Recent Advances in Information Extraction from Historical Archival Records

Arthur Matei[1,2][0009−0009−6028−7502], Tim Hallyburton[1][0009−0008−0944−9193], Lukas Hennies[3][0000−0003−1516−8183], Christoph Rass[3][0000−0001−9492−907X], and Gernot A. Fink[1,2][0000−0002−7446−7813]

[1] TU Dortmund University, Department of Computer Science, Dortmund, Germany
{arthur.matei,tim.hallyburton,gernot.fink}@tu-dortmund.de
[2] LAMARR, Institute of Machine Learning and Artificial Intelligence, Dortmund, Germany
[3] Osnabrück University, Chair of Modern History and Historical Migration Studies, Osnabrück, Germany
{lukas.hennies,christoph.rass}@uni-osnabrueck.de

Abstract. Extracting structured information from historical archival documents is challenging due to the scarce availability of labeled data and the oftentimes convoluted information on the documents. We build on the work of Wolf et al. [26] to further investigate field extraction from CM/1, a dataset of historical WWII care and maintenance application forms, using vision-language models (PaliGemma and Donut) trained on as little as 1% of available annotations. To address data scarcity, we explore two strategies: cross-field pre-training on the same documents and synthetic document generation, where empty document templates are populated with synthesized handwriting. We show pre-training on synthetic documents yields small improvements, but cross-field pre-training shows greater improvements in low-data training. We further find that joint multi-field training slightly benefits larger models, while smaller models perform better when trained on single fields in low-data scenarios. Although our approach can be applied to any field in the forms, we exemplarily demonstrate practical relevance through nationality extraction, a field of high interest for historians in migration studies. Alongside this, we publish CM/1v2, an extended version of the dataset with more annotated fields, namely Nationality, Place of Birth, and Religion, covering both heads of family and family members. Notably, the Religion field poses a unique recognition challenge as it combines pre-printed checkboxes with free-text entries added by applicants. The dataset is available at github.com/rthrmt/cm1v2.

1 Introduction

When large archival collections are digitized, archives typically annotate a small number of key fields, like names and dates, to index the material. However, researchers' questions often require fields far beyond what was annotated for indexing. Annotating thousands of documents from scratch for every such field

is prohibitively expensive, yet a model that already extracts names has seen exactly the same pages, the same handwriting, and the same layouts that the new field appears on. The central question is whether this knowledge transfers: can a model trained on one set of fields bootstrap extraction of a different field with minimal additional annotation? We study this question on historical application forms for care and maintenance from the Arolsen Archives[4], using the CM/1 benchmark introduced by Wolf et al. [26].

The rise of transformer-based end-to-end models [13,6,14] has made this kind of investigation practical [26]. Unlike classical pipelines that chain layout analysis, segmentation, and OCR, end-to-end vision-language models learn directly from document images and field-level labels, making them a natural fit for archival settings where layout or segmentation ground truth is unavailable.

We investigate two strategies for bootstrapping a new field under low-resource conditions. In *cross-field pre-training*, a model is first trained to extract already annotated fields (e.g., Name and Place of Birth) and then fine-tuned on the target field with limited labeled data. In *synthetic pre-training*, the model is instead initialized on synthetically generated full-page documents with handwriting rendered by a GAN [9]. Both strategies are illustrated in Figure 1. We also examine whether training exclusively on one field outperforms training a *multi-field model* jointly.

To check whether these strategies behave differently across model scales, we compare a small-scale and a large-scale model: **Donut** [13], a lightweight encoder-decoder with 201 M parameters, and **PaliGemma** [2], a large vision-language model with 3 B parameters.

Although applicable to any field, we design our experiments around a concrete research question from migration studies: extracting the nationality of applicants.

From a historical perspective, extracting attributes such as 'nationality' from the CM/1 application documents is central to analyzing the dynamics of forced migration and the functioning of the early international refugee regime [11]. In the postwar period, 'nationality' was not a static identity marker but rather a bureaucratic category of action - deployed both by institutions to regulate mobility and by those categorized as 'displaced persons' to negotiate their own status within that very regime [12]. Methodologically, however, working with such data requires caution: researchers should be aware that the information recorded in these files is often the product of complex negotiation processes and that national categories may reflect political intentions rather than social realities [5]. The data extracted from these documents cannot, in other words, be understood in a positivistic sense as objective records of fact; rather, they constitute traces of situated encounters between applicants, caseworkers, and institutional frameworks, each shaped by divergent interests and shifting political conditions [11,12].

The CM/1 files preserved in the Arolsen Archives clearly document this dynamic. Applicants frequently 'interpreted' their nationality strategically to se-

[4] https://collections.arolsen-archives.org/en/archive/3-2-1

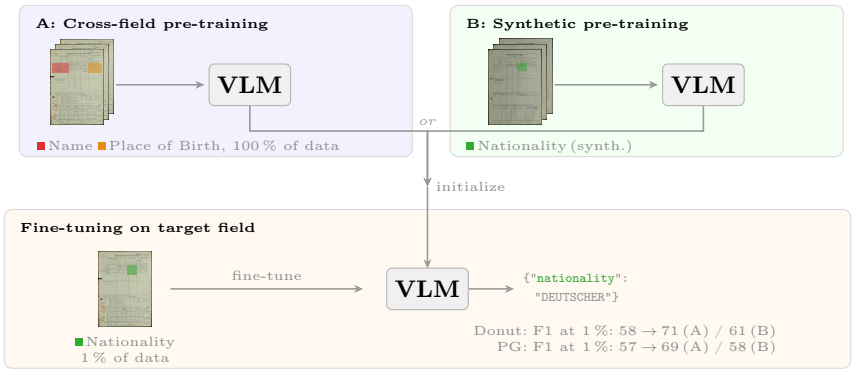

Fig. 1. Overview of the proposed knowledge-transfer strategies. **(A)** Cross-field pre-training uses a model trained on existing fields (Name, Place of Birth) with abundant annotations. **(B)** Synthetic pre-training uses empty document templates filled with GAN-generated handwriting. Both strategies initialize a model that is then fine-tuned on limited target-field data (here: Nationality at 1 %).

cure admission to UNRRA or IRO programs. Many Soviet DPs of Turkic origin — for example Uzbeks, Kazakhs, or Turkmen — declared themselves Turkish citizens to avoid forced repatriation to the Soviet Union; other groups, such as Jewish DPs, used 'Jewish' as a national container to support their migration goals [25]. At the same time, the categories inscribed in these documents were never solely the product of applicant self-representation; they emerged from interactions between applicants, caseworkers, institutional mandates, and shifting geopolitical pressures [11].

Automated extraction of such fields takes on a transformative role precisely because it enables a 'bird's-eye view' of these large-scale processes [20]. By transforming millions of documents into machine-readable datasets, researchers can systematically link micro-historical biographies with macro-historical movement patterns [3]. This, in turn, opens up central research questions that have remained difficult to address at scale: How have DPs adapted their narratives to shifting bureaucratic mandates [11]? To what extent did subjective or racist prejudices of officials shape recognition practices [25]? Analyzing these data makes it possible to understand those referred to as 'displaced persons' not as passive objects of institutional processing but as purposeful actors who shaped the international migration regime from below [20].

Lastly, we also publish **CM/1 v2**, an extended version of the dataset with annotations for three additional fields: Nationality, Place of Birth, and Religion, for both heads of family and family members. These fields each come with distinct recognition challenges. Nationality requires producing normalized German demonyms rather than literal transcriptions, Place of Birth demands extracting only the town while ignoring co-located province and country, and Religion mixes pre-printed checkboxes with free-text entries that break the expected form layout. Although CM/1 v2 contains a substantial amount of annotated data, this is

exceptional in archival research. Our experiments therefore focus on low-resource settings, using as little as 1% of the available training data (roughly 900 samples), a realistic scale for historians approaching a new extraction task.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, models, and training strategies. Section 4 presents our experimental evaluation. Section 5 summarizes our findings.

2 Related Work

Our work builds on three research directions: the application of large vision-language models to handwritten and historical text recognition, knowledge transfer strategies for adapting such models to new tasks with limited supervision, and synthetic data generation for document analysis.

2.1 LVLMs for Handwritten Text Recognition

End-to-end vision-language models have become a dominant paradigm for document understanding [13,2]. Unlike modular pipelines with separate stages for layout analysis, text detection, and OCR, LVLMs directly map document images to structured text outputs, making them attractive when layout or segmentation annotations are unavailable.

Crosilla et al. [4] recently benchmarked both proprietary and open-source multimodal LLMs against the established Transkribus platform on modern and historical handwriting datasets in English, French, German, and Italian. Their findings highlight two issues that are directly relevant to CM/1: first, proprietary models such as Claude 3.5 Sonnet substantially outperform open-source alternatives in zero-shot settings, but raise privacy concerns when applied to sensitive archival material. Second, all models exhibit a strong language bias toward English due to pre-training data composition, which poses a challenge for multilingual collections like CM/1, where form fields may be printed in English or German while applicants wrote in their native language. The study also found that LLMs have limited ability to autonomously correct their own transcription errors, suggesting that fine-tuning remains essential for high-accuracy extraction on domain-specific documents.

2.2 Transfer Learning for Document Understanding

Transfer learning has long been recognized as a key strategy for improving performance in low-resource settings by leveraging knowledge acquired on related tasks or domains [31]. In document analysis, the most common form is initializing from large-scale pre-training before fine-tuning on a downstream task, as exemplified by LayoutLM [27] and LayoutXLM [28], which combine text, layout, and image modalities.

A conceptually related but less explored direction is *within-document* or *cross-field* transfer, where a model trained on one set of fields in the same collection is repurposed to extract a different field. This arises naturally when an

institution already has a model for common indexing fields and wants to bootstrap a new field without additional annotation. Because source and target fields share the same documents, writers, and layouts, the learned representations are directly reusable, potentially offering stronger initialization than transfer from an unrelated dataset.

Recent work on task transfer in vision-language models shows that fine-tuning on one perception task can both help and hurt performance on others, depending on task similarity [21]. In the document domain, Singh et al. [22] demonstrated that even simple transfer from synthetic to real data with as few as ten documents yields competitive layout understanding.

2.3 Synthetic Data for Document Analysis

When real annotated data is scarce, synthetic data generation offers a way to produce large training corpora with automatic ground truth. In the document analysis community, this idea has been pursued along two main axes: style transfer approaches that adapt the appearance of existing or template-based documents, and code-generation frameworks that render documents programmatically.

Style transfer and GAN-based approaches. Vögtlin et al. [24] proposed a two-step pipeline: template documents are first created with specified content and structure, then CycleGAN-based style transfer using unpaired real historical images makes them visually realistic. Pre-training on such synthetic documents outperformed both random initialization and transfer from real handwriting datasets like IAM [17]. At the word and line level, GAN-based handwriting synthesis has advanced considerably. HiGAN+ [9] achieves realistic one-shot style transfer by disentangling textual content from calligraphic style through writer-specific auxiliary losses and local patch refinement. A recent comprehensive survey by Diaz et al. [7] covers the full spectrum of handwriting synthesis methods from 2019 to 2024, noting that GAN-based models, including ScrabbleGAN, HiGAN+, and variants, now produce word images realistic enough to improve downstream HTR when used for data augmentation. Vidal-Gorène et al. [23] evaluated ScrabbleGAN and CycleGAN for line-level synthesis across Latin, medieval French, Armenian, and Arabic scripts, confirming the viability of GAN-generated data for out-of-domain historical documents.

Code-generation and template-based synthesis. On the programmatic side, Yang et al. [29] introduced CoSyn, where a text-only LLM generates code (Python, HTML, LaTeX) to render diverse text-rich images such as charts, tables, and documents. Models trained on the resulting 400K synthetic images achieved state-of-the-art performance on multiple document understanding benchmarks. Archibald and Martinez [1] proposed the DELINE8K pipeline, which integrates preprinted text, handwriting, and document backgrounds to generate semantically segmented training pages for form classification and layout segmentation.

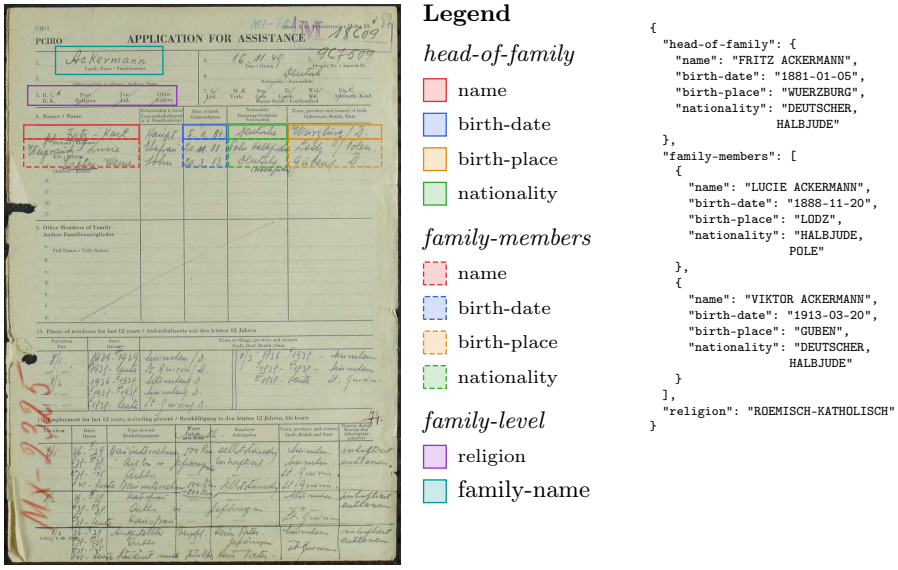

Fig. 2. Example CM/1 application form with annotated fields. Head-of-family fields are marked with solid borders, family-member fields with dashed borders. Fields of the same type share the same color. The corresponding JSON annotation is shown on the right.

Relation to our approach. Our synthesis pipeline combines elements of both directions: like Vögtlin et al., we render text onto manually annotated document templates with known field positions; like CoSyn, synthesis is fully automated through a pipeline that samples field values and writer styles. Field values are drawn from the real annotations to guarantee realistic content, but we do not enforce logical consistency between fields, as our goal is to improve downstream extraction rather than to model the joint distribution of field values. Instead of style transfer, we use HiGAN+ [9] trained on IAM to generate handwritten field images pasted onto empty form templates. This approach preserves exact ground truth while introducing realistic stylistic variation, yielding fully reproducible synthetic training data with pixel-perfect annotations for all fields.

3 Information Extraction from Historical Documents

3.1 CM/1 v2

While the original fields in CM/1 [26] are useful for indexing, they are insufficient for concrete research questions in, e.g., migration studies. We therefore extend the dataset with three additional fields of high interest to historians: Nationality, Place of Birth, and Religion, forming CM/1 v2.

Table 1. Number of samples per dataset split, including training subsets, validation, and test sets, following *CM1-NAME* and *CM1-DATE* by Wolf et al. [26]

Dataset	Train				Val	Test	Total
	1%	10%	25%	100%			
CM/1v2	904	9 042	22 606	90 427	5 000	10 000	105 427

Fields in CM/1 applications fall into two categories: family-level and per-member. Religion applies to the whole family, while the remaining fields are per member. The newly included fields are described below:

Nationality. Applicants wrote their nationality in varying languages depending on the form language and their proficiency. Target annotations, however, are given as German masculine demonyms, e.g. *Pole* or *Deutscher*. Annotations are thus normalized to German demonyms, which aids retrieval but prevents learning a classical text recognition model. When applicants stated multiple nationalities, we sort them alphabetically in the annotations to reduce ambiguity.

Place of Birth. Applications ask for town, province, and country of birth, but the annotations provided by Arolsen Archives only include the town. A model must therefore learn to extract the town while ignoring province and country, which occupy the same crowded text region.

Religion. Unlike the previous free-text fields, Religion is family-level and typically answered by checking a box from a list of common religions (catholic, protestant, jewish, etc.). If an applicant's religion was not listed, they wrote it by hand next to the pre-printed options. The model must therefore handle both checked boxes and free-text entries that break the expected form layout.

Figure 2 shows an example with bounding boxes illustrating the field locations alongside annotations in JSON format. All annotations are upper-cased; training and evaluation are therefore case-insensitive. Fields absent from a form type or left blank are annotated as *null*.

CM/1 v2 shares the same document collection as CM/1 [26] and focuses exclusively on the first page of each application, analogous to the CM1-NAME and CM1-DATE benchmarks. Cover pages are omitted as their information is already captured by CM1-COVER. Train, validation, and test splits are identical to the original. Table 1 gives an overview of the number of samples in each split.

3.2 Document Synthesis

To generate training data for the new fields, we synthesize full-page samples based on empty document templates for each document type. Templates were obtained by manually removing all handwritten content from real scans using

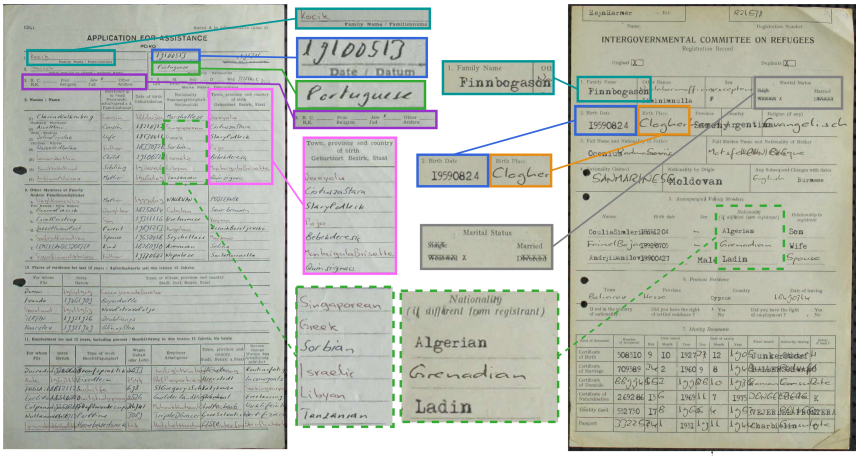

Fig. 3. Example of synthetically generated full-page documents. Handwriting is rendered by HiGAN+ while typewriter styles shown in the right image are created using image fonts. Word images are pasted onto an empty document template. All fields are synthesized individually, we show color-coded bounding boxes for the fields of interest in this work. The corresponding field values are stored as metadata for training and evaluation.

standard image manipulation tools, leaving only the printed layout and static text. Text regions in these templates were manually annotated and associated with unique field labels and bounding box coordinates. Field values are sampled uniformly from a database of annotated field values as transcribed from the original documents with sensible extensions for low-resource fields. The data synthesis pipeline handles the generation of field-level data, based on the type of field, e.g., date-fields, name-fields, location-fields, and stores the sampled values for each field along with associated bounding boxes as metadata. We utilize HiGAN+ [9], a handwriting generation model trained on the IAM dataset [17]. Furthermore, we employ font-based rendering using a variety of handwriting and typeface fonts. This combination captures various handwriting styles and simulates typewritten fields, which are prevalent in CM/1 for fields filled out by clerks or officers rather than applicants. For each document, distinct writers are sampled and associated with a style vector for GAN-based rendering. Fields are rendered in each writer's style and pasted onto the template at the corresponding bounding box locations. The resulting image is compressed and artificially degraded to match the quality of the original documents. Figure 3 shows an example of a synthesized document.

3.3 Vision-Language Models

We consider two open-weight end-to-end models at different scales: Donut as a lightweight model suitable for consumer-grade hardware, and PaliGemma as a larger model with extensive multi-stage pre-training.

Donut. Donut [13] follows an encoder-decoder design, with roughly 201 M parameters, tailored for end-to-end document understanding tasks such as classification, visual question answering, and information parsing. On the encoder side, a Swin Transformer [16] processes the input image through hierarchical stages with shifted windows, progressively building a rich spatial representation that covers both fine-grained details and broader layout structure. A BART-based [15] decoder consumes these visual features and produces structured output tokens in an XML-like syntax, directly mappable to JSON key-value pairs. During pre-training, the model is exposed to a mixture of real documents with machine-generated OCR labels and fully synthetic pages, giving it broad coverage of different scripts, fonts, and page layouts.

PaliGemma. PaliGemma3B [2] is a multimodal model that connects a SigLIP-So400m image encoder [30] to the Gemma-2B language model [19], yielding approximately 3 B parameters in total. Input images are processed by a Vision Transformer (ViT) [8] whose output embeddings are linearly mapped into the token space of the language model, so that visual and textual information can be jointly attended to during generation. We use the checkpoint additionally fine-tuned on DocVQA [18] at the maximum supported resolution of 896 px to preserve fine details in document scans. Despite being an order of magnitude larger than Donut, PaliGemma remains trainable on a single high-end GPU, and its richer pre-training provides stronger generalization in low-data regimes.

Fine-Tuning with LoRA. While Donut can be fully fine-tuned, PaliGemma requires a parameter-efficient approach, for which we employ Low-Rank Adaptation (LoRA) [10]. Instead of updating all model weights, LoRA introduces small, trainable low-rank matrices parallel to the existing layers while keeping the original parameters frozen. Computation is performed by the original large weights as well as the newly introduced low-rank matrices in parallel. Afterwards, the output of both is simply added, giving the smaller low-rank matrices the ability to steer the computation of their corresponding larger weight matrix without the need to update the large weights themselves. This reduces the number of trainable parameters by orders of magnitude compared to full fine-tuning, substantially lowering computational and memory demands [10]. For example, in this work instead of updating all 3 B parameters of PaliGemma, we only need to train around 5.1 M parameters. LoRA has been shown to achieve performance comparable to full fine-tuning across a variety of tasks. In our setup, LoRA is applied to all linear layers of both the encoder and decoder components.

3.4 Cross-field Pre-training and Pre-training on Synthetic Documents

In many practical archival settings, annotated data for a newly requested field is scarce, while annotations for other fields on the same documents may already exist. We explore two strategies that tackle this problem from different angles: cross-field pre-training on the same documents, leveraging information in already indexed files, and pre-training on abundant synthetic documents.

Cross-Field Pre-training. Since each CM/1 page contains multiple annotated fields, a model trained on one field has already learned to parse the layout and segment handwriting from the background. Cross-field pre-training exploits this: source and target share the very same documents and writers, so the model only needs to redirect its reading toward a new page region and output vocabulary. Concretely, we first train on the full annotations for one or more source fields and then fine-tune on the target field with limited data.

Pre-training on Synthetic Documents. Alternatively, synthetically generated documents offer another source of pre-training data. Using the pipeline described in section 3.2, we generate full-page document images that closely resemble real scans in layout and visual complexity, so the model already learns to locate the target field within a realistic page structure during pre-training. Because subsequent fine-tuning targets the same field, only the writing style and data distribution shift when transitioning to real data.

4 Experiments

We design our experiments around the scenario that a user wants to extract a new field from an existing document collection, potentially having a model already trained on common indexing fields like names. We target the Nationality field as a realistic basis for research questions in migration studies. Since historians typically have far less annotated data than CM/1 offers, our fine-tuning experiments consider only the 1%, 10%, and 25% splits, where 1% (roughly 900 samples) represents a realistic annotation budget. We additionally consider synthesizing data for pre-training to reduce the need for real annotations. All experiments focus on head-of-family fields to keep the scope manageable.

4.1 Metrics

We measure Character Error Rate (CER) and F1 score. Model outputs are parsed and flattened into key-value pairs (e.g., `{"name": "Mustermann"}` becomes `<name, Mustermann>`); metrics are computed on the values, not the keys. For multi-field extraction, we compute them per pair.

CER measures the character-level edit distance between prediction and annotation, normalized by annotation length; lower is better.

The F1 score measures extraction accuracy at the field level, treating each key-value pair in the flattened JSON as one individual extraction result. The F1 score is defined by the harmonic mean between precision P and recall R:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}, \quad P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{1}$$

A prediction is counted as a *true positive* (TP) if the field key is present in both the prediction and the ground truth and the predicted value matches the ground truth value exactly. A prediction is a *false positive* (FP) if the field

key is absent from the ground truth, or if the key is present but the value does not match. A *false negative* (FN) occurs when a field key present in the ground truth is missing from the prediction. If the model output cannot be parsed as valid JSON, all ground truth fields are counted as false negatives. In single-field extraction, this is equal to accuracy.

While CER captures fine-grained transcription quality and is sensitive to small character-level improvements, F1 reflects whether the model correctly identifies and extracts complete field values. We use CER as our primary metric for early stopping during training, as it provides a more locally sensitive signal, while F1 is the measure to assess end-to-end extraction quality. Both metrics are reported multiplied by 100 for readability.

4.2 Training Setup

For our experiments we tried to stay as close as possible to the training setup of Wolf et al. [26] to ensure comparability. All models were trained on a single NVIDIA A100 GPU with 80GB of VRAM. We use AdamW with a learning rate of 1×10^{-4} for PaliGemma and 3×10^{-5} for Donut, combined with cosine annealing. Donut additionally uses linear warmup over the first 10% of steps. The batch size is 1 for PaliGemma and 4 for Donut, both with gradient accumulation of 4. Training runs for at most 10 epochs with early stopping after 3 validations without CER improvement; most models converged after around 5 epochs. We chose CER over F1 for early stopping as it is more sensitive to small improvements and represents a more local error. Validation is performed every epoch for the 1–25% splits and every quarter epoch for 100%. The best checkpoint by validation CER is used for testing. Because of its size, PaliGemma is trained with LoRA, while Donut is fully fine-tuned. We used a LoRA rank of 8 and an alpha of 16 for all linear layers of the model. PaliGemma's input resolution is fixed to 896px by its architecture, while Donut is trained with an input resolution of 2560×1920px. Percent-based splits, i.e., 1%, 10% and 25% of data are sampled deterministically, guaranteeing comparability between models trained on the same subset. PaliGemma is steered via the prompt template `Extract {field} from this document in JSON format.`, where `{field}` is replaced with the target field name. For multi-field extraction, field names are concatenated with a comma. The model outputs JSON directly. Donut requires no prompt and always outputs all fields it was trained on in its custom XML-like format, which is converted to JSON via its `token2json` function.

4.3 Results

We present our results in three major sections. First, we present baselines for our newly introduced fields as well as some baselines for models that were trained on multi-field extraction. Then, we answer the question, whether a user that is only interested in a single field should train for that field individually or if a multi-field model achieves a better performance. Lastly, we show that starting with a model that is already able to extract other fields from the dataset gives a

Table 2. Comparison of trainig to extract different field combinations from CM/1 applications. Training and test is performed on the same fields. Scores are multiplied by 100.

Model	Fields	1%		10%		25%		100%	
		CER ↓	F1 ↑	CER ↓	F1 ↑	CER ↓	F1 ↑	CER ↓	F1 ↑
PaliGemma	Name	33.71	18.64	21.63	29.83	17.09	38.74	10.64	56.69
	Nationality (Nat)	36.90	57.03	23.98	75.38	19.91	78.54	16.62	82.31
	Place of Birth (PoB)	59.28	15.39	45.59	28.87	40.23	30.62	33.18	40.05
	Religion	47.54	58.84	27.31	78.50	20.39	82.82	18.85	84.47
	Name,PoB	38.03	22.75	27.17	34.78	22.90	43.45	17.12	55.81
	Name,Nat,PoB	38.62	42.68	25.82	56.19	21.49	62.84	16.46	70.63
Donut	Name	51.22	8.01	26.61	31.56	17.79	44.79	9.90	61.96
	Nationality (Nat)	40.73	57.64	23.79	76.24	20.63	79.54	15.60	84.06
	Place of Birth (PoB)	56.56	18.49	41.20	31.83	36.18	38.20	29.86	46.07
	Religion	67.87	44.80	29.53	78.62	20.44	83.93	16.48	86.14
	Name,PoB	51.22	16.17	26.85	40.76	21.57	51.09	15.53	61.67
	Name,Nat,PoB	50.47	35.39	25.17	60.41	20.46	66.66	14.83	74.43

clear advantage and significantly boosts performance in very low data scenarios. We also compare this in-data transfer learning with fine-tuning from a model pre-trained on synthetic data.

Baselines Table 2 presents single-field and multi-field baselines for the newly introduced fields, using the same data splits as Wolf et al. [26], where 100% corresponds to approximately 90K samples.

Contrary to Wolf et al. [26], Donut does not exhibit severely degraded performance at 1% and 10% of data. Both models achieve non-trivial results even at 1%, where they are roughly comparable: PaliGemma has a slight edge for Name, Religion, and multi-field settings, while Donut performs marginally better on Nationality and Place of Birth. From 10% onward, Donut consistently overtakes PaliGemma across all single fields, despite being more than an order of magnitude smaller. The same pattern holds for multi-field extraction: PaliGemma leads at 1% but Donut surpasses it with more data available.

Regarding field difficulty, Nationality and Religion are the easiest fields for both models. For Nationality, the normalization to German masculine demonyms turns the task from open-ended transcription into something closer to classification over a limited vocabulary. Interestingly, even the much smaller Donut handles this implicit translation well. Religion benefits similarly from a small set of possible values, and both models handle its dual nature of pre-printed checkboxes and free-text off-layout entries well. Place of Birth is the most challenging field. Applicants filled in town, province, and country, but only the town is annotated, so the model must learn to ignore the surrounding information. The visually crowded layout of this region further complicates extraction.

Table 3. Inference on fields defined by the **Fields** column, but F1 score calculated only for the Nationality field. Higher is better.

Model	Fields	1%	10%	25%	100%
PG	Nat (baseline)	57.03	75.38	78.54	82.31
	Name,Nat,PoB	60.61	74.63	80.57	84.20
D.	Nat (baseline)	57.64	76.24	79.54	84.06
	Name,Nat,PoB	51.76	75.61	79.87	84.11

Multi-Field vs. Single-Field Learning Table 3 asks whether a user interested only in Nationality is better served by single-field training or by a joint multi-field model. As an example, we train on Name, Nationality, and Place of Birth jointly and evaluate F1 for Nationality only in both cases.

With all training data available, joint multi-field training is marginally better for both models: PaliGemma improves from 82.31 to 84.20 F1 and Donut from 84.06 to 84.11. The joint training objective appears to provide some complementary document understanding that benefits Nationality extraction even when it is not the only training target.

In low-data regimes the two models behave differently. For Donut, single-field training is clearly preferable at 1% (57.64 vs. 51.76 F1), but the gap closes quickly and the two approaches are essentially equivalent from 25% onward. PaliGemma, by contrast, already benefits from multi-field training at 1% (60.61 vs. 57.03 F1) and maintains a slight advantage at higher data percentages. PaliGemma's stronger pre-training apparently allows it to exploit the joint objective even with very limited data, whereas the added task complexity initially hurts Donut in the 1% regime.

In summary, the results suggest that single-field training tends to be the safer choice for smaller models, while larger models with richer pre-training may already benefit from joint multi-field training even in low-resource settings.

Cross-field Pre-training for Nationality Extraction Since real annotated data for a new field is typically scarce, we focus this experiment on the 1%, 10%, and 25% splits. Having access to 90K fully annotated samples, as CM/1 v2 provides, is a fortunate exception rather than the norm and most humanities and social science projects operate with far fewer labeled documents. The 100% split is therefore excluded here as it represents an unrealistic upper bound for the target application scenario.

Table 4 compares two pre-training strategies against the Nationality baseline. In cross-field pre-training, a model is first trained on the full Name and Place of Birth data (100% split) and then fine-tuned on Nationality with limited real data. In synthetic pre-training, the model is pre-trained to extract the Nationality on 500K synthetically generated documents for one epoch before fine-tuning on real data. The synthetic sample count is chosen so that one epoch yields roughly the same number of gradient updates as five epochs on 100% of real data, the

Table 4. F1 scores on the test split of CM/1v2 after fine-tuning for Nationality with prior training on other CM1 fields defined in the **Fields** column first. Higher is better.

Model	Fields	1%	10%	25%
PG	Nat (baseline)	57.03	75.38	78.54
	Nat synthetic	57.91	74.21	79.62
	Name,PoB	68.80	78.53	81.00
Donut	Nat (baseline)	57.64	76.24	79.54
	Nat synthetic	61.48	76.20	80.65
	Name,PoB	70.75	79.03	80.73

typical convergence point observed in our baseline experiments, making both pre-training budgets comparable.

Cross-field pre-training yields substantial gains in the 1% regime for both models. PaliGemma improves from 57.03 to 68.80 F1 (+11.8 points) and Donut from 57.64 to 70.75 F1 (+13.1 points). Name and Place of Birth expose the model to the same documents, writers, and handwriting styles as Nationality, giving it a strong initialization for the target task without requiring any Nationality-specific annotations. As more Nationality data becomes available, the advantage diminishes but remains positive: at 25% Donut still gains 1.2 F1 points over the baseline. Notably, with cross-field pre-training and only 25% of Nationality annotations, both models come within 1–3 F1 points of their respective 100% single-field baselines (Table 2). However, at 25% most of the improvement is already attributable to the additional target data rather than the pre-training itself.

Synthetic pre-training shows a more modest effect. For Donut, it improves the 1% baseline from 57.64 to 61.48 F1 (+3.8 points) and yields small gains at 25%, while leaving performance roughly on par with the baseline at 10%. For PaliGemma, gains are marginal at 1% (57.03 to 57.91) and 25% (78.54 to 79.62), while at 10% synthetic pre-training slightly underperforms the baseline (74.21 vs. 75.38). In both cases, cross-field pre-training is substantially more effective, particularly at 1% where it outperforms the synthetic variant by 9–11 F1 points.

4.4 Discussion

A recurring finding across our experiments is that Donut, despite being more than an order of magnitude smaller than PaliGemma, matches or surpasses it from 10% of data onward. This suggests that for information extraction with moderate training data, a lightweight fully fine-tuned model can be preferable to a larger LoRA-adapted one in both accuracy and computational cost. Only at 1% does PaliGemma's richer pre-training provide an advantage on most fields.

Our single-field vs. multi-field comparison reveals a model-size-dependent trade-off in low-data regimes: Donut benefits from focused single-field training, while PaliGemma can absorb the additional complexity of joint extraction

even with limited data. With sufficient annotations, the gap narrows for Donut, whereas PaliGemma retains a small gain of around 2 F1 points from multi-field training.

Regarding knowledge transfer, cross-field pre-training is clearly the more effective strategy, outperforming synthetic pre-training by a wide margin at 1%. The gap between the two likely has multiple causes. Cross-field models see real document images with authentic handwriting, while the synthetic pipeline relies on GAN-generated writing that may not fully capture the variability and degradation of the original scans. Perhaps more importantly, the synthetic pipeline samples nationalities uniformly, whereas the true distribution is heavily skewed, giving the model a weaker prior for the frequent classes that dominate the test set. Correcting this by estimating the true distribution from real data is circular, as the distribution of nationalities is precisely what a historian would want to discover, not assume beforehand.

5 Conclusion

We investigated how to adapt vision-language models to extract new fields from historical document collections with minimal annotation effort, using nationality extraction from WWII compensation forms as a concrete use case driven by migration studies research.

Our experiments show that cross-field pre-training is the most effective low-resource strategy, yielding gains of over 13 F1 points at 1% of data by leveraging existing annotations for related fields. Synthetic pre-training provides smaller improvements, likely due to the domain gap between generated and real handwriting as well as mismatched field value distributions. We further find that smaller models like Donut are competitive or superior to PaliGemma once moderate training data is available, and that the choice between single-field and multi-field training depends on model capacity. To support this and future work, we publish CM/1v2, extending the CM/1 dataset with three new fields: Nationality, Place of Birth, and Religion.

References

1. Archibald, T., Martinez, T.: DELINE8K: A Synthetic Data Pipeline for the Semantic Segmentation of Historical Documents. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024, vol. 14806, pp. 289–304. Springer Nature Switzerland (2024)
2. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
3. Bondzio, S., Rass, C., Tames, I.: People on the move. Revisiting events and narratives of the European refugee crisis (1930s–1950s). In: Borggräfe, H. (ed.) Freilegungen. Wege, Orte Und Räume Der NS-Verfolgung, Jahrbuch Des International Tracing Service, vol. 5, pp. 36–55. Wallstein Verlag (2016)

4. Crosilla, G., Klic, L., Colavizza, G.: Benchmarking large language models for handwritten text recognition. Journal of Documentation **81**(7), 334–354 (2025)
5. Dahinden, J., Fischer, C., Menet, J.: Knowledge production, reflexivity, and the use of categories in migration studies: Tackling challenges in the field. Ethnic and Racial Studies **44**(4), 535–554 (2021)
6. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., Morariu, V.: End-to-end document recognition and understanding with Dessurt. In: Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, vol. 13804, pp. 280–296. Springer (2022)
7. Diaz, M., Mendoza-García, A., Ferrer, M.A., Sabourin, R.: A survey of handwriting synthesis from 2019 to 2024: A comprehensive review. Pattern Recognition **162**, 111357 (Jun 2025)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
9. Gan, J., Wang, W., Leng, J., Gao, X.: HiGAN+: Handwriting imitation GAN with disentangled representations. ACM Transactions on Graphics **42**(1), 11:1–11:17 (2022)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
11. Huhn, S.: Negotiating forced migration in the IRO's 'Care and Maintenance' (CM/1) files. One setting, three underlying aims, (at least) four actors, and multiple forms of human agency. IMIS Working Papers (12), 1–32 (2021)
12. Huhn, S., Rass, C.: Displaced person(s): The production of a powerful political category. Ethnic and Racial Studies **48**(4), 718–739 (2025)
13. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: OCR-free document understanding transformer. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol. 13688, pp. 498–517. Springer (2022)
14. Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 18893–18912. PMLR (2023)
15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics (2020)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
17. Marti, U.V., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002)
18. Mathew, M., Karatzas, D., Jawahar, C.: DocVQA: A dataset for VQA on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2200–2209 (2021)

19. Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P., Hussenot, L., et al.: Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
20. Rass, C., Tames, I.: Negotiating the aftermath of forced migration: A view from the intersection of war and migration studies in the digital age. Historical Social Research **45**(4), 7–44 (2020)
21. Sachdeva, B., Uppal, K., Java, A., Balasubramanian, V.N.: Understanding Task Transfer in Vision-Language Models. arXiv preprint arXiv:2511.18787 (Nov 2025)
22. Singh, K., Navaratnam, T., Holmer, J., Schaub-Meyer, S., Roth, S.: Is Synthetic Data all We Need? Benchmarking the Robustness of Models Trained with Synthetic Images. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2505–2515. IEEE (Jun 2024)
23. Vidal-Gorène, C., Camps, J.B., Clérice, T.: Synthetic Lines from Historical Manuscripts: An Experiment Using GAN and Style Transfer. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds.) Image Analysis and Processing - ICIAP 2023 Workshops, vol. 14366, pp. 477–488. Springer Nature Switzerland (2024)
24. Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., Ingold, R.: Generating Synthetic Handwritten Historical Documents With OCR Constrained GANs. In: International Conference on Document Analysis and Recognition. pp. 610–625. No. arXiv:2103.08236, Springer, arXiv (May 2021)
25. Wehner, J., Rass, C.: Disputed (non-)Belonging: Migrant agency in the European displacement crisis 1945–56. Journal of Contemporary History (2025)
26. Wolf, F., Tüselmann, O., Matei, A., Hennies, L., Rass, C., Fink, G.A.: CM1 – a dataset for evaluating few-shot information extraction with large vision language models. In: Document Analysis and Recognition – ICDAR 2025. Lecture Notes in Computer Science, Springer (2025)
27. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (Aug 2020)
28. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. arXiv preprint arXiv:2104.08836 (Sep 2021)
29. Yang, Y., Patel, A., Deitke, M., Gupta, T., Weihs, L., Head, A., Yatskar, M., Callison-Burch, C., Krishna, R., Kembhavi, A., Clark, C.: Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation (May 2025)
30. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (2023)
31. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE **109**(1), 43–76 (Jan 2021)