaptation. By comparing the proposed DiffLine approach with the baselines (Table1), we observe that, overall, DiffLine produces the most effective synthetic images for improving HTR performance. In contrast, all other HTG baselines appear to degrade performance, increasing both CER and WER. A few synthetic image lines generated by DiffLine and the baselines are presented in Appendix A.5.

Interestingly, Table 1 also shows that generating high-quality synthetic data capable of significantly reducing CER and WER requires a sufficient amount of annotated data for the given manuscript, an observation that reduces the usefulness of synthetic data. In the second scenario, which corresponds to a high-resource manuscript, we observe that the synthetic line images generated by DiffLine are of high quality. When used alone (i.e., without the manuscript-specific real data), they reduce the CER by 34.77% and the WER by 41.05%. However, when both synthetic and real data are used together, the improvements remain significant but are smaller, 5.69% for CER and 6.27% for WER.

On the other hand, when targeting a manuscript with limited training data (first scenario), DiffLine has less information to learn from, resulting in lower-quality synthetic images and smaller reductions in CER and WER. Specifically, when both the real and synthetic data are used to train TrOCR, CER is reduced by only 0.04% and WER by 1.15%. However, for all other HTG baselines in this low-resource setting, both CER and WER increase.

Handwritten text recognition To evaluate AdapterTrOCR in each scenario, we train the model on datasets augmented with synthetic data generated by DiffLine, using the handwriting style of

Table 2: Comparison between the proposed AdapterTrOCR and the line-level HTR models TrOCR, HTR-VT and PyLaia. For the comparison with the end-to-end HTR model ViTLP, AdapterTrOCR is applied only to the lines detected by the fine-tuned YOLO model with a confidence score greater than 70% (see Appendix A.2).

	1st Scenario - underrepresented manuscripts			2nd Scenario - well-represented manuscript		
Method	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$
ViTLP	3.53	27.43	64.63	**40.02**	6.22	17.43
AdapterTrOCR	**3.89**	**25.53**	**61.35**	39.54	**5.34**	**15.42**
HTR-VT	3.87	25.35	65.25	33.63	10.53	17.46
PyLaia	1.34	29.40	67.54	26.34	12.42	20.07
TrOCR	4.33	23.32	60.03	37.79	4.47	13.59
AdapterTrOCR	**4.66**	**20.22**	**55.53**	**47.49**	**3.05**	**10.48**

Table 3: Ablation results for AdapterTrOCR based on the style and historical Latin adaptations. Additionally, we include the ablation when the style adapter is trained without the synthetic data - third line. All models are trained on the training data augmented with the synthetic data of the associated scenario.

	1st Scenario - underrepresented manuscript			2nd Scenario - well-represented manuscript		
Method	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$	$Acc(\uparrow)$	$CER(\downarrow)$	$WER(\downarrow)$
TrOCR	4.33	23.32	60.03	37.79	4.47	13.59
TrOCR + language adapter	4.00	22.73	60.34	40.8	4.15	12.88
TrOCR + style adapter (without synthetic data)	4.00	22.99	59.12	**48.55**	3.16	10.55
TrOCR + style adapter	4.00	20.78	57.71	48.34	3.15	10.67
AdapterTrOCR (style and language adapter)	**4.66**	**20.22**	**55.53**	47.49	**3.05**	**10.48**

the manuscript associated with the respective scenario. The results reported in Table 2 confirm that the proposed AdapterTrOCR outperforms all other HTR baselines in both line-level and end-to-end HTR tasks. Notably, TrOCR consistently ranks as the next best-performing model after AdapterTrOCR, supporting our decision to adopt it as the foundation for our modular adaptation.

In Table 3, we observe that incorporating both historical Latin and style adaptations increases accuracy by 7.62–25.66%, while reducing CER by 13.29-31.76% and WER by 7.49–22.88%. Among the two, adapting to the handwriting style of the manuscript proves to be the most significant factor for achieving strong HTR performance. In the second scenario (rich in ground-truth data), augmentation of the data used to train the style adapter generates minimal gains. However, in the first scenario, where the manuscript has limited annotations, the use of synthetic data becomes essential, reducing CER and WER by 6.77% and 2.38%, respectively.

5 Conclusion, Limitations and Future Work

Conclusions. In this work, we presented AdapterTrOCR, a modular extension of the TrOCR architecture for recognizing historical Latin handwritten texts. The model introduces two adapter modules: one for historical language adaptation and another for handwriting style adaptation. This design enables effective transfer from a modern English HTR model to historical Latin manuscripts. To mitigate the scarcity of annotated data, we complemented this approach with an HTG model for producing synthetic line images that mimic manuscript-specific handwriting styles. Together, modular adaptation and synthetic data substantially improved recognition accuracy, particularly for underrepresented manuscripts. Moreover, the proposed framework is flexible and can be extended to other languages, making it a valuable tool for cultural heritage preservation.

Limitations and future work. As training data plays an essential role in generating accurate HTR results, a straightforward research direction is to improve the quality of synthetic images. Our DiffLine model enhances recognition performance without polluting the training data, but it still struggles with displaying the text correctly. Given this, a promising direction for future work is to address the degradation in text accuracy observed in synthetic line images, where the left side is generally more accurate than the right, likely due to weaker alignment between characters and pixels as the text progresses.

6 Ethics Statement

This work develops an HTR model for recognizing Latin text in digitized historical documents from the 16th–18th centuries. Our data does not contain personal or sensitive information about living individuals, and follows the terms set by the holding institutions. This research aims to support the preservation of cultural heritage and improve scholarly access, with no expected harmful use cases.

7 Reproducibility statement

Implementation details of our models and experiments are described in Sections 4.1, with further information about the training setup in the Appendix A.3. Upon acceptance, we will release the full source code, trained models, and datasets used in our experiments to ensure reproducibility.

References

Savita Ahlawat, Amit Choudhary, Anand Nayyar, Saurabh Singh, and Byungun Yoon. Improved handwritten digit recognition using convolutional neural networks (CNN). *Sensors*, 20(12):3344, 2020. doi: 10.3390/S20123344. URL https://doi.org/10.3390/s20123344.

Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial generation of handwritten text images conditioned on sequences. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 481–486. IEEE, 2019.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulic. Composable sparse fine-tuning for cross-lingual transfer. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 1778–1796. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.125. URL https://doi.org/10.18653/v1/2022.acl-long.125.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1086–1094, 2021.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18392–18402. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01764.

Kartik Chaudhary and Raghav Bali. Easter2.0: Improving convolutional models for handwritten text recognition. *CoRR*, abs/2205.14879, 2022. doi: 10.48550/ARXIV.2205.14879. URL https://doi.org/10.48550/arXiv.2205.14879.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. SPAN: A simple predict & align network for handwritten paragraph recognition. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part III*, volume 12823 of *Lecture Notes in Computer Science*, pp. 70–84. Springer, 2021. doi: 10.1007/978-3-030-86334-0_5. URL https://doi.org/10.1007/978-3-030-86334-0_5.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013.

Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVIII*, volume 15116 of *Lecture Notes in Computer Science*, pp. 410–427. Springer, 2024. doi: 10.1007/978-3-031-73636-0_24. URL https://doi.org/10.1007/978-3-031-73636-0_24.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Haisong Ding, Bozhi Luan, Dongnan Gui, Kai Chen, and Qiang Huo. Improving handwritten OCR with training samples generated by glyph conditional denoising diffusion probabilistic model. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part IV*, volume 14190 of *Lecture Notes in Computer Science*, pp. 20–37. Springer, 2023. doi: 10.1007/978-3-031-41685-9_2. URL https://doi.org/10.1007/978-3-031-41685-9_2.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yongkun Du, Zhineng Chen, Yuchen Su, Caiyan Jia, and Yu-Gang Jiang. Instruction-guided scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(4):2723–2738, 2025. doi: 10.1109/TPAMI.2025.3525526. URL https://doi.org/10.1109/TPAMI.2025.3525526.

Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. Scrabblegan: Semi-supervised varying length handwritten text generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4324–4333, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 8010–8020. IEEE, 2024. doi: 10.1109/WACV57701.2024.00784. URL https://doi.org/10.1109/WACV57701.2024.00784.

Ji Gan, Weiqiang Wang, Jiaxu Leng, and Xinbo Gao. Higan+: handwriting imitation gan with disentangled representations. *ACM Transactions on Graphics (TOG)*, 42(1):1–17, 2022.

Wenhao Gu, Li Gu, Ching Yee Suen, and Yang Wang. Metawriter: Personalized handwritten text recognition using meta-learned prompt tuning. *CoRR*, abs/2505.20513, 2025. doi: 10.48550/ARXIV.2505.20513. URL https://doi.org/10.48550/arXiv.2505.20513.

Dongnan Gui, Kai Chen, Haisong Ding, and Qiang Huo. Zero-shot generation of training data with denoising diffusion probabilistic model for handwritten chinese character recognition. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part II*, volume 14188 of *Lecture Notes in Computer Science*, pp. 348–365. Springer, 2023. URL https://doi.org/10.1007/978-3-031-41679-8_20.

Laziz Hamdi, Amine Tamasna, Pascal Boisson, and Thierry Paquet. VISTA-OCR: towards generative and interactive end to end OCR models. *CoRR*, abs/2504.03621, 2025. doi: 10.48550/ARXIV.2504.03621. URL https://doi.org/10.48550/arXiv.2504.03621.

Dang Hoai Nam, Huynh Tong Dang Khoa, and Vo Nguyen Le Duy. Writevit: Handwritten text generation with vision transformer. *arXiv e-prints*, pp. arXiv–2505, 2025.

Tobias Hodel, Phillip Benjamin Ströbel, Andreas Fischer, Anna Scius-Bertrand, Anna Janka, Jonas Widmer, Beat Wolf, Patricia Scheurer, and Martin Volk. Bullingers briefwechsel zugänglich machen: Stand der handschriftenerkennung. In *DHd2023: Open Humanities, Open Culture. Konferenzabstracts (tentative title)*, Luxembourg/Trier, Luxembourg/Germany, 2023. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Lei Kang, Pau Riba, Yaxing Wang, Marçal Rusinol, Alicia Fornés, and Mauricio Villegas. Ganwriting: content-conditioned generation of styled handwritten word images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pp. 273–289. Springer, 2020.

Lei Kang, Pau Riba, Marcal Rusinol, Alicia Fornes, and Mauricio Villegas. Content and style aware generation of text-line images for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8846–8860, 2021.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: Non-recurrent handwritten text-line recognition. *Pattern Recognit.*, 129:108766, 2022. doi: 10.1016/J.PATCOG.2022.108766. URL https://doi.org/10.1016/j.patcog.2022.108766.

Liesbeth Keijser. 6000 ground truth of voc and notarial deeds / 3.000.000 htr of voc, wic and notarial deeds. Zenodo, 2024. URL https://doi.org/10.5281/zenodo.11209325. Dataset.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Kaushal Kishor, Rohan Tyagi, Rakhi Bhati, and Bipin Kumar Rai. Develop model for recognition of handwritten equation using machine learning. In *Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022*, pp. 259–265. Springer, 2023.

Filipe Lauar and Valentin Laurent. Spanish trocr: Leveraging transfer learning for language adaptation. *CoRR*, abs/2407.06950, 2024. doi: 10.48550/ARXIV.2407.06950. URL https://doi.org/10.48550/arXiv.2407.06950.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 3530–3539. ACM, 2022. doi: 10.1145/3503161.3547911. URL https://doi.org/10.1145/3503161.3547911.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13094–13102. AAAI Press, 2023. doi: 10.1609/AAAI.V37I11.26538. URL https://doi.org/10.1609/aaai.v37i11.26538.

Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. HTR-VT: handwritten text recognition with vision transformer. *Pattern Recognit.*, 158:110967, 2025. doi: 10.1016/J.PATCOG.2024.110967. URL https://doi.org/10.1016/j.patcog.2024.110967.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Gyanendro Loitongbam and Stuart E Middleton. Tabular context-aware optical character recognition and tabular data reconstruction for historical records. *International Journal on Document Analysis and Recognition*, 2025.

Zhiming Mao, Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Visually guided generative text-layout pre-training for document intelligence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 4713–4730. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.264. URL https://doi.org/10.18653/v1/2024.naacl-long.264.

Alexander Mattick, Martin Mayr, Mathias Seuret, Andreas Maier, and Vincent Christlein. Smartpatch: improving handwritten word imitation with patch discriminators. In *International Conference on Document Analysis and Recognition*, pp. 268–283. Springer, 2021.

Martin Mayr, Marcel Dreier, Florian Kordon, Mathias Seuret, Jochen Zöllner, Fei Wu, Andreas Maier, and Vincent Christlein. Zero-shot paragraph-level handwriting imitation with latent diffusion models. *arXiv preprint arXiv:2409.00786*, 2024.

Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. Wordstylist: styled verbatim handwritten text generation with latent diffusion models. In *International Conference on Document Analysis and Recognition*, pp. 384–401. Springer, 2023.

Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Diffusionpen: Towards controlling the style of handwritten text generation. In *European Conference on Computer Vision*, pp. 417–434. Springer, 2024.

Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten text generation from visual archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22458–22467, 2023.

Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, Alessio Tonioni, and Rita Cucchiara. Zero-shot styled text image generation, but make it autoregressive. *CoRR*, abs/2503.17074, 2025. doi: 10.48550/ARXIV.2503.17074. URL https://doi.org/10.48550/arXiv.2503.17074.

Joan Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pp. 67–72. IEEE, 2017. doi: 10.1109/ICDAR.2017.20. URL https://doi.org/10.1109/ICDAR.2017.20.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4510–4520. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00474. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html.

M. Hadi Sepanj and Paul Fiegth. Sinsim: Sinkhorn-regularized simclr. *CoRR*, abs/2502.10478, 2025. doi: 10.48550/ARXIV.2502.10478. URL https://doi.org/10.48550/arXiv.2502.10478.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017. doi: 10.1109/TPAMI.2016.2646371. URL https://doi.org/10.1109/TPAMI.2016.2646371.

Peter Stotz and Phillip Ströbel. bullinger-digital/gwalther-handwriting-ground-truth: Initial release (v1.0), 2021. URL https://doi.org/10.5281/zenodo.4780947.

Phillip Benjamin Ströbel, Simon Clematide, Martin Volk, and Tobias Hodel. Transformer-based HTR for historical documents. *CoRR*, abs/2203.11008, 2022. doi: 10.48550/ARXIV.2203.11008. URL https://doi.org/10.48550/arXiv.2203.11008.

Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. OMNIPARSER: A unified framework for text spotting, key information extraction and table recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 15641–15653. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01481. URL https://doi.org/10.1109/CVPR52733.2024.01481.

Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 12216–12224. AAAI Press, 2020. doi: 10.1609/AAAI.V34I07.6903. URL https://doi.org/10.1609/aaai.v34i07.6903.

Yiming Wang, Hongxi Wei, Heng Wang, Shiwen Sun, and Chao He. Gl-gan: Perceiving and integrating global and local styles for handwritten text generation with mamba. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2434–2444, 2025.

Zi-Rui Wang and Jun Du. Fast writer adaptation with style extractor network for handwritten text recognition. *Neural Networks*, 147:42–52, 2022. doi: 10.1016/J.NEUNET.2021.12.002. URL https://doi.org/10.1016/j.neunet.2021.12.002.

Curtis Wigington, Chris Tensmeyer, Brian L. Davis, William A. Barrett, Brian L. Price, and Scott Cohen. Start, follow, read: End-to-end full-page handwriting recognition. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pp. 372–388. Springer, 2018. doi: 10.1007/978-3-030-01231-1_23. URL https://doi.org/10.1007/978-3-030-01231-1_23.

Jianjun Xu, Yuxin Wang, Hongtao Xie, and Yongdong Zhang. OTE: exploring accurate scene text recognition using one token. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 28327–28336. IEEE, 2024.

doi: 10.1109/CVPR52733.2024.02676. URL https://doi.org/10.1109/CVPR52733.2024.02676.

Mohamed Yousef and Tom E. Bishop. Origaminet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 14698–14707. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01472. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yousef_OrigamiNet_Weakly-Supervised_Segmentation-Free_One-Step_Full_Page_Text_Recognition_by_learning_CVPR_2020_paper.html.

Jan Zdenek and Hideki Nakayama. Jokergan: memory-efficient model for handwritten text generation with text line awareness. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 5655–5663, 2021.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations. *CoRR*, abs/2306.14870, 2023. doi: 10.48550/ARXIV.2306.14870. URL https://doi.org/10.48550/arXiv.2306.14870.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=NCrFA7dq8T.

Shuai Zhao, Ruijie Quan, Linchao Zhu, and Yi Yang. CLIP4STR: A simple baseline for scene text recognition with pre-trained vision-language model. *IEEE Trans. Image Process.*, 33:6893–6904, 2024. doi: 10.1109/TIP.2024.3512354. URL https://doi.org/10.1109/TIP.2024.3512354.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 9785–9800. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.493. URL https://doi.org/10.18653/v1/2025.naacl-long.493.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 697–706. IEEE, 2021. doi: 10.1109/WACV48630.2021.00074. URL https://doi.org/10.1109/WACV48630.2021.00074.

Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14235–14244. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01368. URL https://doi.org/10.1109/CVPR52729.2023.01368.