Complex Layout Classification in the Wild: A Low-Resource Approach with Layout-Preserving Augmentations

Sharva Gogawale, Iddo Hakim, Gal Grudka, Mohammad Suliman, Omer Ventura, Daria Vasyutinsky-Shapira, Berat Kurar-Barakat, and Nachum Dershowitz

School of Computer Science and AI, Tel Aviv University, Ramat Aviv, Israel
{sharvag,iddoh,galgrudka,suliman,omerventura}@mail.tau.ac.il
{dariashap,berat,nachumd}@tauex.tau.ac.il

Abstract. Many digitized corpora suffer from low resources because annotations may be scarce, page scans are noisy and of poor resolution, or layouts are structurally complex in ways that negatively affect the quality of automatic transcription. Developing robust classification models for low-resource languages is inhibited by the lack of large-scale annotated data and by the frequent semantic complexity of page layouts. To this end, we have curated a complex-layout dataset, manually classified into eight distinct layout types based on their separator regions. To overcome data scarcity, we propose a novel training strategy in the form of a CNN-based classifier that employs strong, domain-aware augmentations to improve generalization. We utilize narrow anisotropic Gaussian masking to suppress incidental textual details while preserving essential separations, compelling the model to learn global geometric arrangements. Additionally, we implement reflection-induced label transformations to enrich the training distribution while maintaining label consistency across asymmetric categories. The results demonstrate that layout-specific augmentations can substantially improve page-level layout classification under severe annotation scarcity.

Keywords: Document layout analysis · complex layout classification · historical documents · data augmentation · low-resource learning

1 Introduction

Document layout analysis is a fundamental task in document understanding, which serves as a critical prerequisite for downstream pipelines, such as optical character recognition (OCR), automated transcription [16], entity extraction, and digital reconstruction of reading orders. Within layout analysis, layout classification refers to the task of assigning the layout of a page to a specific category. This is especially useful for historical documents, as it allows each page to be directed to the best processing pipeline. Then, later steps (e.g. segmentation, OCR, and reading-order reconstruction) can be adapted to the specific layout.

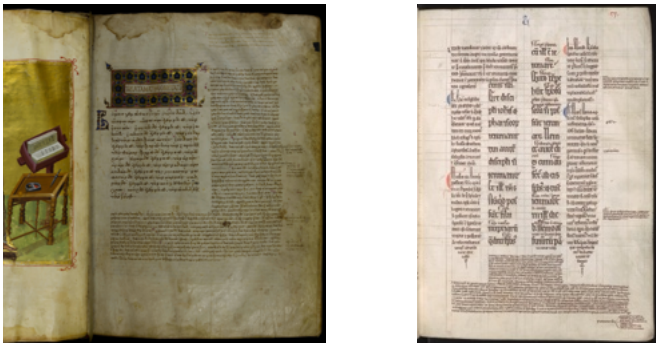

Fig. 1: Sample parchment pages with complicated layouts. Left: Greek Gospels with catena (10th c. Byzantium), National Library of Greece. Right: Gospels in Latin with glosses (England, 1300–1350), Cambridge, Queens' College, MS 26.

Modern documents such as digital PDFs, invoices, or scientific papers usually follow clean and well-organized grid structures, often called "Manhattan" layouts. In contrast, historical documents are much more irregular and complex, and often contain a high level of visual noise. This makes layout classification not merely a visual segmentation task but an important preliminary step in separating the logical structures from physical attributes. Correspondingly, layout analysis can be broadly categorized into visual-based physical layout analysis and semantics-based logical analysis [7]. Visual-based analysis relies on appearance and geometric cues to differentiate regions with different visual attributes, whereas logical analysis builds upon this segmentation by organizing regions according to the semantic role and relationships expressed in the text [15].

Complex layouts (Fig. 1) are often found in medieval manuscripts before the era of print (until the mid-16th century), and are present in recently printed books, too. Developing robust models for classifying such layouts can support layout-aware segmentation and OCR pipelines for both manuscripts and printed books. Because of the uniformity of printed scripts, it is easier to first solve this problem for printed books, and then proceed to manuscripts.

Two natural layouts for a page containing a main text along with a translation or commentary are two columns or top and bottom. We refer to these layouts, as well as one region per page, as "simple." But more complex options suggest themselves, especially when the main part is considerably shorter than the supplementary parts. Rather than two vertical columns, the supplement may be wrapped around the main text in several ways: L-shaped and its mirror image (frequently on facing pages); C-shaped and its mirror image (already attested in the 8th century Codex Zacynthius, containing the Gospel of Luke with a running catena commentary); O-shaped with the main text in the center. These arrangements culminate in the Glossa Ordinaria or "Talmudic" layout (included in our catch-all Y layout), with three (or more) main components: a central text, with one commentary on the outside (L or C) and another on the inside (L or C), with

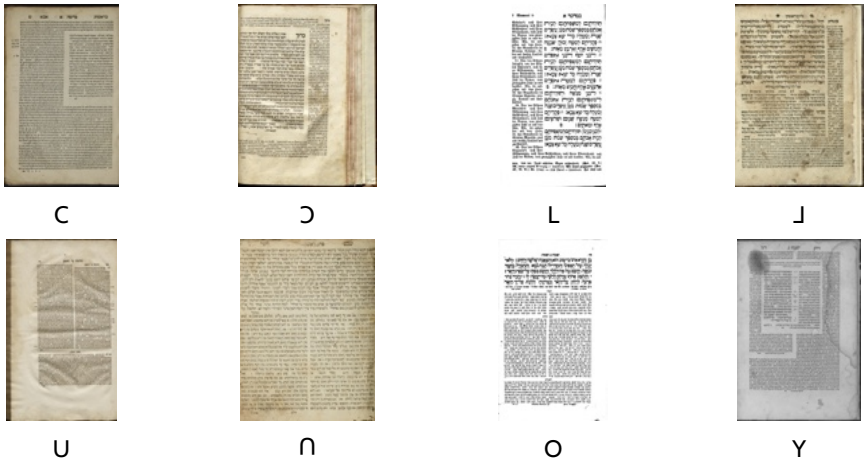

Fig. 2: Samples of complex layout classes.

facing pages usually having a mirrored layout. Also, instead of a full-width main text on the top of the page with upper and lower texts, the upper alone may be split into two columns in what we call the U format, or likewise, the lower text may be split in an inverted U format. When there are three components, top, bottom, and middle, one or more may be split into two columns. This variety of layouts continued to be used when books with such texts were printed. Fig. 2 presents illustrative examples of (Hebrew) books for each class.

Of special interest for manuscripts is the layout of the colophon page. These are unique for every manuscript, but their classification, as well as that of the many other decorative formats (e.g. *tashqil* in Samaritan manuscripts or micrography in Hebrew ones), are beyond the scope of this paper.

Despite strong progress in document layout analysis, most benchmark datasets focus on layout analysis for English-language articles and similarly standardized document types. The recent advancements in deep learning-based layout analysis have been driven by large-scale annotated benchmarks like PubLayNet [17] and DocLayNet [9], which predominantly focus on English-language scientific articles or modern document types. For instance, in [11], document page classification is addressed using deep neural networks that leverage visual layout cues to distinguish broad structural categories. However, these improvements do not easily apply to low-resource historical collections with complex layouts. Fully supervised methods need large amounts of annotated data and often fail to handle unseen layout structures. Digitizing such material faces many challenges, and supervised methods trained on available datasets struggle with the unseen, non-rectangular layouts of medieval manuscripts.

It has been shown that pixel-based segmentation approaches often perform poorly when dealing with complex document layouts, especially when simple structural assumptions fail to capture the diverse organization of the page [3].

Fig. 3: Complex layout samples with faulty segmentation.

Consequently, performance degradation is commonly observed in datasets containing intricate layout configurations. Even popular open-source text recognition environments like kraken and eScriptorium [14], which are designed for historical and non-Latin scripts, rely mostly on pixel-level segmentation. This makes it hard for them to reliably separate meaningful regions in documents with complex layouts. As illustrated in Fig. 3, these systems often fail to distinguish between semantically distinct text blocks in complex C- or L-shaped configurations. Thus, there is a critical need for robust classification models capable of learning from minimal samples without requiring massive-scale annotation efforts. These limitations motivate methods that can learn from tiny annotated datasets while focusing on stable geometric cues such as separator structure rather than overfitting to incidental textual content or style.

To address these challenges, we study complex-layout classification in a low-resource historical setting by curating a compact dataset and proposing a training strategy that emphasizes layout-defining separator geometry. The main contributions are as follows: (1) We curate the CLC dataset, a specialized corpus of 155 high-resolution Hebrew pages manually annotated with eight distinct, semantically complex layout classes. (2) We propose a *Narrow Anisotropic Gaussian Augmentation* strategy that suppresses incidental textual details while preserving essential separator structures, compelling the model to learn global geometric arrangements rather than overfitting to font or content. (3) We provide a systematic evaluation under severe data scarcity, including backbone comparison, augmentation ablation, external blind-corpus evaluation and a preliminary label-efficiency analysis.

2 Related Work

Layout analysis benchmark datasets. Layout analysis has benefited substantially from the release of large-scale annotated benchmarks. PubLayNet [17] is a widely used public dataset built primarily from modern scientific articles, enabling strong supervised training for document segmentation and structural understanding. DocLayNet [9] expands this direction by covering more diverse

and visually complex layouts drawn from a variety of public document sources. The PRImA [1] layout analysis dataset is comprised of heterogeneous document types (e.g. magazines and technical articles), comprising 305 pages grouped into six categories, namely, Financial Reports, Manuals, Scientific Articles, Laws & Regulations, Patents, and Government Tenders; the majority of documents are in Latin scripts. More recently, M^6Doc [3] explicitly targets diversity in both layout and language, including multi-layout settings (rectangular, Manhattan, non-Manhattan, and multi-column Manhattan) and multilingual content (Chinese and English). Collectively, these datasets have shaped the dominant evaluation protocols and model designs for layout analysis under modern, well-resourced conditions.

Historical layout datasets. Several datasets have been compiled to address historical documents with unique challenges due to noise, degradation, and layout complexity. NewsEye [5] comprises Austrian newspaper pages from the 19th and early 20th century with carefully corrected text, including 148 training pages and 13 validation pages, with ground truth created using Transkribus by the Austrian National Library. DIVA-HisDB [13] includes 150 pages from three medieval Latin manuscripts from the Carolingian period. A Brazilian historical newspaper dataset contains 140 page images sourced from the Brazilian National Digital Library (BND) [11]. For non-Latin scripts, SCUT-CAB [2] addresses the complex layouts of ancient Chinese books with 4000 manually annotated images, while the HJDataset [12] provides a large-scale resource for historical Japanese documents, including hierarchical structures and reading-order information. These datasets highlight the growing recognition that historical layout analysis requires specialized resources; they also underscore how fragmented the landscape remains across scripts and traditions. U-DIADS-Bib [19] introduced a few-shot benchmark for ancient manuscript layout segmentation. The ICDAR 2025 FEST competition [18] targeted few-shot text-line segmentation of handwritten documents. Our focus is on page-level separator-driven layout classification as an upstream routing step before segmentation and OCR.

Page-level classification. Approaches to page-level layout classification have evolved from feature engineering to deep learning. Early work often relied on manually engineered features. A study on page layout classification for legal documents categorized pages into structural types such as single column, marginal single column, and two column formats. The results showed that feature-engineered methods (e.g. SVM and random forest) leveraging geometric layout cues achieved accuracies around 94%, emphasizing the value of explicit structural features [6]. With the advent of deep learning, convolutional neural networks (CNNs) became the dominant strategy. A prior study demonstrated that CNNs could effectively discriminate between diverse categories such as letters, memos, invoices, etc., by analyzing whole-page visual cues [1]. Additionally, frameworks for graphical object detection [10] have leveraged transfer learning and domain adaptation to identify graphical elements like tables and figures, mitigating the scarcity of annotated training data.

Hebrew layout resources. In the specific context of Hebrew document layout analysis, resources are notably scarce. The primary existing benchmark is NetLay [4], which contains approximately 1,300 pages from printed Hebrew (or Hebrew-alphabet) books. NetLay utilizes a multi-label encoding approach with VGG16 to classify pages into four categories, namely no-text, single-column, double-column, and complex. However, the label space remains relatively coarse with respect to separator-defined, non-rectangular historical layouts (e.g. C- or L-shaped configurations), which motivates finer-grained datasets and modeling assumptions that explicitly target separator geometry. Motivated by these gaps, our work focuses on separator-defined complex layout classification for low-resource historical Hebrew pages, complementing existing Hebrew resources (e.g. NetLay) with a finer-grained label space tailored to non-rectangular and multi-region structures. Our goal is to provide a targeted, publicly usable dataset and baseline that isolates the layout-classification problem under severe annotation constraints and emphasizes global geometric structure.

3 Complex Layout Classification (CLC) Dataset

We curate a compact, task-specific corpus to address complex layout classification with the high-level goal of improving downstream text transcription systems by routing pages to layout-aware pipelines. The dataset comprises 155 high-resolution document images sourced from the National Library of Israel's digital collections and printed books from various periods. To evaluate the efficacy of our model in a low-resource setting, we work with a small but representative dataset, a common constraint in computational humanities where large-scale annotated data are scarce.

The images are manually categorized into our eight classes. The complexity of these layouts is not merely structural but also semantic; pages often contain separate but related works by different authors, thereby necessitating models that can capture complex, multi-block hierarchical relationships. They use two main script types: square (block letters) and "Rashi" (curvilinear). To support robust learning from small data, all images are retained at their native scan resolution. Table 1 summarizes the distribution across the eight layout classes, the script types present in each, and the annotation source. Fragments in non-Hebrew characters are a common phenomenon in complex layouts, as such texts typically include commentaries or translations into local vernaculars or appear in scholarly editions.

Ground truth annotations were established through a rigorous protocol. To maximize structural diversity, candidate pages were restricted to a single representative page per printed book. Initial categorizations were performed by trained annotators, and all labels were subsequently reviewed and verified by an expert paleographer, resolving the disagreements to produce a single ground-truth label per image. The complete dataset, including the images, book identifiers, and the ground truth, is available at https://github.com/TAU-CH/midrash_clc.

Table 1: CLC dataset statistics. Sq = square (block letters); Ra = Rashi (curvilinear). Pages sourced from the National Library of Israel digital collections.

Class	Count	% Total	Script Types
C	21	13.5	Sq, Ra
Ↄ	19	12.3	Sq, Ra
L	28	18.1	Sq, Ra
⅃	22	14.2	Sq, Ra
O	22	14.2	Sq, Ra
U	16	10.3	Sq, Ra
∩	13	8.4	Sq, Ra
Y	14	9.0	Sq, Ra
Total	**155**	**100.0**	—

4 Method

We address the task of complex-layout classification, where document images are to be assigned to one of eight predefined layout types. Let $\mathcal{C}$ denote the set of layout classes, $\mathcal{C} = \{\mathsf{C}, \mathsf{Ↄ}, \mathsf{L}, \mathsf{⅃}, \mathsf{U}, \mathsf{∩}, \mathsf{O}, \mathsf{Y}\}$. These classes correspond to page layouts containing text regions shaped as C, horizontally reflected C, L, horizontally reflected L, U, vertically reflected U, or O, with irregular region forms grouped under the Y class.

The primary challenge in classifying document images into these eight classes is the severe scarcity of annotated data, with only a handful of samples available per class. To address this limitation, we propose a CNN-based framework that employs strong augmentation strategies to improve generalization.

In layout classification, the defining feature of each class lies in its separator regions, which partition the page into characteristic geometric layouts defined in $\mathcal{C}$. In contrast, the textual content, including font type, size, or semantic information does not determine the layout category. Therefore, any augmentation procedure that preserves separator regions encourages the model to focus on the global arrangement of separators rather than incidental textual details.

In particular, we adopt two complementary families of augmentations: (1) Narrow anisotropic Gaussian based masking, which applies multi-oriented anisotropic Gaussian filters with a narrow standard deviation to suppress text regions while preserving separator regions; and (2) horizontal and vertical reflections, which enrich the diversity of training samples while ensuring label consistency in asymmetric categories. Together, these augmentations significantly expand the effective training distribution and enable robust CNN learning despite the limited availability of annotated data.

Narrow anisotropic Gaussian based masking. We augment the document images using a multi-oriented anisotropic Gaussian filter that preserves separator

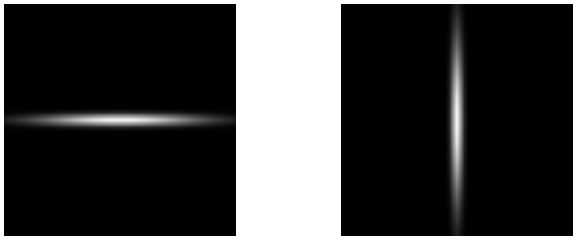

Fig. 4: Narrow anisotropic Gaussians used for masking ($\sigma_x = 1$; for visualization, $\sigma_y = 15$ is shown). Left: horizontal ($\theta = 0$). Right: vertical ($\theta = \pi/2$).

regions. For orientation θ and standard deviations σ_x, σ_y, the kernel is

$$G_{\sigma_x,\sigma_y,\theta}(x,y) \;=\; \frac{1}{2\pi\sigma_x\sigma_y}\exp\!\left(-\frac{(x\cos\theta + y\sin\theta)^2}{2\sigma_x^2} - \frac{(-x\sin\theta + y\cos\theta)^2}{2\sigma_y^2}\right)$$

Given a document image $I(x, y)$, where (x, y) are pixel coordinates, containing low-contrast text on a high-contrast background, we fix $\sigma_x = 1$ and filter $I(x, y)$ with $G_{1,\sigma_y,\theta}$ at orientations $\theta \in \{0, \pi/2\}$ (Fig. 4). For each pixel, the maximum response across the two orientations is computed to form the response map $R_{\sigma_y}(x, y)$. The directional response and the subsequent binary mask are defined as: $R_{\sigma_y}(x, y) = \max_{\theta\in\{0,\pi/2\}}(I * G_{1,\sigma_y,\theta})(x, y)$; $M_{\sigma_y}(x, y) = \mathbf{1}[\tilde{R}_{\sigma_y}(x, y) \geq t]$, where $\tilde{R}_{\sigma_y}$ denotes the min-max normalized response map scaled to $[0, 1]$, and $\mathbf{1}[\cdot]$ is the indicator function. We then form the masked augmentation by preserving pixels in separator regions and suppressing the remaining content: $I'(x, y) = 255 \cdot M_{\sigma_y}(x, y) + \big(1 - M_{\sigma_y}(x, y)\big)\,\big(255 - I(x, y)\big)$. Pixels outside the separator mask are inverted to suppress fine-grained text texture while retaining global page structure and high-contrast separator geometry.

The use of $\sigma_x = 1$ together with multi-orientation filtering ensures that separator region pixels consistently yield high responses. In our case, we set $\theta \in \{0, \pi/2\}$, since separator regions in the dataset are oriented only vertically or horizontally. At vertical separator regions, the maximum response is produced by the vertically oriented narrow Gaussian, whereas at horizontal separator regions it is produced by the horizontally oriented narrow Gaussian. Because the filters are narrow, their placement at separator regions near text region boundaries still yields high responses without blurring, despite the presence of neighboring low-contrast pixels within the text regions (Fig. 5).

Horizontal and vertical reflections. To further enrich the training distribution, we apply horizontal and vertical reflections, which preserve separator regions while modifying the overall page structure. In contrast, rotations and shearings are deliberately excluded, as they alter separator orientations and thereby disrupt the structural integrity of the layout. Translations are avoided, since they may shift crucial separator regions outside the field of view, invalidating the class-defining structure. And scalings are omitted because all input images are rescaled to a fixed size prior to training, rendering additional scaling redundant.

Fig. 5: Illustration of separator region preservation in a document with an L-shaped layout. On vertical separator regions, no matter how narrow they are, a narrow vertical Gaussian (red) will lead to a high response. Similarly, on horizontal separator regions, no matter how narrow they are, a narrow horizontal Gaussian (green) will lead to a high response.

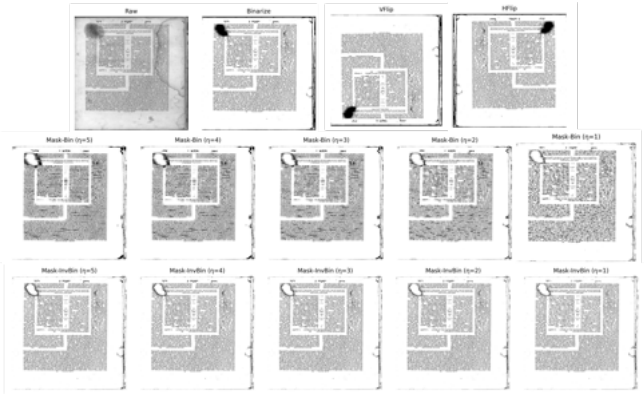

Fig. 6: Comprehensive view of the augmentation suite. A single training image is expanded into eleven distinct intensity representations (binary, masked, and inverse-masked), alongside their valid spatial reflections.

Reflections do require relabeling; however, the label set is closed under reflection, ensuring that the new labels remain within the $\mathcal{C}$. Horizontal reflection exchanges C $\leftrightarrow$ Ↄ and L $\leftrightarrow$ ⅃. Vertical reflection exchanges U $\leftrightarrow$ ∩ and turns L and ⅃ into Y. Two classes, O and Y, are (usually) retained under reflection.

Augmentations. To overcome severe data scarcity, we implement a combinatorial but task-constrained augmentation suite that preserves class-defining separator topology, in contrast to generic appearance perturbations which perturb arbitrary image regions without structural awareness. For each training image, we first binarize, while also extracting an inverse-binarized variant to minimize polarity sensitivity. We then generate masking-based variants by applying our narrow anisotropic Gaussian operator across five strength settings, $\eta \in \{1, 2, 3, 4, 5\}$. This gives both binarized-masked and inverse-binarized-masked modalities, re-

sulting in a transformed pool of $|\mathcal{T}| = 11$ distinct intensity representations per image (1 plain binary, 5 masked binary, and 5 inverse-masked binary) as illustrated in Fig. 6. Subsequently, each of the 11 modalities is subjected to a spatial reflection protocol using deterministic label mappings. Horizontal reflection is universally applied, exchanging asymmetric pairs, while preserving symmetric classes. Vertical reflection is applied exclusively to the subset of classes where structural semantics remain well-defined (U, ∩, O, C, ↄ, Y). By enumerating all valid spatial and intensity combinations, this pipeline effectively acts as a dataset multiplier, yielding 22 to 33 training instances per original image depending on class validity under vertical reflection. Finally, to mitigate the residual class imbalance within this expanded set, we employ weighted random sampling during optimization.

ConvNeXt-based classifier. ConvNeXt [8] is a modernized Convolutional Network which retains inductive biases that matter for layout: locality, translation equivariance, and hierarchical composition—properties well aligned with separator-driven global geometry under scarce labels. A ConvNeXt block replaces classic ResNet bottlenecks with: (i) large-kernel depthwise convolution for long-range spatial aggregation, and (ii) pointwise "MLP-style" mixing for channel interactions, while using LayerNorm/GELU for stable optimization.

Let $x \in \mathbb{R}^{H\times W\times C}$ be an input feature map. A depthwise convolution with kernel K operates per-channel:

$$\mathrm{DWConv}(x)_{i,j,c} = \sum\nolimits_{u,v\in K} w_{u,v,c}\, x_{i+u,\, j+v,\, c}.$$

Channel mixing is then performed by pointwise projections (equivalent to 1×1 convolutions): For a channel vector $\mathbf{z}_{i,j} \in \mathbb{R}^C$ at spatial location (i, j), this operates as a linear transformation $\mathrm{PW}(\mathbf{z})_{i,j} = \mathbf{W}\mathbf{z}_{i,j}$. A simplified ConvNeXt block can be written as:

$$y = x + \mathcal{F}(x), \qquad \mathcal{F}(x) = \mathrm{PW}_2(\sigma(\mathrm{LN}(\mathrm{PW}_1(\mathrm{DWConv}(x))))\,,$$

where LN is LayerNorm, σ is the GELU activation, PW_1 is an expansion layer that increases the channel dimension by a factor of 4, and PW_2 is a projection layer that reduces it back to the original dimension.

For classification, the network outputs logits $g(x) \in \mathbb{R}^8$, and is trained with standard cross-entropy loss.

We fine-tune an ImageNet-pretrained ConvNeXt-Tiny, adapting the stem to 1-channel inputs by averaging RGB pretrained weights, and replacing the final classifier for 8 classes. We use *staged fine-tuning*: we train the classifier head first and then unfreeze the backbone with a smaller learning rate, which reduces overfitting early and improves stability under scarce supervision. We first establish a strong *supervised baseline* by fine-tuning standard image backbones for 8-way page-level layout classification (C, ↄ, L, ⅃, U, ∩, O, Y).

Input images were resized to 224×224 pixels and binarized using Otsu's thresholding prior to ingestion. We utilized a stratified data split, allocating

60% to training, 20% to validation, and 20% to testing. To prevent destabilizing the pretrained weights, we employed a two-stage fine-tuning strategy for a maximum of 100 epochs using a batch size of 16. For the first 10 epochs, we froze all non-classifier parameters and trained the head using the AdamW optimizer (initial learning rate 10^{-3}, weight decay 10^{-4}) with a cosine annealing schedule ($T_{\max} = 70$, $\mathrm{lr}_{\min} = 10^{-5}$). Subsequently, the entire network was unfrozen, and AdamW was reinitialized with a lower learning rate of 10^{-4} (weight decay 10^{-4}) and a new cosine schedule ($T_{\max} = 90$, $\mathrm{lr}_{\min} = 10^{-6}$). Training was regulated by early stopping with a patience of 10 epochs based on validation loss.

Active learning protocol. We also study an active learning protocol to evaluate label efficiency in our low-resource setting, which reflects historical document analysis, where structurally complex layouts often require domain-expert annotation and labels are expensive to obtain. Therefore, even modest reductions in labeling effort are practically useful.

Let $\mathcal{D}_{\text{train}}$ be the training split. At round t, we maintain a labeled set $\mathcal{L}_t \subseteq \mathcal{D}_{\text{train}}$ and an unlabeled pool $\mathcal{U}_t = \mathcal{D}_{\text{train}} \setminus \mathcal{L}_t$. We initialize with $\rho_0 \approx 0.2$ and add $\Delta\rho \approx 0.2$ each round until the pool is exhausted. At each round, we train ConvNeXt-Tiny with our **Full** pipeline (Binary + Reflections + Anisotropic Masking) on the labeled set, then select a batch $\mathcal{S}_t \subset \mathcal{U}_t$ of size B according to an acquisition function $a_t(x)$: $\mathcal{S}_t = \operatorname{argmax}_{S \subseteq \mathcal{U}_t, |S|=B} \sum_{x \in S} a_t(x)$, followed by $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \mathcal{S}_t$ and $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_t \setminus \mathcal{S}_t$. We compare three simple acquisition rules: (i) **Random** ($a_t(x)$ uniform); (ii) **Entropy**, using predictive entropy $a_t(x) = -\sum_{c=1}^{8} p(c \mid x) \log\big(p(c \mid x) + \epsilon\big)$; and (iii) **Margin**, using the negative gap between the two highest class probabilities, $a_t(x) = -\big(p_{(1)}(x) - p_{(2)}(x)\big)$, where $p_{(1)}(x)$ and $p_{(2)}(x)$ denote the largest and second-largest entries of $p(\cdot \mid x)$. For the uncertainty rules we select the B samples with the highest $a_t(x)$.

Learning curve summary. Across labeled budgets $\rho_t = |\mathcal{L}_t|/|\mathcal{D}_{\text{train}}|$, we evaluate accuracy and macro-F_1 on a fixed held-out split and summarize label-efficiency using the area under the learning curve (AULC).

5 Results

Backbone architecture comparison. To determine the optimal architecture for complex layout classification under severe data constraints, we evaluated a diverse set of deep learning models: classical CNNs (VGG16, ResNet-18), a highly optimized mobile architecture (EfficientNet-B0), a pure vision transformer (ViT-B/16), and a modernized CNN (ConvNeXt-Tiny). All architectures were trained using our proposed Full layout-preserving pipeline (Binary+Flips+Masking) to ensure a fair comparison. Because the test set is small (31 pages), we report macro-averaged precision, recall, and F_1-score alongside accuracy to account for class imbalance.

Compared to vision transformers, which typically require extensive supervision to learn strong spatial priors, ConvNeXt relies less on learning global

Table 2: Left: Backbone comparison under our full layout-preserving pipeline on the CLC dataset. Right: NetLay-adapted multi-label baseline models *without* layout-preserving augmentations. Test set: 31 pages. Best in bold.

Model	**Params**	**Acc.**	**P**	**R**	F_1	**NetLay**	**Acc.**	**P**	**R**	F_1
ConvNeXt-Tiny	28.6M	**0.90**	**0.93**	**0.88**	**0.88**	EfficientNet-				
EfficientNet-B0	5.3M	0.45	0.39	0.44	0.41	V2-S	**0.52**	**0.60**	**0.48**	**0.49**
VGG16	138M	0.45	0.39	0.44	0.41	VGG16	0.45	0.42	0.42	0.40
ResNet-18	11.7M	0.23	0.16	0.20	0.17	ViT-B/16	0.23	0.30	0.19	0.17
ViT-B/16	86.6M	0.23	0.19	0.20	0.15					

Table 3: External blind evaluation on the test corpus (20 samples per class).

Class	Wrong	Right	Accuracy (%)	Class	Wrong	Right	Accuracy (%)
⊂	1	19	95	○	13	7	35
⊃	8	12	60	∪	0	20	100
⌊	0	20	100	∩	4	16	80
⌋	0	20	100	Y	0	20	100
				Overall	26	134	83.75

structure from scratch. The large receptive field from depthwise large-kernel convolutions allows the network to capture the global geometric arrangement of separator regions (e.g. distinguishing a ∪ layout from an ⌊ layout), and parameter-efficient depthwise operations reduce overfitting risk and improve efficiency. This dynamic is visually confirmed in Fig. 7, where the network's attention is predominantly localized along separator regions, block boundaries, and inter-column gaps rather than dense textual content. Such qualitative evidence demonstrates that our proposed preprocessing and augmentations successfully compel the classifier to anchor its predictions on global separator geometry.

Table 2 summarizes the performance of the evaluated architectures. We additionally adapt the NetLay multi-label encoding [4] to our 8-class task (Table 2, right); the best result reaches only 52% accuracy and 0.49 macro-F_1 without layout-preserving augmentations, confirming that the separator-aware training strategy is the decisive contribution. ConvNeXt-Tiny performs substantially better than the others, achieving an accuracy of 0.90 and a macro-F_1 of 0.88, substantially outperforming EfficientNet-B0 and VGG16 (acc. 0.45, macro-F_1 0.41), as well as ResNet-18 and ViT-B/16 (macro-$F_1 \leq 0.17$). On a per-class basis, ConvNeXt achieves perfect or near-perfect performance on most layouts (e.g. ⊃/⌋/∪), with remaining errors concentrated in the rarest and most ambiguous classes (Y and ∩), suggesting confusion among separator patterns when examples are extremely limited. In low-resource regimes, ViTs often underperform unless heavily regularized or trained on large datasets. ResNet-18 suffers a similar fate, likely due to its smaller receptive field and aggressive spatial downsampling, which can weaken the thin, delicate separator lines crucial for identifying medieval layout classes. The strong performance of ConvNeXt-Tiny

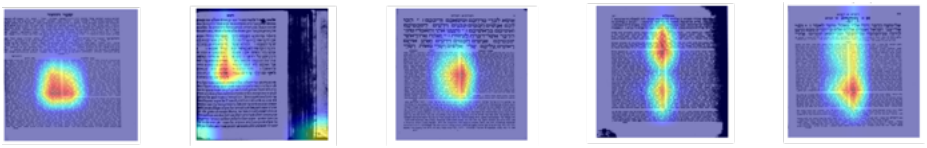

Fig. 7: GradCAM visualizations of the ConvNeXt-Tiny classifier. Warmer colors indicate regions with higher contribution to the predicted layout class.

supports our data-scarcity hypothesis: with limited samples, the model cannot rely on memorizing textures. Instead, CNN inductive biases encourage learning contiguous edges, while the ViT-inspired design helps integrate these cues into coherent global layout structures. The only notable struggle for ConvNeXt-Tiny is the Y class (recall of 0.33, $F_1 = 0.50$), which represents the most intricate layout with central text flanked by multiple distinct elements. This slight degradation highlights that, while ConvNeXt effectively models global structure from minimal data, highly fragmented layouts with inner and outer separators remain challenging for zero-to-few-shot historical document analysis.

Evaluation on a blind corpus. To examine how the classifier behaves outside our CLC dataset, we conducted a blind-corpus sanity check on a large corpus drawn from the book collections of the National Library of Israel (NLI). These were not used for training, validation, or hyperparameter tuning. We emphasize that this is not a fully labeled out-of-domain benchmark; rather, it is a precision-oriented stress test on high-confidence predictions in a large unlabeled archive. Inference was performed over a directory containing 36,215 book folders. For each folder, up to 10 page images were randomly sampled (or fewer when unavailable), yielding a maximum of 362,150 candidate pages. The trained classifier model (Full configuration: Binary+Flips+Masking) was applied without any fine-tuning. To focus on highly confident predictions, we retained only outputs whose maximum softmax probability exceeded 0.99. This procedure identified: 3,932 book folders containing at least one page classified as complex; and 9,339 page images satisfying the confidence threshold.

Because the corpus lacks ground truth layout annotations, we conducted a manual visual evaluation on a stratified sample of the high-confidence predictions. For each of the eight classes, 20 images were randomly selected. Each image was visually inspected and assigned its true layout class, which was then compared to the model prediction. See Table 3. Out of 160 inspected samples, 134 were correct, giving an overall accuracy of 83.75% on this verified subset. Performance was lower for Ↄ and especially for O layouts.

Ablation. To systematically quantify the contributions of our proposed layout-preserving augmentations, we conducted an extensive ablation study using the best-performing ConvNeXt-Tiny architecture. All experiments are evaluated on the same 8-way test set (31 pages).

Table 4: Ablation results on CLC using ConvNeXt-Tiny (31 pages), showing **Acc**uracy and macro-averaged **P**recision, **R**ecall, and F_1. Best in **bold**.

Config	**Acc.**	**P**	**R**	F_1
Binary only	0.32	0.34	0.31	0.29
Binary+Flips	0.45	0.44	0.47	0.42
Binary+Masking	0.68	0.69	0.61	0.59
Binary+Std. Aug.	0.68	0.76	0.62	0.63
Full (Masking+Flips)	**0.90**	**0.93**	**0.88**	**0.88**

η	**Acc.**	**P**	**R**	F_1
η=1	0.55	0.56	0.53	0.52
η=3	0.71	0.82	0.70	0.69
η=5	0.81	0.85	0.79	0.79

Ablation Settings. We consider (i) **Binary only**, which uses only binarized inputs without geometric augmentations; (ii) **Binary+Flips**, which additionally applies horizontal/vertical reflections with label-consistent remapping for asymmetric layouts; (iii) **Binary+Masking**, which uses only the proposed narrow anisotropic Gaussian masking to suppress incidental text while preserving separator structure; (iv) **Full**, which combines Binary+Flips+Masking; and (v) **Mask strength** sweeps, where we vary a masking hyperparameter $\eta \in \{1, 3, 5\}$ controlling the strength of anisotropic suppression.

Quantitative results. Table 4 shows that Binary preprocessing alone performs poorly (acc. 0.32, macro-F_1 0.29), indicating that simple foreground/background normalization is insufficient to learn separator-defined global geometry from tiny supervision. Notably, classes with nested semantic regions (⅃ and ⅄) completely fail to converge (F_1: 0.00) in this low-resource regime. Reflection-based augmentation in the **Binary+Flips** strategy, together with a fixed label-transformation mapping, improves accuracy to 0.45. While this successfully expands the dataset and enforces orientation invariance (boosting Ⅽ's F_1 from 0.44 to 0.62), the model remains heavily distracted by local textual textures, resulting in marginal gains for complex asymmetric layouts. We also evaluate a **Binary+Standard Augmentations** baseline using common transforms (random rotation, affine, perspective, and Gaussian blur), which achieves 0.68 accuracy and 0.63 macro-F_1. While this exceeds Binary+Flips, the full pipeline with separator-preserving masking and label-consistent reflections yields substantially stronger performance, confirming that task-aware augmentation design matters beyond generic regularization.

In contrast, **Binary+Masking** produces a substantial jump (acc. 0.68, macro-F_1 0.59), supporting the hypothesis that the task is fundamentally *separator-driven*. By stochastically suppressing high-frequency textual components while preserving the contiguous geometric paths of separators, the masking forces the CNN's inductive bias to focus strictly on layout-defining topological boundaries. Hence, the model is compelled to base its predictions on global geometric structures, such as long separator strokes, block boundaries, and their spatial arrangement, rather than local content- or style-specific patterns.

Finally, combining all components (**Full**) achieves the best overall performance (acc. 0.90, macro-F_1 0.88), indicating that flips and masking are comple-

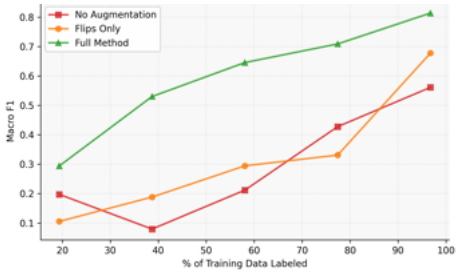

Fig. 8: Active learning under label scarcity. Macro-F_1 learning curves are shown for three augmentation settings, demonstrating that the full layout-preserving pipeline improves performance throughout the low-label regime.

mentary: flips increase effective sample diversity and promote invariance, while masking acts as a structure-preserving regularizer that improves generalization.

Effect of masking strength. Varying the masking hyperparameter highlights a clear sensitivity: weaker masking (η=1) underperforms (acc. 0.55, macro-F_1 0.52), while moderate-to-strong masking improves robustness (η=3: acc. 0.71, macro-F_1 0.69; η=5: acc. 0.81, macro-F_1 0.79). This trend suggests that sufficiently strong suppression is required to consistently remove text cues; otherwise, the model partially relies on residual texture signals that do not transfer across pages. This directional filtering acts as a line-tracking mechanism: it aggressively blurs orthogonal high-frequency noise (text) while constructively reinforcing continuous low-frequency paths (separators). Nevertheless, the best performance is obtained by the **Full** configuration, implying that masking alone is not a substitute for geometric augmentation and label-consistent symmetry handling.

Minimizing cost via active learning. Annotation remains a bottleneck in historical document analysis, since complex layouts often require domain-aware inspection. We therefore evaluate whether active learning can improve label efficiency when expanding CLC.

Figure 8 shows that the dominant factor in label-efficient learning is the augmentation strategy. Under entropy acquisition, the full layout-preserving pipeline achieves a macro-F_1 AULC of 0.61, compared with 0.30 for Binary+Flips and 0.28 for Binary only. This large shift indicates that separator-preserving masking and label-consistent reflections improve learning consistently across all labeling budgets, rather than only at the final training phase.

Evaluating the acquisition rules reveals a secondary effect relative to the augmentation strategy. On accuracy, the AULC is identical for Random, Entropy, and Margin acquisition (0.65), with all three converging to the same final accuracy of 0.87. However, macro-F_1 is more sensitive to the acquisition choice: Margin yields the highest AULC (0.62), compared with 0.61 for both Entropy and Random. Its advantage appears mainly in the mid-budget regime: at 60%

labeled data, Margin reaches a macro-F_1 of 0.69 (compared to 0.59 for Random), and at 80% it reaches 0.76 (compared to 0.73 for Random). At the earliest post-initial budget, however, Margin slightly trails Random, suggesting that acquisition-rule differences are less reliable when the labeled set is extremely small. Overall, these results suggest that the primary label-efficiency gain comes from the separator-aware augmentation pipeline, while margin-based acquisition provides a lightweight mechanism for prioritizing structurally informative pages during dataset expansion.

6 Conclusion

We addressed the challenging task of complex-layout classification in low-resource historical documents, motivated by the practical need to route heterogeneous documents to layout-aware transcription and segmentation pipelines. To support controlled evaluation under annotation scarcity, we curated the CLC dataset, consisting of 155 high-resolution pages spanning eight separator-defined layout categories. We introduced layout-preserving augmentations that suppress incidental textual detail while retaining separator structure, and combined them with label-consistent reflection transformations to expand the effective training distribution without introducing label noise. Masking, combined with label-consistent spatial reflections, achieved 90% accuracy on our dataset. Furthermore, an external blind evaluation on the massive NLI corpus demonstrated strong generalization behavior yielding an 83.75% accuracy on a manually verified high-confidence subset, suggesting that the classifier can generalize beyond the curated dataset. Finally, our preliminary active-learning study shows that the augmentation pipeline is the primary driver of label efficiency, while margin-based acquisition provides a lightweight strategy for prioritizing structurally informative pages during dataset expansion. Future work will extend the approach to handwritten manuscript pages and integrate script-type recognition for layout-aware OCR routing.

Acknowledgments. This research was funded in part by the European Union as part of "MiDRASH" [https://www.midrash.eu] (ERC project no. 101071829, with principal investigators: Nachum Dershowitz, Tel Aviv University; Judith Olszowy-Schlanger, EPHE-PSL; Avi Shmidman, Bar-Ilan University; and Daniel Stökl Ben Ezra, EPHE-PSL). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: ICDAR. pp. 296–300 (2009)

2. Cheng, H., Jian, C., Wu, S., Jin, L.: SCUT-CAB: A new benchmark dataset of ancient Chinese books with complex layouts for document layout analysis. In: ICFHR (2022)
3. Cheng, H., Zhang, P., Wu, S., Zhang, J., Zhu, Q., Xie, Z., Li, J., Ding, K., Jin, L.: M^6Doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In: CVPR. pp. 15138–15147 (Jun 2023)
4. Gogawale, S., Bambaci, L., Kurar-Barakat, B., Vasyutinsky Shapira, D., Stökl Ben Ezra, D., Dershowitz, N.: NetLay: Layout classification dataset for enhancing layout analysis. Magazén: Intl. J. for Digital and Public Humanities **5**, 1–14 (2024)
5. Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T.T.H., Hackl, G., Moreno, J.G., Doucet, A.: A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In: SIGIR '21. pp. 2328–2334 (2021)
6. Josi, F., Wartena, C., Heid, U.: Preparing legal documents for NLP analysis: Improving the classification of text elements by using page features. In: 8th Intl. Conf. on Natural Language Processing (NATP). vol. 12 (Jan 2022)
7. Lee, J., Hayashi, H., Ohyama, W., Uchida, S.: Page segmentation using a convolutional neural network with trainable co-occurrence features. In: ICDAR. pp. 1023–1028 (2019)
8. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR (2022)
9. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.W.J.: DocLayNet: A large human-annotated dataset for document-layout segmentation. In: KDD '22. pp. 3743–3751 (2022)
10. Saha, R., Mondal, A., Jawahar, C.V.: Graphical object detection in document images. In: ICDAR. pp. 51–58 (2019)
11. Santos Júnior, E.S.d., Paixão, T., Alvarez, A.B.: Comparative performance of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 for layout analysis of historical documents images. Applied Sciences **15**(6) (2025)
12. Shen, Z., Zhang, K., Dell, M.: A large dataset of historical Japanese documents with complex layouts. In: CVPR Workshops. pp. 2336–2343 (Jun 2020)
13. Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., Ingold, R.: DIVA-HisDB: A precisely annotated large dataset of challenging medieval manuscripts. In: ICFHR. pp. 471–476 (2016)
14. Stokes, P., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., Gargem, E.H.: The eScriptorium VRE for manuscript cultures. Classics@ Journal **18** (2021)
15. Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Giles, C.L.: Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: CVPR. pp. 4342–4351 (2017)
16. Zhang, C., Ibrayim, M., Hamdulla, A.: A methodological study of document layout analysis. In: VRHCIAI. pp. 12–17 (2022)
17. Zhong, X., Tang, J., Yepes, A.J.: PubLayNet: largest dataset ever for document layout analysis. In: ICDAR. pp. 1015–1022 (Sep 2019)
18. Zottin, S., De Nardin, A., Branca, G., Piciarelli, C., Foresti, G.L.: ICDAR 2025 competition on few-shot text line segmentation of ancient handwritten documents (FEST). In: ICDAR. pp. 586–602. Cham (2026)
19. Zottin, S., De Nardin, A., Colombi, E., Piciarelli, C., Pavan, F., Foresti, G.L.: U-DIADS-Bib: A full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts. Neural Comput. Appl. **36**(20), 11777–11789 (Jan 2024). https://doi.org/10.1007/s00521-023-09356-5