# Joint Transcription and Decryption of Images of Ciphered Handwritten Documents

Marino Oliveros Blanco

February 16, 2026

**Abstract**

Since antiquity, humans have used ciphers — systematic methods of encoding information to conceal its meaning — to protect sensitive communications. Historical ciphered manuscripts present a significant challenge at the intersection of cryptography, linguistics, and computer vision. Current decipherment approaches rely on a two-stage pipeline of transcription followed by decryption, where errors from the initial stage propagate to the final output. This work develops Direct Image Decryption, a novel end-to-end approach that directly maps encrypted manuscript images to decrypted text, bypassing the intermediate transcription step. We implement a synthetic data generation pipeline producing cipher-like images and evaluate both traditional two-stage and Direct Image Decryption architectures, demonstrating that our approach outperforms the traditional pipeline on both the real cipher manuscript and synthetically generated data.

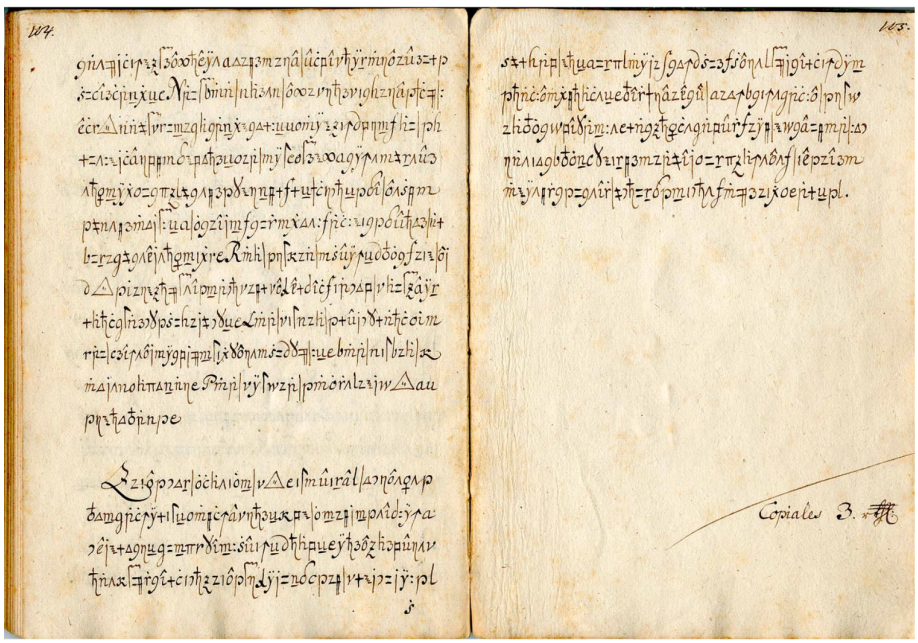
**Keywords** Historical cryptography, handwritten text recognition, neural networks, Copiale manuscript.

## 1 INTRODUCTION

Historical encoded manuscripts represent a unique challenge at the intersection of computer vision, cryptographic analysis, and linguistics. Thousands of encrypted manuscripts reside in archives worldwide — diplomatic correspondence, private letters, and manuscripts tied to secret societies — their contents remaining hidden for centuries. Deciphering these sources demands a joint pipeline: vision techniques to process handwritten imagery, cryptographic methods to model substitution schemes, and linguistic knowledge to validate the plaintext. Scholars across these disciplines routinely encounter the same core difficulties. Among these cryptographic artifacts, the Copiale Cipher stands as a particularly remarkable case study. This 18th-century encrypted manuscript, discovered in Germany, consists of 105 pages filled with symbols and abstract glyphs that puzzled cryptographers for over two centuries. It is a homophonic substitution cipher, where individual plaintext letters are represented by multiple cipher symbols drawn from an alphabet of over 90 distinct glyphs, a technique designed to thwart frequency analysis.

• Contact E-mail: marino.oliveros@autonoma.cat
• Supervised by: Alicia Fornés
• Academic Year 2025/26

The manuscript remained undeciphered until 2011, when Kevin Knight and Beáta Megyesi revealed its contents using computational techniques [10], demonstrating that the text encoded German and contained the initiation ritual of a secret society known as "the Oculists". Their approach, like most current methods, relied on a two-step pipeline: first transcribing symbols from images to text representations, then decrypting those transcriptions to obtain the plaintext. This dependency creates a critical vulnerability. Errors introduced during transcription inevitably propagate to the decryption stage, compounding inaccuracies. The transcription phase also requires extensive manual labor and expertise, limiting scalability and increasing the chance for human error. The Rammanacoil cipher study by Dinnisen and Kopal illustrates this challenge, where researchers manually transcribed entire manuscript pages before attempting decryption [7]. These problems represent a systemic bottleneck in historical cryptanalysis, reinforcing the need for approaches that generalize beyond individual manuscripts. Current approaches rely on a two-stage pipeline, but fundamental problems persist: error propagation from transcription to decryption and limited availability of real manuscript data to train deep learning models. This work proposes Direct Image Decryption, a paradigm shift that learns the direct mapping from encoded manuscript images to decrypted plaintext in a single end-to-end model, offering elimination of transcription-related error propaga-



| Plaintext (German) | Ciphertext |
|---|---|
| A | |
| Ä | |
| B | |
| C | |
| D | |
| E | |
| F | |
| G | |
| H | |
| I | |
| J | |
| K | |
| L | |
| M | |
| N | |
| O | |
| Ö | |
| P | |
| R | |
| S | |
| T | |
| U | |
| Ü | |
| V | |
| W | |
| X | |
| Y | |
| Z | |
| SCH | |
| SS | |
| ST | |
| CH | |
| repeat previous consonant | |
| EN / EM | |
| space | |

Fig. 1: Final page of the Copiale manuscript along with its homophonic cipher key

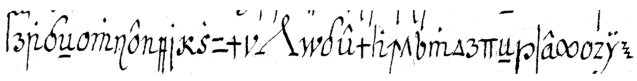


Fig. 2: Sample image of a Copiale manuscript sentence available at the 2024 ICDAR Competition

tion, removal of the manual transcription bottleneck, and the possibility of capturing visual features lost during symbol transcription. While this work evaluates on the Copiale Cipher as its primary benchmark, the proposed framework is cipher-agnostic, and should be tested on multiple encrypted manuscripts to represent a promising direction. The primary objectives of this work are threefold. First, we develop a comprehensive synthetic data generation pipeline capable of producing realistic Copiale-like manuscript images, creating over 115,000 training samples from historical German texts including Goethe's Faust and Kant's Critique of Pure Reason. Second, we implement and evaluate a CRNN-based transcription model with CTC loss, achieving 91.5% token accuracy on synthetic data and 91.1% on the original Copiale manuscript, establishing a strong baseline for the traditional approach. Third, we implement and compare the Direct Image Decryption approach against the two-stage transcription-decryption pipeline, demonstrating that Direct Image Decryption achieves improvement in token accuracy on both synthetic and original manuscript data, while also reducing Word Error Rate (WER) and Character Error Rate (CER) across all evaluated datasets, thus validating our hypothesis that eliminating the transcription step reduces error propagation and improves overall decipherment performance.

The remainder of this paper is organized as follows. Section 2 reviews related work in historical cipher decipherment and handwritten text recognition. Section 3 describes our synthetic data generation methodology. Section 4 presents the architecture and training details of our transcription and decryption models. Section 5 reports decryption of the 1-step pipeline Direct Image Decryption in both real and synthetic data. Section 6 discusses the implications and results of both approaches across a series of experiments conducting a comprehensive comparison between methods. Finally, Section 7 concludes and discusses the different possible improvements and future work for the project.

## 1.1 Objectives

The primary objective of this project is to investigate whether direct image-to-plaintext decryption (Direct Image Decryption) can outperform the traditional two-stage transcription-decryption pipeline for historical encrypted manuscripts. To achieve this overarching goal, we establish the following specific objectives:

1. **Synthetic Data Generation for Training:** Develop a pipeline for generating synthetic Copiale-like manuscript images that closely resemble the visual characteristics of the original 18th-century document. This includes implementing appropriate augmentation techniques to simulate aging, degradation, and handwriting variations, producing a dataset of over 115,000 images from historical German texts.

2. **Transcription Model Development:** Create and optimize a high-performance handwritten text recognition (HTR) model capable of accurately transcribing Copiale cipher symbols. The goal is to achieve token accuracy exceeding 90% on both synthetic and real manuscript data to establishing a strong baseline that eliminates the argument that poor transcription performance undermines the comparison with Direct Image Decryption.

3. **Two-Stage Pipeline Implementation:** Implement a complete transcription-decryption two-stage pipeline architecture using state-of-the-art techniques. This

serves as the baseline against which Direct Image Decryption will be compared. We want to achieve a fair evaluation by using the best possible traditional approach instead of a suboptimal implementation.

4. **Direct Image Decryption Architecture Development:** Design and implement an end-to-end model that directly maps encrypted manuscript images to decrypted plaintext, bypassing the intermediate transcription step entirely. The aim is to eliminate error propagation, handwritten character segmentation errors and capture visual features lost during symbol transcription.

5. **Comparative Evaluation:** Conduct quantitative and qualitative comparisons between the Direct Image Decryption approach and the traditional two-stage pipeline across multiple datasets, including synthetic data and the original Copiale manuscript. The evaluation metrics for our comparison include character accuracy, edit distance, word error rate (WER), and character error rate (CER).

6. **Generalization Assessment:** Test the trained models on out-of-distribution data, such as texts by different authors (e.g., Novalis), different languages (English) and different sequence lengths, to evaluate the strength and generalization capabilities of both approaches.

### 1.2 Methodology and timeline

For the timeline, the project was divided in 3 parts, according to the follow-up sessions of the TFE, appreciable in the Gantt Chart present in the appendix A as the light to darker blue and purple lines. Part 1 consists of the project kick off and state-of-the-art research, crucial steps to ensure the best project handling possible, the environment was set up and I conducted synthetic data generation, we will delve deeper on this task later. Part 2 focuses on the model creation tasks both main pipelines, Transcription-Decryption and Direct Image Decryption. Lastly, in Part 3 we built our conclusions and report, experimented, troubleshoot and improved our model.

On a more technical note, I have been working with multiple datasets of both synthetically generated data and the original Copiale manuscript obtained from the ICDAR Competition 2024 (ICDAR)[13], consisting of a little over 2000 segmented line images in grayscale of the original manuscript, these images are annotated with their transcriptions; the name of the symbols, not their decryption, for that we have to assign each image to their respective lines from the decrypted file.

The code used for synthetic data generation is based on work I conducted during my internship at the Computer Vision Center (CVC). Additionally the augmentation strategies employed are drawn from CVC's research.

The environment in which the research is being conducted is a virtual machine server owned by CVC, which consists of a placeset for datasets and my own working space. The technical specifications of said work environment are as follows: 8GPUs of which 6 NVIDIA GeForce RTX 3090 and 2 Quadro RTX 6000. This environment altogether ensures sufficient data diversity, and computational power to guarantee the successful realization of the project.

## 2 Related Work

The field of historical cipher decipherment has evolved from purely manual cryptanalysis to sophisticated computational approaches. This section reviews the current state of the art, examining both traditional two-stage methodologies and recent advances in neural decipherment, it also highlights the persistent challenges that motivate our Direct Image Decryption approach. In this section we analyze the current State-Of-The-Art approaches.

### 2.1 Historical Cipher Decipherment

The computational decipherment of historical manuscripts gained significant momentum with the aforementioned breakthrough work on the Copiale Cipher [10]. Their approach established the now-standard two-stage pipeline: first transcribing cipher symbols from manuscript images into machine-readable text, then applying cryptanalytic techniques to recover the plaintext. This methodology successfully revealed that the Copiale manuscript encoded German text describing the rituals of an 18th-century secret society.

The DECRYPT project has since become a cornerstone initiative in historical cryptology, creating standardized datasets and benchmarks for evaluating decipherment approaches [6]. However, as Dinnisen and Kopal demonstrated in their work on the Rammanacoil cipher, the transcription phase remains heavily dependent on manual effort, with their team requiring extensive human labor to transcribe the Dutch manuscript before attempting decryption [7]. This bottleneck significantly limits the scalability of decipherment efforts, particularly for lengthy or complex manuscripts, as well as non orthodox ciphers.

### 2.2 Neural Approaches to Cipher Decipherment

Recent advances in neural architectures have transformed both the transcription and decryption components of the traditional pipeline. For decryption, Kambhatla et al. [9] introduced neural language models with beam search algorithms, demonstrating improved accuracy over frequency-analysis methods by learning statistical patterns of natural language. Building on this foundation, Aldarrab and May [1] developed transformer-based multilingual models achieving remarkable results: less than 1% error on synthetic ciphers and 5.47% on the authentic Borg cipher. Aldarrab's doctoral work [2] further explored end-to-end approaches while acknowledging the persistent challenge of bridging synthetic and real historical data, particularly regarding handwritten text segmentation.

For transcription, Convolutional Recurrent Neural Networks (CRNNs) with Connectionist Temporal Classification (CTC) loss [8] have emerged as the dominant architecture for sequence recognition [14]. These models combine convolutional layers for visual feature extraction with recurrent layers for sequence modeling, eliminating the need

for explicit character segmentation. Yin et al. [15] pioneered automatic segmentation and transcription for cipher manuscripts, addressing challenges such as unusual symbol sets, degraded conditions, and the absence of dictionary-based error correction available to standard HTR systems. More recently, attention-based approaches have shown superior performance: Bluche et al. [4] introduced an end-to-end model combining LSTMs with attention mechanisms for paragraph-level recognition, achieving state-of-the-art results without requiring explicit segmentation or language models. This attention-based paradigm directly influences our Direct Image Decryption approach, which similarly employs attention to jointly optimize visual feature extraction and decryption.

The ICDAR Competition [13] on ciphers has contributed standardized datasets and evaluation protocols for historical cipher recognition. However, the scarcity of real manuscript data—rarely more than a few thousand lines—continues to necessitate heavy reliance on synthetic data generation. Generation which often fails to capture and resemble the full complexity of the authentic manuscripts.

Despite these advances, all current neural decryption methods maintain fundamental dependency on accurate transcription, creating error propagation that directly affect decryption performance.

### 2.3 Limitations of Current Approaches

Three fundamental limitations persist in the current cipher decipherment methodologies. First, error propagation from transcription to decryption compounds inaccuracies and limits overall system performance. A single transcription error can cascade through the decryption process and corrupt the interpretation of surrounding text.

Second, the transcription bottleneck requires much manual effort or large quantities of training data. For many historical ciphers, neither enough man power nor sufficient annotated data is readily available. This is the most important of the limitations, not having enough high quality homogeneous data is what makes decrypting documents such as the Voynich Manuscript [5], part of the Zodiac Killer Ciphers (particularly the brief Z13 and Z32, 13 and 32 characters long respectively) [11] or the Beale Ciphers (3 ciphertexts containing roughly 2,500 characters) a daring task [3].

Lastly, the gap between synthetic and real data remains substantial. While synthetic data enables the training of large-scale models, it typically fails to replicate the exact characteristics of historical manuscripts—including aging effects, ink degradation, writing style variations, and document damage. Models trained primarily on synthetic data often exhibit worse performance when applied to the original historical ciphers.

These limitations motivate the exploration of alternative approaches that bypass the transcription stage entirely. Approaches that to escape the difficulties of handwritten character segmentation, directly map visual features of the encrypted manuscript images to decrypted plaintext. This end-to-end approach, which we term Direct Image Decryption, represents a fundamental departure from established methodologies, constituting the core contribution of this work.

## 3 Synthetic Data Generation

The scarcity of annotated historical cipher manuscripts presents the most fundamental challenge for training deep learning models. With only approximately 2,000 segmented line images available from the original Copiale manuscript through the ICDAR Competition dataset [13], we developed a comprehensive synthetic data generation pipeline to produce training samples of sufficient quantity and quality for a robust model development of both the transcription-decryption pipeline and the Direct Image Decryption model. This pipeline takes lines of text as input and generates augmented images of the original text encoded into Copiale, along with the transcriptions and original decrypted plaintext.

Our synthetic data must satisfy three requirements. First, have visual similarity to the original Copiale manuscript, including appropriate symbol shapes, spacing, and overall appearance. Second, the underlying text must reflect the linguistic patterns of 18th-century German, as this was the language encoded in the original cipher. Third, the data must incorporate realistic degradation effects—including noise, ink variations, and aging marks—to reduce the gap between the synthetic samples and authentic historical documents.

### 3.1 Text Source Selection

To ensure authenticity, we selected historical German texts that chronologically and stylistically align with the Copiale manuscript, or what is considered Old German. Our primary corpus contains four major works: Goethe's *Faust* (1808-1832), Kant's *Critique of Pure Reason* (1781), *the Lutheran Bible* (1534 original, 1760 revision used), and Adalbert Stifter's *Nachsommer* (1857). These texts provide over 115,000 lines of period-appropriate German text in total.

Text preprocessing involved multiple stages to ensure compatibility with the Copiale cipher vocabulary. We filtered the corpus to retain only the 106 characters present in the original cipher, removing modern punctuation marks and characters absent from the original Copiale manuscript such as asterisks, parentheses, and accented characters uncommon in the cipher. Line segmentation ensured that generated images contained 12-40 characters, matching the length distribution observed in the original manuscript.

For additional evaluation, we generated extra datasets using texts in another language, including excerpts from Bret Eaton Ellis' *American Psycho* and Steinbeck's *East of Eden* both in English. We also created another synthetically generated dataset with 1300 images for testing purposes based on the shorter, poetic verses of the Old German text *Hymns of the Night* by Novalis; these will be discussed later on in the Experiments section.

### 3.2 Visual Representation Pipeline

The encoding from text to cipher-like images employs the *"Copiale.ttf"* font file, which maps standard Unicode characters to their corresponding cipher glyphs. This font encodes the visual representation of each cipher symbol, so automatic generation of cipher-like text from the plaintext input can be achieved. For example, the trigraph "sch"

(common in German) maps to a dagger-like symbol through an intermediate ASCII representation before rendering as the final glyph. Using this font, if you were to type "T" the dagger-like symbol would appear.

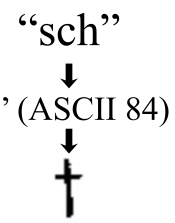



Fig. 3: Input to Copiale encoding

The mapping process utilizes the vocabulary file from the DECRYPT project [6], which defines the correspondence between input text sequences and their cipher symbol representations. This ensures that our synthetic data maintains the same symbol-to-meaning relationships as the authentic Copiale manuscript.

Initial image generation produces clean, high-resolution renderings of cipher text. However, these generated images differ substantially from aged historical manuscripts. To bridge this gap, we apply a comprehensive augmentation pipeline that simulates the degradation effects observed in the authentic document.

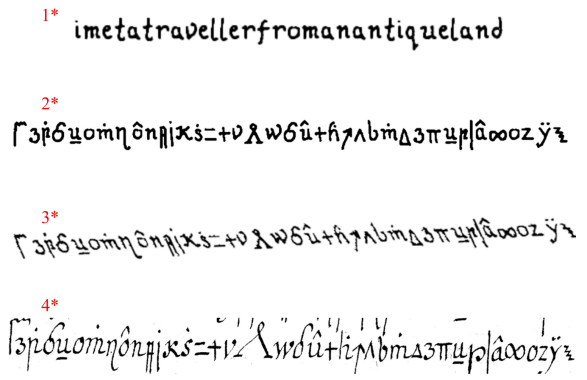



Fig. 4: Comparison of 1* Plaintext in Copiale font, 2* Encoded text into Copiale (non-augmented), 3*Encoded text into Copiale (augmented), and 4*Original manuscript image

### 3.3 Augmentation Strategy

Our augmentation approach applies multiple transformations to simulate realistic manuscript aging and variation:

**Degradation effects** include Gaussian noise to simulate paper texture and scanning artifacts, random erosion and dilation operations to replicate ink spread and fading, gamma correction for brightness variations, and Kanungo noise patterns that model document degradation, including spots, stains, and fiber patterns in aged paper.

**Geometric transformations** apply random rotation (±3 degrees) to simulate natural writing slant variations, shearing operations to introduce perspective distortions, random scaling to vary character sizes within realistic bounds, and random cropping to generate diverse image boundaries and eliminate edge writings corresponding to other lines or the presence of noise.

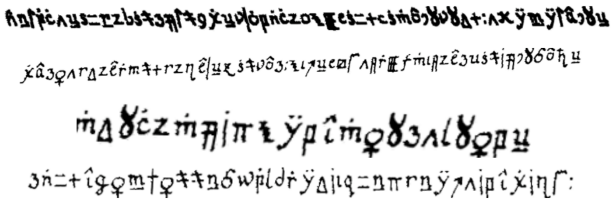


Fig. 5: Different augmentation effects

The augmentation parameters were slightly tuned to produce realistic variations as similar as possible to the original manuscript. Figure 5 shows some of the range of augmentation effects obtainable through our augmentation pipeline.

### 3.4 Dataset Statistics and Organization

Our primary synthetic dataset, referred to as the "Faust" dataset, comprises 115,000 line images generated from the aforementioned historical German corpus. The dataset employs an 80/10/10 split for training, validation, and testing respectively. The "Faust" synthetic dataset also contains the transcription data and original decrypted plaintext per image. The original Copiale manuscript is formed of 2000 grayscale images of the original manuscript along with the transcription and decrypted plaintext.

Figure 4 presents a visual comparison between authentic Copiale manuscript images from the ICDAR dataset and our synthetically generated samples. The augmentation section of the pipeline (3*) effectively achieves correct symbol morphology, spacing patterns, and degradation effects. The main differences remain in the degree of aging marks and ink consistency, the paper detailing, texture and exact tracing techniques of the author or authors, with real manuscripts exhibiting more pronounced historical wear. How to limit these differences will be discussed in the Conclusions & Future Work section.

The vocabulary distribution in our synthetic data closely resembles that of the original manuscript. Common Old German letter combinations such as "u__" (representing "EN/EM"), "CapitalLambda" (representing "T"), and "z" (representing "D") appear with frequencies proportional to their usage in 18th-century German texts. This frequency similarity ensures that models trained on the synthetically generated data encounter symbol patterns consistent with the ones present in the real manuscript images.

## 4 Transcription & Decryption Pipeline

The two-stage pipeline represents the traditional approach to cipher manuscript decipherment, converting cipher

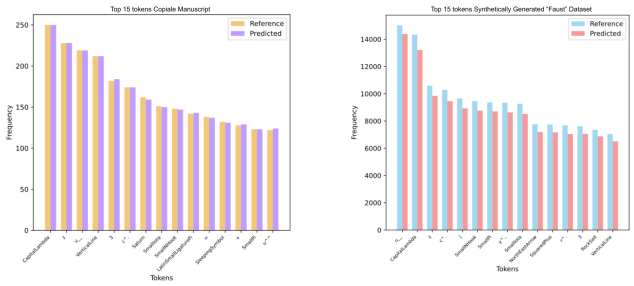



Fig. 6: Token frequency: Original Copiale manuscript vs. Synthetically generated "Faust" dataset

manuscript images into decrypted plaintext through transcription followed by decryption. This section describes both components of the pipeline, analyzing their architecture and training strategies.

Figure 7 illustrates the complete two-stage pipeline from manuscript images to plaintext German text.

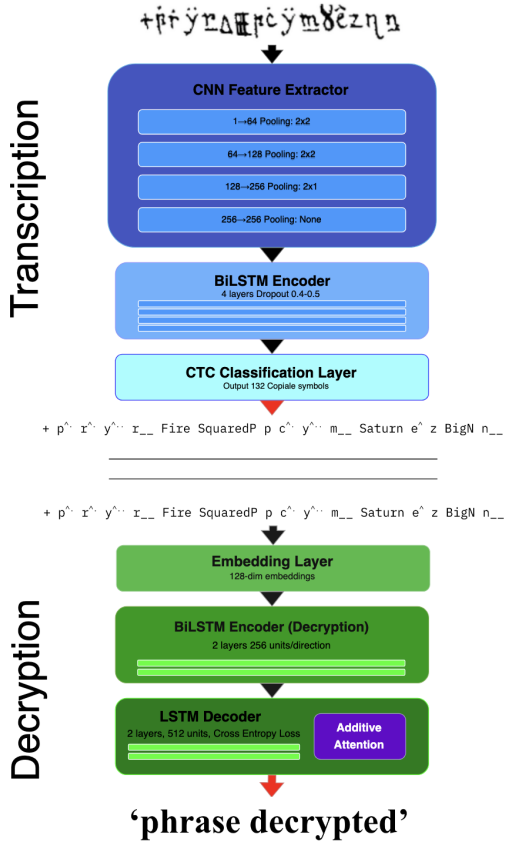



Fig. 7: Complete two-stage pipeline: manuscript image → transcription + transcription → decryption → plaintext

## 4.1 Stage 1: Transcription

The transcription component converts cipher manuscript images into sequences of symbol tokens, enabling subsequent decryption through language model-based approaches and cryptanalysis. Figure 8 demonstrates a successful transcription example.

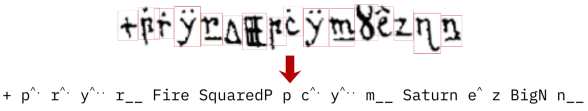



Fig. 8: Example of successful transcription converting input image to cipher symbol token sequence

The goal of achieving high-quality transcription is critical for fair comparison with Direct Image Decryption, dismissing the potential counterclaim that "Direct Image Decryption is not better, the transcription approach simply under performs."

### 4.1.1 Model Architecture

The transcription model employs a Convolutional Recurrent Neural Network (CRNN) architecture with Connectionist Temporal Classification (CTC) loss [14]. Input images are resized to 64 pixels height (maintaining aspect ratio), padded/cropped to 800 pixels width, and normalized to [0,1].

**CNN Feature Extractor:** Four convolutional blocks (1→64→128→256→256 channels) progressively extract hierarchical features with 3×3 convolutions, batch normalization, ReLU activation, and pooling (2×2, 2×2, 2×1, none). Output feature maps are reshaped to (batch_size, width/4, 2048).

**Bidirectional LSTM:** 4 layers with 256 hidden units per direction process CNN features bidirectionally to disambiguate visually similar symbols. Dropout (0.4-0.5) provides regularization, producing 512-dimensional vectors per timestep.

**CTC Classification:** The vocabulary comprises 132 Copiale cipher tokens plus special tokens (blank, padding, UNK). Greedy CTC decoding selects the most probable token at each timestep, then collapses repetitions and removes blanks.

The CTC loss [8] enables alignment-free training by marginalizing over all possible alignments:

$$\mathcal{L}_{\text{CTC}} = -\log P(y|x) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p_t(\pi_t|x) \tag{1}$$

where $y$ is the target symbol sequence, $x$ is the input image, $T$ is the sequence length, $\pi$ is an alignment path, $\mathcal{B}^{-1}(y)$ is the set of valid alignments, and $p_t(\pi_t|x)$ is the probability of symbol $\pi_t$ at timestep $t$.

### 4.1.2 Training Configuration

AdamW optimizer [12] with learning rate $3\times10^{-4}$, weight decay $1\times10^{-4}$, batch size 8, gradient clipping (max norm 1.0), and ReduceLROnPlateau scheduler (factor 0.1, patience 5 epochs). Training runs 100 epochs with early stopping. Transcription results will be analyzed in depth in Section 6.

### 4.2 Stage 2: Decryption

The decryption component takes the transcribed symbol sequences and generates the decrypted German plaintext using sequence-to-sequence architecture with attention mechanisms—effectively implementing the inverse of the Copiale homophonic substitution cipher.

#### 4.2.1 Model Architecture

Encoder-decoder architecture with additive attention. The encoder uses bidirectional LSTM (2 layers, 256 units per direction) to create contextual representations. The decoder generates plaintext character-by-character through matching LSTM architecture, with attention focusing on relevant encoded cipher portions. Character-level embeddings of dimension 128 for both input and output. Output projection maps decoder states to German alphabet vocabulary (uppercase, lowercase, special characters, control tokens: SOS, EOS, PAD, UNK). The decryption model uses the same vocabulary structure as transcription for its input, assuring seamless integration. Approximately 100-200 cipher symbol tokens; depending on the dataset. The model is optimised with AdamW (lr=0.001, weight decay=0.0001) using a ReduceLROnPlateau scheduler, with teacher forcing annealed linearly from 0.5 to 0.0 across epochs and gradient clipping applied at a max norm of 1.0.

#### 4.2.2 Training Strategy

AdamW optimizer (learning rate $3\times10^{-4}$, weight decay $1\times10^{-4}$, dropout 0.4), batch size 16, 15-35 epochs with early stopping based on validation edit distance. ReduceLROnPlateau scheduler reduces learning rate when validation plateaus.

The training process uses the transcription model's vocabulary for encoder input, creating unified token space across both stages.

## 5 Direct Image Decryption

Direct Image Decryption learns direct image-to-plaintext mapping in a single end-to-end model, addressing three limitations: error propagation from transcription to decryption, information loss during symbolic conversion (discarding visual cues like symbol confidence and spacing), and architectural complexity of maintaining separate models. By jointly optimizing visual feature extraction and decryption without intermediate discrete decisions, the model discovers which visual features are most relevant for decipherment.

### 5.1 Architecture Overview

The architecture comprises a CRNN feature extractor of 5 blocks, producing sequential visual representations, and an attention-based decoder generating plaintext characters autoregressively. The critical difference from transcription: intermediate representations remain continuous—the model never commits to discrete cipher symbol decisions. This allows end-to-end gradient flow. A joint optimization of visual feature extraction and decryption.

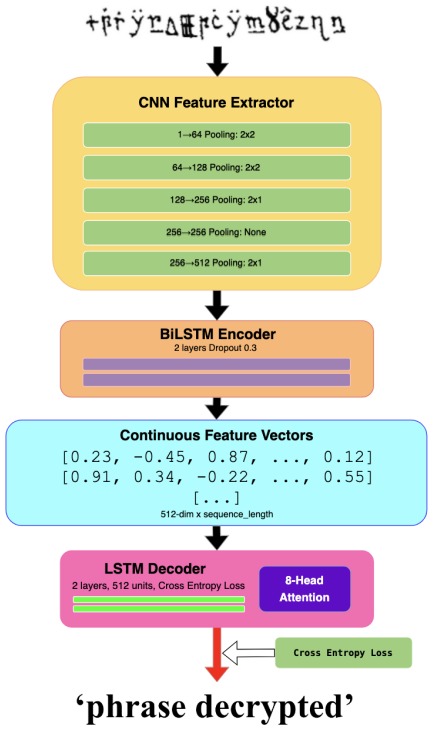



Fig. 9: Direct Image Decryption architecture: images processed through CRNN feature extraction, decoded directly to plaintext through attention-based LSTM decoder

### 5.2 CRNN Feature Extractor

The feature extractor extends the successful CRNN architecture from transcription with modifications to support end-to-end training. It processes grayscale images through a deeper five-block CNN structure, expanding upon the four-block design used in transcription (Section 4).

Following CNN blocks, feature maps of dimension (batch_size, 512, height/16, width/4) are reshaped to (batch_size, width/4, 512 × height/16). A 2-4 layer bidirectional LSTM (256 units per direction) produces contextualized visual representations of dimension (batch_size, sequence_length, 512).

The CRNN initializes with pretrained transcription model weights, providing strong starting point for visual feature extraction. However, unlike the transcription pipeline where CRNN weights remain fixed during decryption training, Direct Image Decryption allows fine-tuning during end-to-end optimization. The feature extractor output—sequences of 512-dimensional vectors captures cipher symbols, spatial relationships, and visual characteristics, without committing to discrete symbol decisions.

### 5.3 Attention-Based Decoder

The decoder generates German plaintext characters autoregressively, conditioning on encoded image features and previously generated characters. It employs 2-layer LSTM (512 hidden units) with 8-head multi-head attention mechanism.

At each timestep, the decoder receives embedding of the previous character (dimension 128) or SOS token. The LSTM produces a query vector, and multi-head attention (8 heads, per-head dimensionality 64) computes scores

between query and encoded image features, determining which manuscript regions are most relevant for predicting the current character. Multiple heads allow simultaneously attending to different regions.

Attended features are concatenated with the LSTM output, passed through linear projection (dimension 512), then final output projection to German alphabet vocabulary. Cross-entropy loss trains the model with gradients flowing backward through attention, decoder, and feature extractor. Additionally, teacher forcing during training stabilizes learning; inference uses autoregressive generation until an EOS token is found.

### 5.4 Training Configuration

End-to-end training uses AdamW [12] (learning rate $3\times10^{-4}$, weight decay $1\times10^{-4}$, dropout 0.3), batch size 16 (larger than transcription's 8—CTC loss absence reduces memory requirements), 35-50 epochs with early stopping based on validation edit distance.

Training begins by loading pretrained CRNN weights when available, providing a base initialization though subsequently fine-tuning. This transfer learning significantly accelerates convergence versus random initialization. Gradient clipping (max norm 1.0) prevents instability given the long backpropagation path. ReduceLROnPlateau scheduler (factor 0.1, patience 5 epochs) enables fine-grained optimization.

The model is trained end-to-end to directly generate plaintext from images, jointly optimizing all parameters:

$$\mathcal{L}_{\text{PH}} = -\sum_{t=1}^{T} \log P(w_t | w_{<t}, I; \theta) \tag{2}$$

where $I$ is the input manuscript image, $w_t$ is the plaintext character at position $t$, $w_{<t}$ represents all previous characters, and $\theta$ represents all model parameters including both the CRNN feature extractor and the decoder. This objective directly optimizes for accurate plaintext generation without intermediate transcription objectives, allowing gradients to flow through the entire network.

## 6 EXPERIMENTS

This section presents experimental results comparing both architectures across multiple scenarios. Firstly, we compare architectural components and differences between approaches. We then evaluate performance on synthetic data, analyzing the transcription-decryption pipeline and Direct Image Decryption independently. Secondly, we assess both approaches on the original Copiale manuscript. Finally, we test hypothesis through experiments on sequence length variation and cross-language text.

### 6.1 Architecture Comparison

To understand the performance differences between the two-stage pipeline and Direct Image Decryption, we examine their architectural similarities and differences.

The transcription model uses a four-block CNN (1→64→128→256→256 channels) with bidirectional LSTM (2-4 layers, 256 units per direction), while Direct Image Decryption extends this with a deeper five-block CNN (1→64→128→256→256→512 channels) before its BiLSTM encoder. Both decoding stages employ 2-layer LSTMs with 512 hidden units and 128-dimensional embeddings, though they differ in attention mechanisms: the decryption model uses simpler additive attention, while Direct Image Decryption employs 8-head multi-head attention. These architectural difference translate into maintaining strict parity for controlled comparison versus optimizing each approach independently.

Table 1 summarizes the key architectural components and highlights where the approaches differ.

Despite these capacity differences, the fundamental comparison remains: both approaches use CRNN-based visual encoding followed by attention-based decoding. The critical distinction: Two independent models versus single end-to-end training. The two-stage pipeline commits to discrete symbol decisions early, preventing gradient flow from decryption errors back to visual feature extraction. Direct Image Decryption maintains continuous representations, allowing end-to-end optimization where visual features adapt directly to decryption requirements. Whether this end-to-end learning advantage outweighs the architectural differences is the central empirical question answered in the subsequent results sections.

Comparison with the State-Of-The-Art could not be done, as currently baselines for a Direct Image Decryption approach do not exist. Papers on this matter either delve into transcription or decryption separately; a conjoined baseline does not exist.

### 6.2 Evaluation Metrics

All models are evaluated using four metrics computed case-insensitively (text normalized to lowercase before comparison, treating "Copiale", "COPIALE", and "cOpIaLe" as equivalent). This ensures fair evaluation focused on content accuracy rather than capitalization conventions. Decryption success rate (1-NED) is occasionally reported for interpretability, also used in the sequence length experiment in Section 6.5.

**Token Accuracy:** Percentage of exactly correct predictions. For transcription, tokens are cipher symbols; for decryption/Direct Image Decryption, tokens are plaintext characters. Higher is better [0.0-1.0].

**Normalized Edit Distance (NED):** Levenshtein distance (insertions, deletions, substitutions) normalized by sequence length. Lower is better [0.0-1.0].

**Word Error Rate (WER):** Token-level error computed as (S+D+I)/N, where S=substitutions, D=deletions, I=insertions, N=reference tokens. Despite its name, operates at token level. Lower is better [0.0-1.0].

**Character Error Rate (CER):** Character-level equivalent of WER, providing finer-grained evaluation. Lower is better [0.0-1.0].

### 6.3 Results on Synthetic Data

We evaluate both approaches on synthetic data to assess their performance under controlled conditions. All models were trained on the 115,000-image "Faust" dataset (Section 3) and tested on two datasets: the held-out "Faust" test set

TABLE 1: ARCHITECTURAL COMPARISON BETWEEN TWO-STAGE PIPELINE AND DIRECT IMAGE DECRYPTION.

| Component | Two-Stage Pipeline | Direct Image Decryption |
|---|---|---|
| CNN Blocks | Transc: 4 blocks (1→64→128→256→256) | 5 blocks (1→64→128→256→256→512) |
| Pooling | 2×2, 2×2, 2×1, none | 2×2, 2×2, 2×1, none, 2×1 |
| CNN Output | 256 channels | 512 channels |
| BiLSTM | Transc: 4 layers; Decr: 2 layers (256 units/dir, dropout 0.4-0.5) | 2 layers (256 units/dir, dropout 0.3) |
| Decoder LSTM | 2 layers, 256 units | 2 layers, 512 units |
| Attention | Additive (linear) | 8-head multi-head |
| Embeddings | 128-dim | 128-dim |
| Optimizer | AdamW | AdamW |
| Learning Rate | Transc: 3e-4; Dec: 1e-3 | 1e-3 |
| Weight Decay | Transc: 1e-4; Dec: 1e-4 | 1e-4 |
| Batch Size | Transc: 8; Dec: 16 | 16 |
| Training | Two sequential models | Single end-to-end |
| Output | Transc: 132 symbols; Dec: 62 chars | 62 chars directly |
| Loss | Transc: CTC; Dec: Cross-entropy | Cross-entropy |
| Representation | Discrete tokens | Continuous features |
| CRNN | Frozen after transcription | Fine-tuned end-to-end |
| Gradients | Blocked at symbols | End-to-end flow |
| Training Time | 10-12 hours total | 8-10 hours |

and the out-of-distribution Novalis dataset (1,300 images of Old German poetry). This evaluation strategy tests both in-distribution performance and generalization to different text styles.

### 6.3.1 Transcription Performance

The transcription model achieves strong performance on the "Faust" test set, as shown in Table 2. With 91.5% token accuracy and only 4.3% normalized edit distance, the model successfully recognizes the vast majority of cipher symbols. The similarity between WER (7.5%) and CER (7.6%) indicates uniform error distribution across the sequence.

TABLE 2: TRANSCRIPTION MODEL PERFORMANCE ON "FAUST" SYNTHETIC TEST SET

| Metric | Score |
|---|---|
| Token Accuracy | 0.915 |
| Edit Distance | 0.043 |
| WER | 0.075 |
| CER | 0.076 |

Figure 10 illustrates representative examples of transcription performance, showing both successful cases and failure modes. The best cases demonstrate accurate recognition even with degraded image quality, while worst cases reveal challenges with visually similar symbols and heavy augmentation artifacts.

### 6.3.2 Comparison: Two-Stage vs. Direct Image Decryption

Table 3 presents the end-to-end decryption performance comparison between the two-stage transcription-decryption pipeline and Direct Image Decryption on both synthetic test sets.

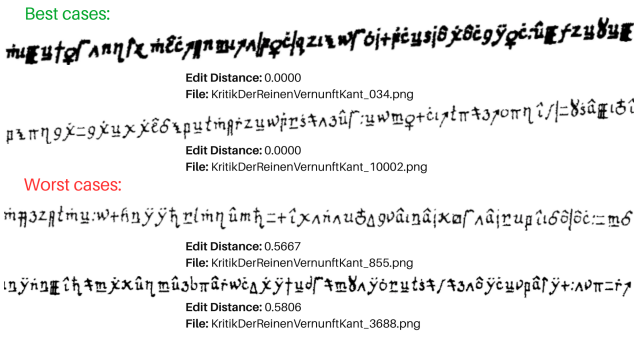



Fig. 10: Best and worst case examples of transcription on synthetic data

**Faust Test Set Results:** On in-distribution data, Direct Image Decryption outperforms the two-stage pipeline across most metrics. The 1.1% improvement in token accuracy (91.3% → 92.4%) validates our hypothesis that eliminating the transcription bottleneck reduces error propagation. The dramatic 49% reduction in WER (0.206 → 0.105) demonstrates Direct Image Decryption's superior ability to generate coherent character sequences. Interestingly, CER is slightly higher for Direct Image Decryption (6.5% vs. 5.6%), suggesting that while Direct Image Decryption makes fewer token-level errors, some character-level substitutions persist.

**Novalis Out-of-Distribution Results:** The performance gap widens substantially on the Novalis dataset, which contains different vocabulary, sentence structures, and poetic phrasing compared to the training data. Direct Image Decryption achieves 75.8% token accuracy compared to 69.5% for the two-stage approach—a 6.3% absolute improvement. This larger advantage on out-of-distribution data suggests that Direct Image Decryption's end-to-end learning enables better generalization. The two-stage pipeline's performance degradation (91.3% → 69.5%) is

TABLE 3: END-TO-END DECRYPTION COMPARISON ON SYNTHETIC DATA

| Dataset | Metric | 2-Stage | Direct Decr. |
|---|---|---|---|
| **Faust** | Token Acc. | 0.913 | **0.924** |
| | Edit Dist. | 0.045 | **0.038** |
| | WER | 0.206 | **0.105** |
| | CER | **0.056** | 0.065 |
| **Novalis** | Token Acc. | 0.695 | **0.758** |
| | Edit Dist. | 0.190 | **0.162** |
| | WER | 0.597 | **0.316** |
| | CER | 0.204 | **0.183** |

more severe than Direct Image Decryption's (92.4% $\rightarrow$ 75.8%), indicating that error propagation from transcription compounds when encountering unfamiliar text patterns.

The WER reduction is particularly striking on Novalis (59.7% $\rightarrow$ 31.6%), demonstrating that Direct Image Decryption maintains better sequence-level coherence even when faced with novel vocabulary and grammatical structures. These results provide strong evidence that the end-to-end paradigm offers robustness advantages beyond simple accuracy improvements.

## 6.4 Results on Original Copiale Manuscript

We evaluate both approaches on approximately 2,000 line images from the authentic 18th-century Copiale manuscript, provided through the ICDAR Competition dataset [13]. This evaluation probes model behavior under real historical conditions, including ink degradation, handwriting variability, and layout irregularities that cannot be fully reproduced through synthetic data generation. Unlike synthetic benchmarks, this setting represents the true target domain for historical cipher decipherment.

### 6.4.1 Transcription Performance

The transcription component generalizes well to the original manuscript, as shown in Table 4. The model achieves 91.1% token accuracy, only 0.4 percentage points lower than its performance on synthetic data. Edit distance (2.3%), WER (1.7%), and CER (1.4%) remain low, indicating robust symbol recognition despite manuscript aging and handwriting variation.

These results confirm that the visual recognition of Copiale glyphs transfers effectively from synthetic to real data. Glyph shapes remain sufficiently consistent for the model to identify individual symbols reliably, and transcription errors alone do not explain the failures observed in end-to-end decipherment.

TABLE 4: TRANSCRIPTION MODEL PERFORMANCE ON ORIGINAL COPIALE MANUSCRIPT

| Metric | Score |
|---|---|
| Token Accuracy | 0.911 |
| Edit Distance | 0.023 |
| WER | 0.017 |
| CER | 0.014 |

### 6.4.2 End-to-End Decryption Performance

In contrast to transcription, end-to-end decryption performance degrades substantially on the authentic manuscript, as summarized in Table 5. Both approaches perform significantly worse than on synthetic data, with absolute accuracies well below practical usability.

The two-stage pipeline achieves 39.6% token accuracy, while Direct Image Decryption reaches 51.4%, corresponding to an absolute improvement of 11.8 percentage points (30% relative). Direct Image Decryption also reduces WER from 89.0% to 76.0% and CER from 43.0% to 39.3%, indicating improved sequence-level coherence and fewer catastrophic decoding failures.

Despite this improvement, both models remain incorrect more often than correct on the Copiale manuscript. These results demonstrate that authentic historical cipher decipherment remains a challenging open problem under current data constraints.

TABLE 5: END-TO-END DECRYPTION COMPARISON ON COPIALE MANUSCRIPT

| Metric | 2-Stage | Direct Decr. | $\Delta$ |
|---|---|---|---|
| Token Acc. | 0.396 | **0.514** | +11.8% |
| Edit Dist. | 0.428 | **0.503** | +17.5% |
| WER | 0.890 | **0.760** | -13.0% |
| CER | 0.430 | **0.393** | -3.7% |

### 6.4.3 Analysis: The Data Scarcity Problem

The performance collapse from 91-92% (synthetic) to 40-51% (real) reveals a fundamental challenge: insufficient training data on real manuscripts. Critically, when we train models on only 20,000 synthetic images instead of 115,000, performance drops to approximately 60%, 8,000 images to 30% —demonstrating that the models require large-scale data to learn robust decipherment, regardless of whether that data is synthetic or real.

This observation reframes our understanding of the synthetic-to-real gap. The problem is not primarily that synthetic data is qualitatively inadequate (although also substantial), but rather that we have **57 times less real data** (2,000 images) than the 115,000 synthetic images used for training. The models trained on synthetic data achieve 91-92% accuracy because they have sufficient examples to learn robust patterns. When applied to real manuscripts, they fail not because the domain is fundamentally different, but because they were never trained on enough real examples.

Several factors contribute to the performance degradation on limited data:

**Insufficient statistical coverage:** With only 2,000 real manuscript images, the models encounter symbol combinations, writing variations, and degradation patterns during testing that were never seen during training. Deep learning models require tens of thousands of examples to generalize robustly. Our 20,000-image experiment confirms this: reducing synthetic training data by 93% (115k $\rightarrow$ 8k) causes a 35+ percentage point accuracy drop (92% $\rightarrow$ 53%), comparable to the synthetic-to-real degradation we observe.

**Compounding error propagation:** When trained on limited data, both the visual feature extraction and language modeling components underfit. Small errors in visual recognition compound with weak language model priors, leading to catastrophic failure cascades. The two-stage pipeline's 39.6% accuracy reflects this: 9% transcription errors (already elevated due to data scarcity) multiply through a decryption model that lacks robust linguistic patterns learned from adequate training examples.

**Linguistic domain specificity:** While the Copiale manuscript's esoteric content (secret society rituals, symbolic terminology) differs from our Faust/Kant/Bible/Nachsommer training corpus, this vocabulary mismatch would be learnable given sufficient real manuscript data. The issue is not that the domains are incompatible, but that 2,000 examples provide insufficient coverage of the Copiale's specific linguistic patterns.

TABLE 6: IMPACT OF TRAINING CORPUS DATASETS AND SCALE ON DECRYPTION PERFORMANCE

| Corpus | Language/Era | Images | Accuracy |
|---|---|---|---|
| Faust | 18th-C German | 115,000 | 92.4% |
| American Psycho | 20th-C English | 20,000 | 53.7% |
| East of Eden | 20th-C English | 8,000 | 31.7% |
| Copiale (Real) | 18th-C German | 2,000 | 51.4% |

**Domain characteristics captured by scale:** Real manuscript characteristics such as ink flow, pressure variations, aging effects, could be learned from data rather than engineered through augmentation, but only with adequate training examples. Synthetic augmentation serves as a proxy when real data is scarce, not as a fundamentally different data source. Here we pose the question of whether generative AI (e.g., GANs) augmentation could obtain higher quality synthetic data?

Despite severe data scarcity, Direct Image Decryption's 11.8% improvement over the two-stage baseline remains meaningful. This advantage persists even when both models are starved for data, validating that end-to-end learning reduces error propagation regardless of training set size. Critically, the advantage is most pronounced on the challenging real-world data: 1.1% improvement on abundant synthetic data (115k images) versus 11.8% improvement on scarce real data (2,000 images). This suggests that end-to-end gradient flow provides robustness benefits that become more valuable precisely when training data is limited.

The core challenge facing automatic historical cipher decipherment is therefore clear: we need more labeled real manuscript data. The 51% token accuracy achieved by Direct Image Decryption represents the performance ceiling achievable with 2,000 training examples. Scaling to 10,000-50,000 real manuscript images—through digitization efforts, or semi-supervised learning techniques—would likely yield the higher accuracy needed for practical deployment. Until such data becomes available or we generate higher quality synthetic samples, automatic decipherment systems serve as tools for augmenting human expert analysis rather than replacing it.

## 6.5 Sequence Length Experiment

How does sequence length affect decryption success? We aim to answer this question through the creation and evaluation of different models trained on various synthetic and real data distributions. To isolate the impact of sequence length, we evaluated the models on different dataset variations.

First, we tested the performance on variations of the "Faust" dataset categorized by length: very short sequences (3–12 characters), normal sequences based on standard Copiale line lengths (12–40 characters), and long sequences (40–70 characters). As illustrated in the green graph in Fig. 11, the decryption success rate remains high for shorter sequences but experiences a significant decline as the text length increases, dropping from nearly 1.00 for the shortest sequences to approximately 0.84 for the 70-character range.

We conducted a parallel experiment using the original Copiale manuscript using the Direct Image Decryption architecture. The purple graph in Fig. 11 displays the success rate across three length tiers: 3–12, 12–40, and 40–70 tokens. Notably, the model struggles significantly with extremely short inputs (below 10 tokens), where success rates hover around 0.35. However, performance peaks and stabilizes once the reference length exceeds 10 tokens, maintaining a high success rate near 0.5 for the majority of standard manuscript line lengths.

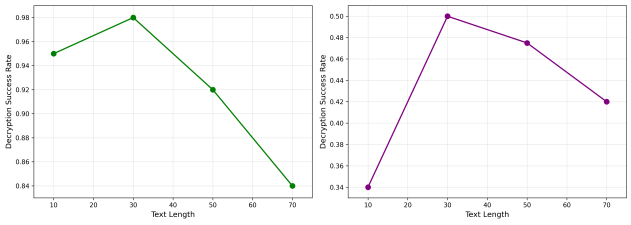



Fig. 11: Decryption success rate relative to sequence length for the synthetic "Faust" dataset (left, green) and the original Copiale manuscript (right, purple)

The divergence between these results suggests that while synthetic models may overfit to specific length patterns found in training, the real manuscript data requires a minimum threshold of linguistic context to achieve reliable decryption. In the synthetic "Faust data", the increased complexity of longer strings drives the error rate up, whereas in the Copiale manuscript, the primary hurdle is the lack of sufficient information in very short snippets.

# 7 CONCLUSION & FUTURE WORK

This research has explored the transition from traditional multi-stage pipelines to end-to-end neural architectures for the decipherment of historical encrypted manuscripts. By introducing the **Direct Image Decryption** approach, we have demonstrated that mapping encrypted manuscript images directly to decrypted plaintext is not only feasible but superior to the traditional transcription-decryption sequence.

Personally, this final degree project has broadened my perspective on the world of AI as a whole. From being able

to work at the CVC, I have gained valuable insights into research and how applicable work is conducted; from work environments to technical discussions. Being surrounded by professionals on diverse topics on this field is enlightening, thus advancing my technical skills. This project covers in technical detail the possible modern improvements applicable to historical artifacts, and it is precisely that intersection which has influenced my future professional interests.

### 7.1 Conclusion

Our experiments lead to several key conclusions regarding automatic cipher decipherment. Direct Image Decryption consistently outperformed the two-stage baseline by a mean of 6% token accuracy on both synthetic and real data, with the advantage becoming more pronounced under challenging conditions—11.8% improvement on the authentic Copiale manuscript versus only 1.1% on abundant synthetic data. This validates our hypothesis that eliminating the intermediate transcription step significantly reduces error propagation, and demonstrates that end-to-end gradient flow provides benefits that become more valuable precisely when training data is limited.

While transcription models generalize effectively to real manuscripts with 91.1% accuracy, decryption performance drops sharply to 51.4%, revealing a fundamental data scarcity paradox. Our analysis proves this is a quantitative rather than qualitative issue as models require tens of thousands of examples to learn robust linguistic patterns, yet we possess 57 times less real data (2,000 images) than synthetic training data (115,000 images). Experiments with reduced synthetic datasets confirm this: dropping from 115,000 to 8,000 images causes comparable performance degradation (92% to 31%) as the synthetic-to-real transfer, demonstrating that the gap stems from insufficient training examples rather than fundamental domain incompatibility. We therefore need data that is plentiful and can capture the visual similarities with the original ciphers.

The consistency of cipher glyphs allows visual recognition models trained on synthetic data to perform reliably on authentic 18th-century handwriting, with transcription accuracy remaining stable across domains. This suggests the primary bottleneck lies in linguistic modeling rather than visual feature extraction. Decryption success is heavily influenced by the visual and linguistic similarity of the training corpus and sequence length, with models struggling on very short sequences that lack sufficient context, or very long sequences that cause attention degradation. These findings underscore that the path forward for automatic historical cipher decipherment lies not in algorithmic innovation alone, but in scaling or mimicking real manuscript data through different approaches. Until such data becomes available, systems like Direct Image Decryption serve as tools for augmenting human expert analysis rather than replacing it, offering a relative improvement in accuracy that can meaningfully reduce the manual effort required for decipherment.

### 7.2 Future Work

While Direct Image Decryption represents a significant step forward, several research avenues remain to bridge the gap between experimental models and practical tools. In lack of large quantities of precious original data: we are only as good as our *"synthetically"* generated data. Future work should explore Generative Adversarial Networks (GANs) for synthetic data generation, training generators to produce manuscript images that capture subtle characteristics like ink flow variations and aging patterns that current augmentation approximates but does not fully replicate.

The Direct Image Decryption architecture should be evaluated on other historical ciphers, such as the Borg or Rammanacoil manuscripts, to determine whether end-to-end mapping advantages generalize beyond homophonic substitution ciphers. Given the scarcity of labeled data, semi-supervised learning techniques could allow models to learn from digitized, untranscribed manuscripts, while human-in-the-loop systems integrating expert feedback would refine linguistic priors and create high-quality training data through interactive correction, accelerating decipherment of unsolved historical manuscripts.

## ACKNOWLEDGEMENTS

I cannot express enough thanks to Dr. Alicia Fornés for her continued support, encouragement, and guidance with this project, for enabling me to embark on this journey into the world of research and cryptography with an always pragmatic and enlightening demeanor. I would also like to extend my gratitude to Dr. Josep Lladós for his invaluable help in providing the opportunities that made this research possible, and to Dr. Ernest Valveny for his steadfast determination in improving our academic experience.

I offer my sincere appreciation for the learning opportunities provided by the Computer Vision Center (CVC) and its research fellows, who by sharing desks and papers have shaped my understanding and appreciation of this field.

This work builds upon the foundational research of the DECRYPT project and leverages datasets from the ICDAR Competition. Special thanks to Jialou Chen, Lei Kang, Pau Torras, and Marçal Rusiñol for making their CRNN implementation and augmentation code available, which proved essential to this research.

A version of this work has been submitted to the International Conference on Historical Cryptology (HistoCrypt 2026) and is currently under review.

My completion of this project could not have been accomplished without the support of my classmates Alejandra Reinares, Eric López, Amritpal Singh, and Amelia Gómez. To my friends Pere, Beto, and Kike; my brother Gabriel; and my parents, Marino and Lourdes—thank you for supporting this educational journey in ways I could have never thought possible.

Finally, I acknowledge the Universitat Autònoma de Barcelona and the Escola d'Enginyeria for their institutional support throughout this final degree project.

## REFERENCES

[1] N. Aldarrab and J. May, "Can Sequence-to-Sequence Models Crack Substitution Ciphers?" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[2] N. Aldarrab, "Automatic Decipherment of Historical Manuscripts," Ph.D. dissertation, University of Southern California, 2022.

[3] J. Gillogly, "A Dissenting Opinion: The Beale Ciphers," *Cryptologia*, vol. 4, no. 2, pp. 116–119, 1980.

[4] T. Bluche, J. Louradour, and R. Messina, "Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention," in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1050–1055.

[5] R. Clemens, "The Voynich Manuscript," *Beinecke Rare Book and Manuscript Library Digital Collections*, Yale University, 2016. Available: `https://brbldl.library.yale.edu/vufind/Record/3519597`

[6] "The DECRYPT Project." Available: `https://de-crypt.org/`

[7] J. Dinnissen and N. Kopal, "Island Ramanacoil a Bridge too Far: A Dutch Ciphertext from 1674," in *Proceedings of the 4th International Conference on Historical Cryptology (HistoCrypt)*, 2021, pp. 48–57.

[8] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[9] N. Kambhatla, A. Mansouri Bigvand, and A. Sarkar, "Decipherment of Substitution Ciphers with Neural Language Models," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 869–874.

[10] K. Knight, B. Megyesi, and C. Schaefer, "The Copiale Cipher," in *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, 2011.

[11] G. Lasry, N. Kopal, and A. Wacker, "Codebreaking Zodiac," *Cryptologia*, 2021.

[12] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[13] "ICDAR Robust Reading Competition: Historical Ciphers Challenge, 2024. Available: `https://rrc.cvc.uab.cat/?ch=27`

[14] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[15] X. Yin, N. Aldarrab, B. Megyesi, and K. Knight, "Decipherment of Historical Manuscript Images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, 2019.

# APPENDIX

## A PROJECT TIMELINE

The project timeline is illustrated in the Gantt Chart shown in Figure 12. The project was divided into three main phases corresponding to the follow-up sessions of the project:

- **Part 1 (Light Blue):** Project kick-off, state-of-the-art research, environment setup, and synthetic data generation. This initial phase established the foundations for the project, including literature review, identification of the core challenges in historical cipher decipherment, and development of the synthetic data generation pipeline.
- **Part 2 (Medium Blue):** Model creation for both main pipelines—Transcription-Decryption and Direct Image Decryption. This phase involved implementing the CRNN-based transcription model, developing the decryption component, and designing the end-to-end Direct Image Decryption architecture. Training and initial evaluation of all models were completed during this period.
- **Part 3 (Dark Blue/Purple):** Testing, comparison between pipelines, conclusions, report writing, and model improvements. This final phase focused on comprehensive evaluation of both approaches, analyzing their relative strengths and weaknesses, and documenting the findings in this report.

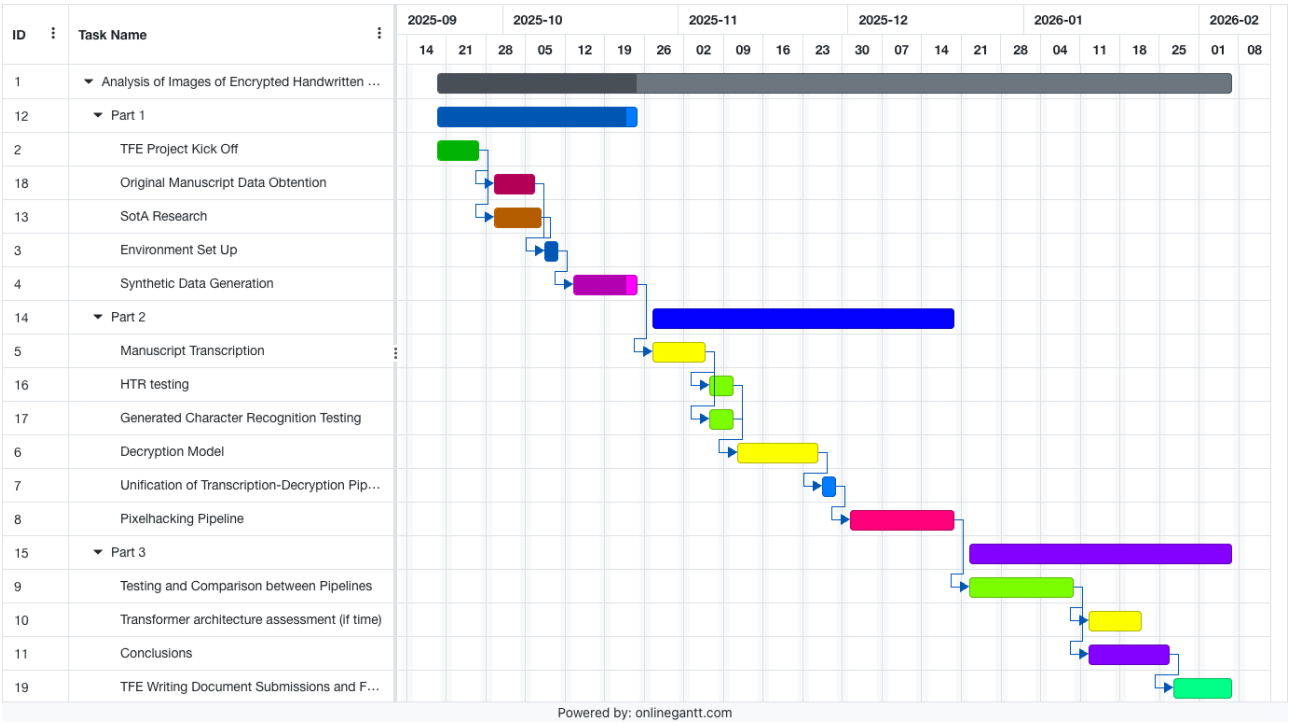



Fig. 12: Gantt Chart

The timeline spans from the project initiation in September 2025 through final documentation and model refinement in January 2026, with careful attention to meeting the deliverable dates for each follow-up session. Weekly meetings with Alicia Fornés, the project supervisor, ensured consistent progress and allowed for timely adjustments to the research direction based on intermediate results.