Learning to Decipher from Pixels — A Case Study of Copiale

Lei Kang
Giuseppe De Gregorio
Computer Vision Center
Universitat Autònoma de Barcelona
{lkang, gdegregorio}@cvc.uab.es

Raphaela Heil
Department of Linguistics
Stockholm University, Sweden
raphaela.heil@ling.su.se

Alicia Fornés
Computer Vision Center
Universitat Autònoma de Barcelona
afornes@cvc.uab.es

Beáta Megyesi
Department of Linguistics
Stockholm University, Sweden
beata.megyesi@ling.su.se

Abstract

Historical encrypted manuscripts require both paleographic interpretation of cipher symbols and cryptanalytic recovery of plaintext. Most existing computational workflows rely on a transcription-first paradigm, in which handwritten symbols are transcribed prior to decipherment. This intermediate step is labor-intensive, error-prone, and not always aligned with the goal of direct plaintext recovery. We propose an end-to-end, transcription-free approach that directly maps handwritten cipher images to plaintext. Using the Copiale cipher as a case study, we introduce the first text-line-level dataset pairing cipher images with German plaintext. We show that pretraining on generic handwriting data followed by cipher-specific fine-tuning substantially improves decipherment accuracy. Our results demonstrate that transcription-free image-to-plaintext decipherment is both feasible and effective for historical substitution ciphers, offering a simplified and scalable alternative to traditional pipelines. https://github.com/leitro/Decipher-from-Pixels-Copiale.

1 Introduction

Historical encrypted manuscripts pose a dual challenge at the intersection of document analysis and cryptology (Yin et al., 2019; Megyesi et al., 2020). As handwritten artifacts with encoded content, they require both visual interpretation and cryptanalytic decoding. Consequently, most computational approaches adopt a transcription-first workflow in which symbols are transcribed into machine-readable ciphertext and then analyzed to recover plaintext (Megyesi, 2020; Méndez et al., 2024). This paradigm has shaped research practices in historical cryptology and digital humanities.

Despite its success, transcription-first processing presents substantial limitations. Historical ciphers often exhibit large and irregular symbol inventories, inconsistent spacing, and idiosyncratic writing conventions, making segmentation and transcription labor-intensive and error-prone. Errors propagate into downstream decipherment and require frequent reference to manuscript images. Moreover, ciphertext is rarely the final scholarly objective; researchers primarily seek readable plaintext, raising the question of whether explicit transcription is always necessary.

Early computational work assumed reliable ciphertext and focused on cryptanalytic modeling. Aldarrab introduced a probabilistic noisy-channel framework (Aldarrab, 2017) and explored early image-based decipherment using segmentation and clustering, identifying character segmentation as a major bottleneck. Yin et al. later formulated decipherment from manuscript images as an integrated task combining visual processing and statistical cryptanalysis (Yin et al., 2019), demonstrating the feasibility of image-based approaches while retaining explicit symbol transcription. Together, these studies reflect a shift toward image-aware pipelines that nonetheless preserve staged processing.

Meanwhile, handwritten text recognition has advanced substantially through LSTM-based methods (Bluche et al., 2017), neural encoder–decoder models (Kang et al., 2018; Kang et al., 2021), transformer-based architectures (Kang

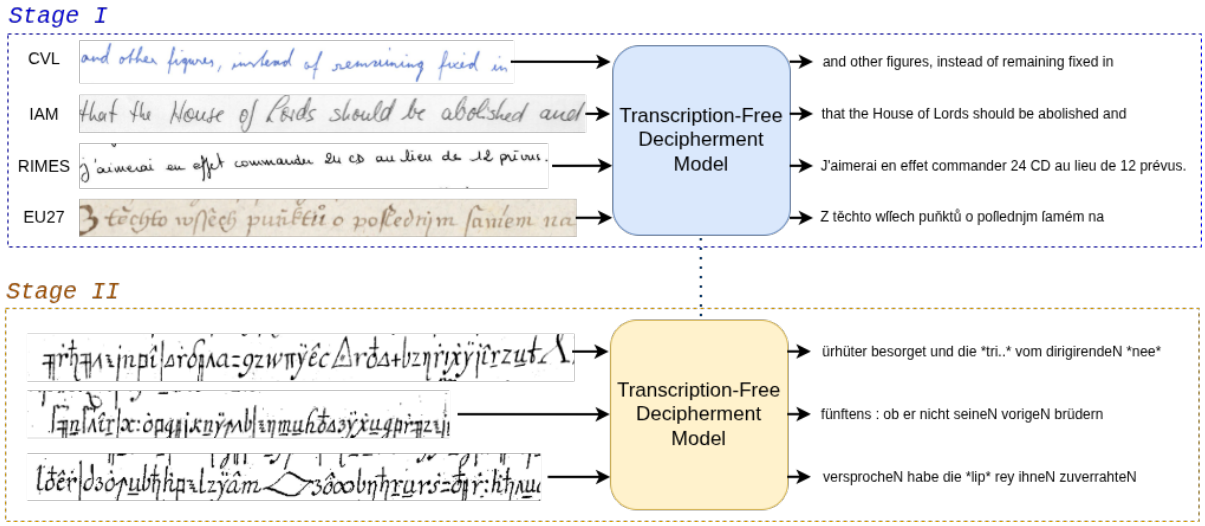


Figure 1: Overview of the training pipeline for our proposed Transcription-Free Decipherment paradigm. Stage I involves pretraining on a unified corpus of publicly available handwritten text-line datasets, followed by Stage II, where the model is fine-tuned on our curated Copiale image-to-plaintext text-line dataset.

et al., 2022), and large-scale systems such as TrOCR (Li et al., 2023). These developments enable direct image-to-text modeling and support end-to-end document understanding without explicit symbolic intermediates.

Building on these advances, we propose an end-to-end, transcription-free decipherment paradigm that directly maps handwritten cipher images to plaintext. Rather than supervising models with cipher symbols, we train them to generate decrypted natural-language text from pixel-level input. We evaluate this approach on the Copiale cipher (Knight et al., 2011; Knight et al., 2012), a well-studied eighteenth-century German manuscript.

Our contributions are threefold. First, we release the first publicly available text-line-level dataset pairing Copiale cipher images with aligned German plaintext. Second, we demonstrate the feasibility of end-to-end image-to-plaintext decipherment for historical manuscripts. Third, we show that pretraining on generic handwriting data followed by cipher-specific fine-tuning substantially improves performance. These results indicate that transcription-free pipelines provide a scalable alternative to traditional multi-stage workflows in computational historical cryptology.

2 Data

2.1 Handwriting Pretraining Data

Pretraining is conducted on a merged corpus comprising 66,492 handwritten text lines drawn from four widely used and complementary benchmarks: IAM (Marti and Bunke, 2002), CVL (Kleber et al., 2013), RIMES (Grosicki and El-Abed, 2011), and EU27 (Kohút and Hradiš, 2025). This combination is designed to maximize diversity in writing styles, languages, scripts, and acquisition conditions, thereby improving the robustness and generalization of the learned representations.

The IAM dataset provides a large collection of English handwritten text lines produced by multiple writers under controlled scanning conditions, and is commonly used as a standard benchmark for offline handwriting recognition. CVL complements IAM by offering high-resolution handwritten documents from a distinct writer population, with variations in writing instruments and stroke dynamics that enrich intra-writer and inter-writer variability. RIMES contributes French handwritten text collected in a realistic mail-processing scenario, introducing linguistic diversity as well as challenges such as cursive writing, ligatures, and noise typical of real-world document workflows. Finally, EU27 extends coverage to a multilingual European setting, incorporating handwriting samples across multiple scripts and orthographic conventions, which is particularly valuable for learn-

ing script-agnostic and language-robust features.

The combined corpus is randomly split into training, validation, and test sets using an 80/10/10 ratio, while preserving the overall distribution of datasets and handwriting styles. Detailed statistics on text-line contributions from each dataset are reported in Table 1.

Table 1: Pretraining data for Stage I.

Dataset	Textline Counts
CVL	13,440
IAM	8,873
RIMES	12,104
EU27	32,075
Total	66,492

2.2 Copiale Image-to-Plaintext Dataset

For cipher fine-tuning, we align handwritten Copiale text-line images with their corresponding German plaintext lines, producing the first dataset that supports direct image-to-plaintext decipherment of the manuscript. The number of lines, minimal and maximal number of characters and words per line for the training, validation and tests sets are described in Table 2. The entire dataset is released publicly to support future research.

Table 2: Fine-tuning data for Stage II: total number of lines, as well as the minimum and maximum number of characters and words per line, for the training, validation, and test sets of the Copiale.

Set	Count	Char Length		Word Length	
		Min	Max	Min	Max
Train	1,269	1	67	1	14
Validation	175	1	65	1	14
Test	370	1	64	1	13

3 Method

3.1 Transcription-Free Decipherment Formulation

We formulate decipherment as a sequence-to-sequence learning problem. In this paradigm, a model learns to transform one ordered sequence into another, without requiring explicit intermediate representations. Given an input image I representing a handwritten cipher text line, the model predicts a plaintext string P in natural language. No explicit ciphertext representation is produced or supervised during training.

3.2 Model Architecture

We adopt TrOCR (Li et al., 2023), a transformer-based encoder–decoder model for handwritten text recognition, and repurpose it for image-to-plaintext decipherment. A vision transformer encoder maps input images to latent visual embeddings, which are autoregressively decoded into plaintext tokens by a transformer decoder. Apart from adapting the output vocabulary, no architectural modifications are required. The overall training pipeline is shown in Figure 1.

3.3 Two-Stage Training Pipeline

We adopt a two-stage training strategy consisting of handwriting pretraining followed by cipher-specific fine-tuning. In Stage I, the model is pretrained on a unified corpus of publicly available handwritten text-line datasets (see Section 2.1), to learn robust and largely style-invariant handwriting representations. In Stage II, the pretrained model is fine-tuned on the Copiale image-to-plaintext dataset using German plaintext supervision (see Section 2.2). This strategy separates handwriting acquisition from task-specific learning: pretraining yields general visual representations independent of cipher symbols, while fine-tuning adapts these features for end-to-end decipherment.

4 Experiments

4.1 Implementation Details

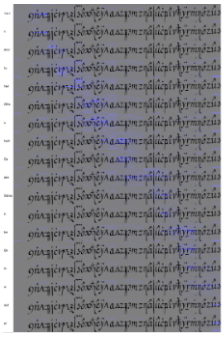
In Stage I pretraining, the model is trained for 5 epochs with a learning rate of 5×10^{-5} and a batch size of 64. For Stage II, we adopt a learning rate of 2×10^{-5} and a batch size of 64, and apply early stopping with a patience of 20 epochs based on validation performance, resulting in a total of 21 training epochs. AdamW is used as the optimizer, and the backbone model is `microsoft/trocr-base-handwritten`. All experiments are conducted on a single NVIDIA 4090 GPU.

4.2 Evaluation Metrics

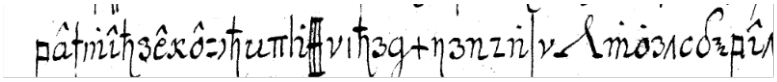
We report Character Error Rate (CER) and Word Error Rate (WER), which are standard metrics for handwriting recognition and sequence prediction. CER is defined as the normalized Levenshtein

Table 3: End-to-end Copiale cipher decoding performance evaluated with CER (%) and WER (%), where lower is better.

Method	Stage I	Stage II	Train Set		Validation Set		Test Set	
			CER↓	WER↓	CER↓	WER↓	CER↓	WER↓
Baseline	–	✓	42.58	93.06	45.85	101.81	46.10	98.48
Ours	✓	✓	**0.19**	**0.28**	**11.02**	**34.14**	**11.03**	**33.03**

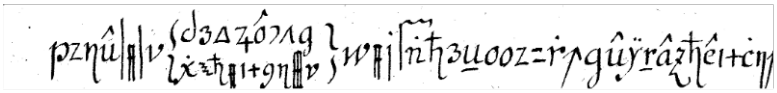


(a)



GT: beschwehre euch dass ihr mir das *nee* wort gebet
Pred: beschwehre euch dass ihr mir das *nee* wort gebet

(b)



GT: dieses { project geheimniss } erfahreN durch eine heimliche
Pred: dieses *bigx* *bigx* neugier *tri..* erfahreN durch eine heimlich

(c)

Figure 2: (a) Attention visualization illustrating alignment between handwritten cipher regions and decoded plaintext tokens. (b) Successful prediction example, where the model output (blue) matches the ground truth (black). (c) Failure case, where predictions (red) deviate from the ground truth (black), illustrating typical error patterns.

distance at the character level, CER $= \frac{S_c+D_c+I_c}{N_c}$, where S_c, D_c, and I_c denote the numbers of character substitutions, deletions, and insertions, respectively, and N_c is the total number of characters in the reference text. Similarly, WER measures error at the word level and is defined as WER $= \frac{S_w+D_w+I_w}{N_w}$, where S_w, D_w, and I_w denote the numbers of word substitutions, deletions, and insertions, and N_w is the total number of words in the reference text.

4.3 Results

The results are shown in Table 3. Handwriting pretraining yields a CER of 5.93% and WER of 17.99% on held-out handwriting data, indicating strong general visual representations. Table 3 compares direct fine-tuning on Copiale against our two-stage approach. The pretrained model dramatically outperforms the baseline, reducing test-set CER from 46.10% to 11.03% and WER from 98.48% to 33.03%. These results show that non-cipher handwriting data substantially improves decipherment accuracy, even though it contains no cryptographic structure.

5 Qualitative Analysis

Qualitative evaluation indicates that the model learns semantically meaningful alignments between handwritten cipher symbols and corresponding plaintext segments, even without explicit ciphertext-level supervision. As shown in Figure 2(a), the attention maps reveal coherent correspondences between spatial regions of the cipher input and decoded plaintext tokens, suggesting that the model captures the underlying structure of the cipher.

Figures 2(b) and (c) present representative successful and failure cases, respectively. In Figure 2(b), the predicted plaintext closely matches the ground truth, demonstrating reliable decoding performance across diverse inputs. In contrast, Figure 2(c) illustrates typical failure patterns, where predictions deviate from the ground truth, highlighting the limitations of the current model.

6 Conclusion

We proposed Transcription-Free Decipherment, an end-to-end approach to historical cipher decipherment that directly maps handwritten cipher images to plaintext. Using the Copiale cipher, we introduced a new publicly available image-to-

plaintext dataset and demonstrated that handwriting pretraining on non-cipher data dramatically improves decipherment performance. By collapsing transcription and decipherment into a single learnable mapping, our approach reduces annotation effort, simplifies processing pipelines, and opens new directions for scalable analysis of historical encrypted manuscripts.

Acknowledgments

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: Echoes of History: Analysis and Decipherment of Historical Writings (DESCRYPT); the Beatriu de Pinós del Departament de Recerca i Universitats de la Generalitat de Catalunya (2022 BP 00256); European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617; the Spanish projects CNS2022-135947 (DOLORES), PID2021-126808OB-I00 (GRAIL) and PID2024-157778OB-I00 (SUKIDI), the Consolidated Research Group 2021 SGR 01559 from the Research and University Department of the Catalan Government, and PID2023-146426NB-100 funded by MCIU/AEI/10.13039/501100011033 and FSE+. Alicia Fornés acknowledges financial support for her general research activities from ICREA under the ICREA Academia (Departament de Recerca i Universitats de la Generalitat de Catalunya).

References

Nada Aldarrab. 2017. Decipherment of historical manuscripts. Master's thesis, University of Southern California.

Théodore Bluche, Jérôome Louradour, and Ronaldo Messina. 2017. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1050–1055. IEEE.

Emmanuele Grosicki and Haikal El-Abed. 2011. Icdar 2011-french handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1459–1463. IEEE.

Lei Kang, J. Ignacio Toledo, Pau Riba, Mauricio Villegas, Alicia Fornés, and Marçal Rusinol. 2018. Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In *German Conference on Pattern Recognition*, pages 459–472. Springer.

Lei Kang, Pau Riba, Mauricio Villegas, Alicia Fornés, and Marçal Rusiñol. 2021. Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition*, 112:107790.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2022. Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*, 129:108766.

Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. 2013. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th international conference on document analysis and recognition*, pages 560–564. IEEE.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2011. The copiale cipher. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 2–9.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2012. The secrets of the copiale cipher. *Journal for Research into Freemasonry and Fraternalism*, 2(2):314.

Jan Kohút and Michal Hradiš. 2025. Practical fine-tuning of autoregressive models on limited handwritten texts. In *International Conference on Document Analysis and Recognition*, pages 22–39. Springer.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

U-V. Marti and Horst Bunke. 2002. The IAM-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5(1):39–46.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, George Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the DECRYPT project. *Cryptologia*, 44(6):545–559.

Beáta Megyesi. 2020. Transcription of historical ciphers and keys. In *3rd International Conference on Historical Cryptology, Histocrypt 2020*, pages 106–115. Linköping University Electronic Press.

Martín Méndez, Pau Torras, Adrià Molina, Jialuo Chen, Oriol Ramos-Terrades, and Alicia Fornés.

2024. Structured analysis and comparison of alphabets in historical handwritten ciphers. In *European Conference on Computer Vision*, pages 330–344. Springer.

Xusen Yin, Nada Aldarrab, Beáta Megyesi, and Kevin Knight. 2019. Decipherment of historical manuscript images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 78–85. IEEE.