

Bounding Box Label Propagation for Re-Annotation of Document Layout Analysis Datasets

Nick Jochum^{*1}, Tobias Alt-Veit¹, Christian Schön¹,
Alexander Lück¹, René Schuster^{2,3}, and Didier Stricker^{2,3}

¹ Insiders Technologies GmbH, Kaiserslautern, Germany
{n.jochum, t.alt-veit, c.schoen, a.lueck}@insiders-technologies.de

² DFKI – German Research Center for Artificial Intelligence
{rene.schuster, didier.stricker}@dfki.de

³ RPTU – University Kaiserslautern-Landau, Germany

* corresponding author

Abstract. Datasets in practical document processing scenarios typically grow over time, and their class annotations undergo continuous refinement. This creates significant re-annotation efforts, which are time-consuming and costly. A promising remedy is to re-annotate only a small subset of available documents manually and apply semi-supervised learning techniques that leverage both labelled and unlabelled data. Although there are numerous approaches to tackle this problem for classification, there exists no adaptation for the problem of re-classifying object detection instances, e.g. for document layout analysis. To this end, we propose Bounding Box Label Propagation (BBLP), a pseudo-labelling framework for object detection. An object encoder integrates visual, textual, and positional embeddings from object detection samples to come up with a joint embedding that can be used for Label Propagation on partially annotated datasets in a plug-and-play fashion. Evaluation results indicate that the proposed approach produces high-quality class annotations of bounding boxes. In the D⁴LA layout analysis dataset, it achieves a mAP of 54.0%, corresponding to 81.6% of fully supervised performance, while using only 10% labelled data. Our work demonstrates the potential of Label Propagation for object detection and lays the groundwork for reducing manual annotation efforts in real-world document processing applications.

Keywords: Label Propagation · Object Detection · Document Layout Analysis · Pseudo-Labelling

1 Introduction

The task of Document Layout Analysis (DLA) is a fundamental step in automatic document processing, as it provides a comprehensive understanding of the physical layout of documents [4]. This serves as an important prerequisite for downstream

tasks such as table analysis [24] and key information extraction [15], as well as the overarching goal of full document understanding [32].

Modern layout analysis approaches address this task by training object detection architectures based on CNNs [12, 27], vision transformers [9], or vision-language models [19, 35, 38] on large annotated document datasets [20, 40].

However, existing research datasets are too static for industrial use case demands. In fact, the required datasets in practice grow continuously and undergo repeated refinements w.r.t. the object classes [11], while the object bounding boxes are often retained. Such refinements are demanding: Splitting a single class into two more refined subclasses already requires a full re-annotation pass through the dataset. These novel classes are often introduced as a refinement of already existing base classes and maintain their bounding box.

Annotation effort can be reduced by re-annotating only a small fraction of the dataset and applying semi-supervised learning [10] to leverage both the re-annotated and the remaining data. However, standard approaches for semi-supervised object detection [2, 30] do not exploit bounding box information on the unlabelled samples, as they assume such information is unavailable, thereby discarding potentially valuable information.

We on the other hand, approach the problem as a transductive classification problem and employ Label Propagation [41, 43] for the re-classification of bounding boxes. This method iteratively propagates class labels from a small labelled sample set to a larger unlabelled one. Our proposed architecture allows using Label Propagation in a plug-and-play fashion for the task of re-classifying object detection labels.

Our Contribution. We introduce Bounding Box Label Propagation (BBLP), a compact framework to adapt Label Propagation to class pseudo-labelling for document layout objects from a limited amount of labelled data.

Therein, we leverage visual, textual, and positional embeddings to come up with a joint multimodal representation of document layout objects. This representation is trained on a generic, large document layout dataset through a surrogate classification problem. As a result, BBLP creates suitable embedding vectors that can be used for Label Propagation on so-far unseen and only partially labelled datasets, even with different class definitions.

We evaluate the performance of BBLP on several large-scale public document datasets. Specifically, we assess the quality of the automatically generated bounding box class annotations using two evaluation protocols: The first one compares the generated pseudo-labels directly with the corresponding ground-truth annotations, while the second one analyses the performance of an object detector trained on these noisy pseudo-labels. Moreover, we demonstrate the importance of each included modality through an extensive ablation study. Our results indicate that BBLP can recover the majority of supervised training performance from only a limited set of labelled data.

Thus, the presented BBLP framework provides a straightforward option to re-annotate large-scale datasets with significantly reduced manual effort.

2 Related Work

Object detection is concerned with simultaneously localizing and classifying objects in images. Early deep learning methods employed a two-stage approach, first proposing candidate regions and then classifying them and fitting a bounding box [12, 27]. Single-stage detectors later emerged as a computationally efficient alternative, directly predicting bounding boxes and class labels in a unified network [22, 26]. More recently, transformer-based architectures have achieved state-of-the-art performance by eliminating rule-based design choices such as anchor boxes and non-maximum suppression [5, 39].

Document Layout Analysis can be viewed as an instance of object detection, where the goal is to identify and classify layout elements such as text blocks, figures, and tables within document images [4]. Traditional approaches relied on rule-based heuristics, but modern approaches usually fine-tune object detection transformers for the downstream layout analysis task. The LayoutLM family introduced multimodal pre-training that jointly models text, layout position, and visual features [38]. Similarly, the Document Image Transformer (DiT) [19] employs self-supervised pre-training on large document corpora.

While large-scale benchmarks for document layout analysis [25, 40] have enabled significant progress, they most often do not capture the dynamics of practical scenarios. In real-world document processing, datasets grow incrementally, and class taxonomies evolve as new document types are encountered, or finer distinctions are required [11]. Each change in class ontology requires a complete re-annotation pass through the dataset, which comes with high manual efforts. This motivates semi-supervised approaches that can propagate annotations from a small manually labelled subset to the remaining data.

Semi-supervised learning has a rich history in classification tasks [6, 10], with pseudo-labelling [28] and consistency regularization [33] emerging as dominant paradigms. Extending these ideas to object detection presents additional challenges, as both localization and classification must be addressed simultaneously.

Contemporary semi-supervised object detection methods predominantly adopt a teacher-student framework. STAC [30] pioneered this direction with an offline pseudo-labelling approach that generates complete bounding box annotations on unlabelled images. Unbiased Teacher [23] introduces an online exponential moving average framework with focal loss [22] to mitigate class imbalance. Soft Teacher [37] further refines this approach by weighting classification losses according to prediction confidence and employing box jittering to improve localization quality. Subsequent works such as Humble Teacher [31], Dense Teacher [42], and PseCo [18] propose various enhancements including soft pseudo-labels, dense prediction maps, and explicit mechanisms for localization precision.

Recent efforts have begun exploring semi-supervised learning specifically for document layout analysis. Kallempudi et al. [17] apply it to graphical object detection in documents, while SemiDocSeg [2] combines pseudo-labelling with domain adaptation for layout segmentation.

Notably, all the aforementioned semi-supervised methods generate pseudo-labels for both class assignments and bounding box coordinates. This design

choice is appropriate when no prior annotations exist, but introduces unnecessary complexity when bounding boxes are already available. Our setting fundamentally differs: we assume that bounding box annotations already exist, but have outdated or incomplete class labels. This scenario arises naturally when class taxonomies evolve due to changing requirements. Rather than generating complete pseudo-boxes, BBLP focuses exclusively on re-classifying existing bounding boxes without modifying their spatial coordinates. This decomposition allows us to leverage Label Propagation, a method specifically designed for classification problems.

Label Propagation is a graph-based semi-supervised learning method that iteratively spreads labels from annotated to unannotated samples based on their similarity. The foundational algorithm was introduced by Zhu and Ghahramani [43], who formulated Label Propagation as computing harmonic functions on a similarity graph. Concurrently, Zhou et al. [41] proposed a related method termed label spreading, which incorporates a normalized graph Laplacian and allows soft label assignments.

The key insight underlying these methods is the smoothness assumption: Samples that are close in feature space are likely to share the same label. By constructing a graph where edge weights reflect pairwise similarities, labels can be propagated along paths of high similarity even when direct neighbours are unlabelled. This transductive approach leverages the global structure of the data distribution rather than relying solely on local decision boundaries.

Label Propagation has seen renewed interest in the deep learning era. Iscen et al. [16] demonstrated that propagating labels in the embedding space of neural networks yields competitive semi-supervised classification results. More recently, Huang et al. [14] showed that combining Label Propagation with simple base predictors can match or exceed graph neural network performance, highlighting the continued relevance of these classical methods.

We build upon this foundation by adapting Label Propagation to the object detection setting. Rather than operating on image-level features, BBLP constructs embeddings for individual bounding box instances by integrating visual, textual, and positional information. This multimodal representation enables effective Label Propagation across layout objects, allowing class labels to spread from a small annotated subset to the full dataset while preserving the original bounding box coordinates.

3 Bounding Box Label Propagation

Our pseudo-labelling framework consists of three core components. Given the complete set of all bounding boxes in a dataset and their corresponding document images, we first create several embeddings tailored specifically towards the DLA task. These embeddings take visual and textual features of individual bounding boxes as well as their position, size, and surrounding context into account. Based on those individual layout object embeddings, we then come up with a uniform embedding that combines all relevant information in a single latent space. Using this uniform embedding, we finally perform automatic Label Propagation to

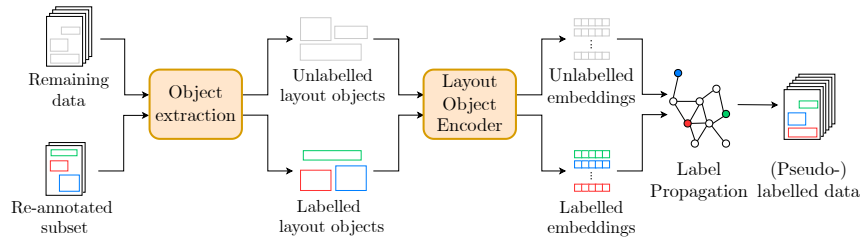


Fig. 1. BBLP framework overview. For all extracted object regions, the proposed multi-modal Layout Object Encoder embeds all layout object into a common vector space. Label Propagation then exploits this representation and generates pseudo-labels for unlabelled objects by transferring label information from the labelled objects.

re-label all objects in the given dataset, based on labels provided for a small subset of objects. Figure 1 provides an overview of the proposed framework.

3.1 Input Modalities

Common DLA tasks include a multitude of different classes, which usually require different modalities to capture all relevant information for classification. Images or diagrams often rely on visual features to correctly locate and classify them in the context of layout analysis. Text-based classes, such as paragraphs or footnotes, are better distinguished by textual features. Headings or image captions in contrast often require positional context, such as their locations within the document or relative positions of neighbouring objects, for accurate distinction. To correctly capture all relevant multi-modal information, we propose a set of embedding functions tailored specifically to the problem of bounding box re-annotation.

Visual Embedding. BBLP leverages a visual embedding model to encode the visual appearance of individual layout objects. Many of those embedding models have the drawback that they rely on a fixed image resolution. As the bounding boxes of layout elements naturally vary in their size and aspect ratio, a visual embedding model in the context of BBLP needs to cope with arbitrary input resolutions. We therefore opted to use the Native-Aspect Flexible Vision Transformer (NaFlexViT), which combines aspects of several recent lines of research [3, 8, 34]. Based on the cropped image segment of each bounding box, it creates a feature vector of size 768 capturing generic visual information associated with a layout element.

Textual Embedding. Textual features form the second building block in our collection of embeddings. To create them, we first have to retrieve the text from a given document. We opted to use the Tesseract open source OCR engine [29] as backbone for text retrieval to facilitate reproducibility. Experiments with

proprietary OCR engines yielded only marginal performance differences ($< 1\%$). Based on the text contained in a given bounding box, a text embedding model is used to create an abstract representation of the textual content. We again opted to use an open weight model, the multilingual E5 text embedding model [36] in its small variant, to retrieve a feature vector of size 384.

Positional Embedding. The BBLP framework extracts positional embeddings directly from annotation data, to capture relative size and position of the layout object, as well as relations to co-occurring layout objects. Specifically, the embedding consists of: (1) the object’s own bounding box normalized by the document width and height, (2) the normalized bounding box of its nearest neighbour under Euclidean closest point distance, (3) their distance normalized by the document’s diagonal, and (4) the total number of objects within the document, normalized by the dataset-wide maximum. Positional embeddings only rely on the given information of bounding box sizes and positions, making them agnostic to class labels.

3.2 Layout Object Encoder

We propose a novel multimodal encoder architecture, the Layout Object Encoder (LOE), designed to project multiple representations of layout objects in a unified embedding space. By training a model using the LOE architecture, BBLP aims to learn a representation integrating visual, textual, and positional information. The model is trained through layout object classification, with the internal representation serving as the layout object embedding.

Figure 2 provides an overview of the LOE architecture. The model processes the positional input embedding vectors through three blocks, each consisting of a batch normalization layer followed by a fully connected layer with 64 output channels and ReLU activation. The LOE then concatenates this vector with the visual and textual input representations and processes the result through two analogous blocks that preserve dimensionality. A final fully connected layer produces classification outputs matching the number of classes. The layout object embeddings consist of the features output at the first fully connected layer after concatenation, as illustrated in Figure 2, separating backbone from classification head. Empirical analysis determined this separation as optimal for generating general and transferable layout object representations, distinct from surrogate-specific classification layers.

Given visual, textual, and positional input representations, BBLP trains the LOE to learn a weighting of the individual input modalities and integrate them into a unified representation. Through this multimodal feature fusion, the LOE learns to predict class labels of layout objects given their different input modalities. By training on a large and diverse DLA dataset, this approach learns a layout object representation generalizing across datasets and class definitions.

Specifically, the LOE is trained on DocLayNet [25]. Every epoch, the embedding quality is validated by performing label propagation on the PRImA

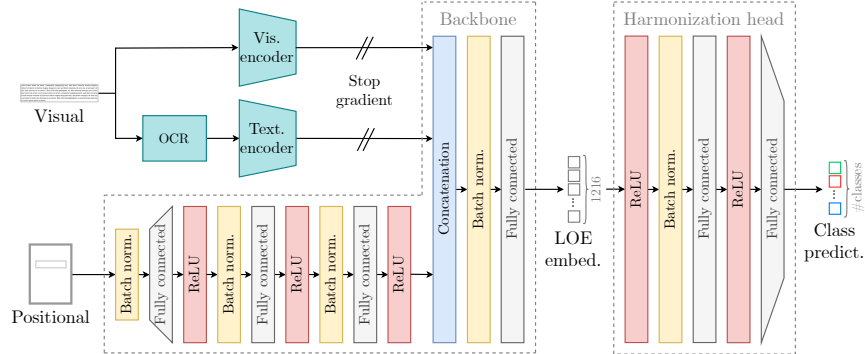


Fig. 2. Overview of the Layout Object Encoder architecture. The encoder maps visual, textual, and positional representations of a layout object to a unified embedding. During training, the harmonization head provides supervision through a surrogate classification task.

layout analysis dataset [1]. For BBLP, we use the LOE checkpoint with the highest pseudo-label accuracy obtained during validation. This selection strategy identifies the checkpoint that generalizes best across datasets and class definitions for predicting pseudo-labels with Label Propagation on unseen datasets.

Implementation Details. The LOE is trained for 30 epochs on the DocLayNet training split with a base learning rate of 10^{-4} , cosine scheduling and cross-entropy loss with label smoothing. Epoch-wise evaluation employs label propagation on 10 problem instances with 10% re-annotated documents, generated from the PRIMa dataset train split. All experiments are executed on a single NVIDIA RTX 3090 GPU.

3.3 Graph-based Label Propagation

Label propagation [41, 43] is an algorithm addressing a transductive classification problem. It exploits the transductive setting where both labelled and unlabelled instances are available upfront and label predictions are only required for the fixed unlabelled set. Based on the smoothness assumption, Label Propagation leverages the spatial distribution across both labelled and unlabelled data, enabling accurate predictions, even in low supervision settings. Since all bounding box annotated layout objects of a dataset are initially available in the considered problem setting, the algorithm is particularly well-suited for our framework.

Specifically, the algorithm constructs a nearest-neighbour graph where nodes represent samples, each connecting to its $k = 20$ nearest neighbours with edge weights reflecting pairwise cosine similarities. In each iteration, the algorithm propagates label information through the graph and fuses it with the initial label information. This process iterates until convergence. Then, the algorithm assigns

Table 1. Summary of experimental datasets showing total number of documents, layout objects, and object categories per dataset. The USAGE column indicates whether we use a dataset for training, validation or testing.

DATASET	DOCUMENTS	OBJECTS	CLASSES	USAGE
DocLayNet [25]	80 863	1 107 470	11	train
PRImA [1]	478	9959	6	val
D ⁴ LA [7]	11 092	146 846	27	test
DocBank _{10%} [20]	50 000	671 596	13	test
PubLayNet _{10%} [40]	35 837	352 289	5	test

to each unlabelled sample the label corresponding to the highest class score. Zhou et al. [41] derived a closed-form solution for the convergent state, enabling an efficient implementation. For further details, we refer to [41, 43].

4 Experimental Results

The evaluation of our BBLP framework addresses two aspects. It measures firstly the quality of the generated pseudo-labels and secondly, their effectiveness as training data for DLA models. We assess each aspect separately using two evaluation protocols. Additional experiments investigate the contribution of individual input modalities and compare different DLA model architectures when training with BBLP pseudo-labels.

4.1 Experimental Setup

Since BBLP learns a layout object representation generalizing well across datasets, the LOE trains, validates, and evaluates on distinct datasets. While BBLP trains on DocLayNet [25] and validates on PRImA, evaluation employs three common DLA benchmarks: D⁴LA [7], DocBank [20] and PubLayNet [40]. To limit computational cost, DocBank and PubLayNet evaluation use 10% subsets, denoted DocBank_{10%} and PubLayNet_{10%}, respectively. Table 1 provides an overview of the size and complexity of each dataset.

To generate a problem instance consisting of a partially re-annotated dataset, evaluation protocols randomly remove label information from a subset of available documents. The remaining documents represent the manually re-annotated fraction, and the objective is to predict pseudo-labels for the removed annotations. Evaluation assesses pseudo-label accuracy across different fractions of manually re-annotated data, denoted by their re-annotated percentage.

4.2 Pseudo-Label Quality

To assess pseudo-label quality, we evaluate BBLP on the train split of each evaluation dataset under the re-annotated fractions of 1, 5, 10, and 20 percent.

Table 2. Evaluation results of BBLP default configuration compared to linear classifier baselines using LOE and visual input features. For each dataset and re-annotation setting, the table reports the mean and standard deviation of pseudo-label accuracies across 10 random instances in percent.

DATASET	RE-ANNOTATION FRACTION			
	1%	5%	10%	20%
<i>Linear classifier baseline (LOE features)</i>				
D ⁴ LA	47.49 ±0.94	54.85 ±0.53	57.35 ±0.58	60.57 ±0.68
DocBank _{10%}	67.24 ±3.60	74.66 ±0.37	75.13 ±0.44	75.55 ±0.40
PubLayNet _{10%}	91.20 ±0.36	93.55 ±0.09	94.04 ±0.04	94.43 ±0.03
<i>Linear classifier baseline (visual features)</i>				
D ⁴ LA	42.40 ±0.94	47.52 ±0.99	48.44 ±1.54	49.32 ±0.86
DocBank _{10%}	67.11 ±3.44	70.80 ±1.17	70.02 ±4.54	72.88 ±3.22
PubLayNet _{10%}	92.92 ±0.15	93.69 ±0.07	93.95 ±0.05	94.21 ±0.06
<i>BBLP (ours)</i>				
D ⁴ LA	57.32 ±0.58	64.71 ±0.38	67.65 ±0.27	69.42 ±0.21
DocBank _{10%}	78.90 ±0.18	81.55 ±0.07	82.08 ±0.07	82.32 ±0.05
PubLayNet _{10%}	93.95 ±0.11	94.75 ±0.05	94.93 ±0.04	95.02 ±0.04

For each setting, we sample 10 problem instances and measure accuracy on the unlabelled fraction of the dataset.

To the best of our knowledge, no prior work in the DLA domain addresses the re-annotation of bounding boxes. Therefore, we cannot compare our approach to existing methods. To establish a baseline, we compare Label Propagation to a simple linear classifier with softmax activation, as a method for generating pseudo-labels from layout object representations. Two variants of the linear classifier are tested. The first variant uses our LOE features, while the second uses visual input features only. Training is started from scratch on the re-annotated fraction of each problem instance for 100 epochs with a learning rate of 0.1. The trained classifier then predicts bounding box pseudo-labels for the unlabelled documents.

Table 2 reports the results for the pseudo-label quality evaluation protocol. Across all datasets and approaches, pseudo-label accuracy increases with additional supervision, with diminishing gain at higher supervision settings. Results are strongest for PubLayNet_{10%}, while D⁴LA exhibits the lowest average accuracy scores, reflecting the respective number of classes. BBLP achieves the highest average accuracies across all datasets and supervision settings, exceeding 90% on PubLayNet_{10%}, around 80% on DocBank_{10%} and below 70% on D⁴LA. Both linear classifier baselines reach considerably lower scores. The margins are most pronounced for D⁴LA, where BBLP outperforms the LOE features linear classifier by roughly 10 percentage points and the visual features baseline by more than 15 percentage points across all supervision settings. On DocBank_{10%} margins remain smaller yet significant, with BBLP surpassing the LOE features and visual

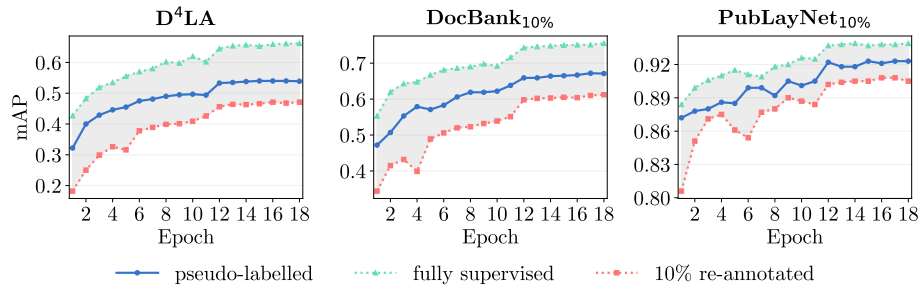


Fig. 3. Training performance of a DINO object detector trained on BBLP pseudo-labels. For each evaluation dataset, curves show mAP scores on the test split during training for models trained on pseudo-labelled, 10%, and 100% ground-truth annotated documents.

features baselines in all settings by 6 and 9 percentage points, respectively. Improvements on PubLayNet_{10%} are notably smaller, as both baselines consistently achieve average accuracies over 91%, yet BBLP yields a consistent advantage.

4.3 Document Layout Analysis with BBLP

Object detection models have demonstrated robustness against label-noise to a certain extent [13]. However, training with BBLP pseudo-labels only remains practical if DLA models can tolerate the inherent amount of label noise. To further investigate the BBLP pseudo-label quality and to examine the effect of the inherent label noise on the downstream task of DLA, we train an object detector on different label configurations. This evaluation compares the generated pseudo-labels against a gold standard and a lower performance bound baseline as DLA training data.

For the train split of each evaluation dataset, we randomly sample a single problem instance using 10% re-labelled data, generate pseudo-labels with BBLP for the unlabelled fraction, and train on the complete dataset, including manually and pseudo-labelled documents. The gold standard baseline model trains on the fully ground-truth annotated dataset, while the lower baseline uses only the 10% ground-truth data, representing the manually re-annotated documents.

After each epoch, we evaluate the current model checkpoint on the corresponding dataset test split and report mAP across all training epochs.

All models trained in this experiment follow the DINO architecture with a ResNet-50 backbone and 4-scale feature maps [39]. In each configuration, the model is initialized with weights pre-trained on the COCO dataset [21] and is trained for 18 epochs on the corresponding evaluation dataset. The remaining hyperparameters follow those in [39].

Figure 3 visualizes the mAP progression on the test splits during training for the different label configuration and evaluation datasets. Across all evaluation datasets and epochs, the pseudo-label trained model consistently exceeds the lower baseline trained on 10% manually re-annotated documents, while remaining

below gold standard trained on the full ground-truth annotated dataset. On D⁴LA, BBLP reaches a converged mAP of 53.9%, compared to 47.1% and 66.0% for the lower and gold standard baseline, respectively. Similarly, on DocBank_{10%} the BBLP pseudo-labels trained model achieves finally 67.1% mAP, surpassing the lower baseline by 5.9 percentage points and falling 8.4 percentage points short to the gold standard baseline. On PubLayNet_{10%}, performance of all three models is comparable. The pseudo-label trained model reaches a mAP score of 92.3% at convergence, which is only 1.8 percentage points above the lower and 1.6 percentage points below gold standard baseline.

4.4 Additional Experiments

To better understand the contribution of each modality and to evaluate an alternative DLA model trained on the generated pseudo-labels, we conduct two additional experiments.

Influence of Individual Modalities. As outlined in Section 3.2, BBLP multi-modal feature encoding integrates visual, textual, and positional information in a joint embedding. We hypothesize that training on a large and diverse dataset enables the LOE to appropriately weight the influence of individual input modalities, yielding a layout object representation superior to any individual input representation. To investigate this hypothesis, this ablation compares the default LOE representation against visual and textual input modality representations, as well as against LOE variants integrating only one or two modalities.

These LOE variants concatenate only the remaining features in the concatenation layer. The resulting feature vector size determines the dimensionality of all subsequent harmonization layers, reducing the overall parameter amount compared to the standard LOE. Training follows the procedure outlined in Section 3.2. For evaluating each representation setting, this ablation adopts the pseudo-label quality evaluation protocol outlined in Section 4.2 and reports results only for the 10% re-labelled setting.

Table 3 reports the pseudo-label accuracies for all modality settings. On D⁴LA and DocBank_{10%}, the original BBLP configuration using all three input modalities reaches the highest pseudo-label accuracies, while on PubLayNet_{10%} it can compete with the visual input embedding. Among single-modality LOE variants, the text modality contributes most to performance on D⁴LA, whereas the visual modality achieves the highest individual scores on DocBank_{10%} and PubLayNet_{10%}. The positional embedding achieves the lowest scores across all datasets. Each multi-modality configuration consistently outperforms the corresponding single-modality configurations, confirming that the LOE leverages complementary information across modalities.

Alternative DLA Model Architecture. The experiment in Section 4.3 demonstrates that DLA models can generally exploit additional supervision provided by BBLP pseudo-labels, despite the inherent amount of label noise.

Table 3. Impact of individual modalities on BBLP pseudo-label accuracy. We report average accuracies and standard deviations with 10% re-labelled data, comparing raw visual and textual input features against features from LOE variants trained on combinations of modalities. Bold values indicate the best and underlined values the second-best result per dataset.

MODALITIES			DATASET		
vis.	text.	pos.	D ⁴ LA	DocBank _{10%}	PubLayNet _{10%}
<i>Input embeddings</i>					
✓			56.64 ± 0.26	81.53 ± 0.07	95.07 ± 0.05
	✓		62.86 ± 0.32	73.47 ± 0.16	84.81 ± 0.10
<i>LOE embeddings</i>					
✓			55.61 ± 0.28	78.73 ± 0.06	94.39 ± 0.03
	✓		61.14 ± 0.33	71.87 ± 0.10	83.81 ± 0.06
		✓	38.25 ± 0.12	63.04 ± 0.06	80.97 ± 0.05
	✓	✓	65.02 ± 0.29	75.56 ± 0.15	88.15 ± 0.12
✓		✓	57.52 ± 0.23	79.19 ± 0.06	94.45 ± 0.03
✓	✓		<u>66.81</u> ± 0.24	<u>81.76</u> ± 0.07	94.85 ± 0.04
✓	✓	✓	67.65 ± 0.27	82.32 ± 0.05	<u>94.93</u> ± 0.04

However, the evaluation considers only one specific model architecture. This experiment evaluates an alternative DLA model architecture trained on the generated pseudo-labels, to verify that noise tolerance generalizes across model architectures, rather than being specific to DINO model architecture.

Specifically, this experiment trains a two-stage Faster R-CNN [27] with a ResNet-50 backbone for 18 epochs. This experiment focuses exclusively on D⁴LA, as the dataset has the highest pseudo-label noise ratio and label noise effects on training are potentially most pronounced. Analogous to Section 4.3, training uses the same set of generated pseudo-labels and compares the model to a gold standard and a 10% manually re-annotated baseline.

Table 4 reports the final mAP scores for all label configurations and both model architectures. Similar to the DINO architecture, the Faster R-CNN pseudo-label configuration consistently outperforms the lower baseline by a significant margin but performs below the gold standard, though the relative margin to the lower baseline is smaller compared to DINO. Furthermore, Faster-CNN converges at notably lower mAP values than DINO across all label configurations, which can be attributed to DINO’s inherent performance advantage over Faster R-CNN, as documented in [39].

5 Discussion

Section 4.2 demonstrates that BBLP can reliably assign correct labels to the majority of layout objects, even under minimal supervision and for datasets with numerous classes such as D⁴LA. A higher fraction of manually re-annotated

Table 4. Performance comparison across DLA architectures. The table reports final mAP scores in percent on D⁴LA after training under three label configurations: BBLP pseudo-labels, fully supervised gold standard baseline and 10% manually re-annotated baseline.

DLA MODEL	PSEUDO-LAB.	FULLY SUP.	10% RE-ANN.
DINO	53.9	66.2	47.1
Faster R-CNN	42.0	54.2	37.9

documents leads to improved pseudo-label accuracy on the remaining documents. This gain is most pronounced under low supervision, as additional labelled samples become increasingly redundant and contribute less novel information with growing supervision. Furthermore, the marginal accuracy gains from additional supervision correlate with the number of classes in the dataset. Such behaviour aligns with the intuition that datasets with many classes require more labelled instances to represent each class adequately. Standard deviations remain negligibly low for all datasets and re-labelled fractions, indicating minimal dependency on the choice of documents for manual re-annotation.

The performance advantage of BBLP compared to the linear classifier baselines likely stems from the transductive nature of the Label Propagation algorithm. While linear classifiers rely exclusively on labelled instances as training data for generating pseudo-labels, Label Propagation additionally leverages the spatial distribution of unlabelled data to produce label predictions. Moreover, the LOE feature baseline outperforms the visual feature baseline on D⁴LA and DocBank_{10%}, but shows comparable performance on PubLayNet_{10%}. This trend likely reflects dataset complexity: PubLayNet_{10%} contains only five visually distinguishable classes, while textual and positional cues integrated in the LOE features prove advantageous for more complex class structures.

Furthermore, experiments demonstrate that BBLP pseudo-labels are well-suited for DLA model training. Even under high noise ratios of almost 30% for the D⁴LA dataset, the models leverage the additional supervision and outperform the baseline trained on only manually re-annotated documents by a significant margin. In comparison to the gold standard baseline, the pseudo-label trained model’s relative mAP surpasses the label accuracy of its training data. This finding supports the hypothesis that DLA models tolerate label noise to a certain extent and exploit incorrectly labelled bounding boxes, as positional information remains correct and semantic similarity often exists between predicted and ground-truth labels. These trends persist across DLA model architectures in our experiments, indicating general applicability.

As the results of the ablation study demonstrate, the quality of LOE features consistently improves when using more modalities. During LOE training, the encoder learns to effectively integrate multi-modal input features in a joint general embedding through appropriate weighting. Across modalities, the textual features have the greatest impact on Label Propagation performance for D⁴LA, while visual embeddings contribute most for DocBank_{10%} and PubLayNet_{10%}. A likely

explanation lies again in the different class definition complexities. D⁴LA contains 27 classes, many of which are text-based with similar visual representations. Conversely, the five classes of PubLayNet_{10%} are plausibly distinguishable by visual features alone. The positional modality achieves the lowest individual performance across datasets, yet consistently improves representation quality when combined with other modalities. Evidently, positional information provides valuable complementary cues, despite being insufficient for classification alone.

Limitations. While we demonstrate the potential of Label Propagation for semi-automatic DLA dataset re-annotation, several limitations need to be acknowledged. The BBLP framework incorporates spatial information from the annotation data, but does not consider class labels prior re-annotation. Given that such refinements typically follow hierarchical patterns, original class labels often hint at semantic relationships and can provide valuable cues to predict refined class labels. Furthermore, this work assesses the quality of pseudo-labels only by label accuracy across the entire dataset. However, layout object classes are naturally highly imbalanced. Per-class evaluation metrics could reveal insights in the structure of label noise, such as common class confusions.

6 Conclusions and Future Work

This work addresses the challenge of reducing manual re-annotation effort for taxonomy adaptation of DLA datasets. Given a small subset of manually re-annotated documents, the proposed BBLP framework generates pseudo-labels for layout objects in the remainder of the dataset. The novel LOE model architecture integrates visual, textual, and positional features into unified layout object representations. BBLP leverages these representations to transfer label information from labelled to unlabelled object instances in the dataset.

Experiments demonstrate that all three modalities provide complementary information, which the LOE effectively integrates into a higher-performing joint representation. Moreover, Label Propagation proves particularly effective for the semi-automatic dataset re-annotation, since the algorithm leverages spacial structure of instance representations across both labelled and unlabelled data. Despite the inherent amount of label noise, DLA models can take advantage of the additional supervision provided by BBLP pseudo-labels, improving performance significantly compared to training on manually re-annotated documents alone.

The identified limitations provide directions for future research. One promising extension would integrate prior class label information in the BBLP framework. For instance, the Label Propagation k -nearest-neighbour graph could enhance edge weights based on original class agreement, facilitating label information flow between semantically related objects. Additionally, we plan to investigate our approach in conjunction with established multimodal encoders, such as LayoutLM [15]. Both directions offer potential for substantially reducing manual re-annotation effort when refining class definitions of DLA datasets.

References

1. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: Proc. 10th Int. Conf. on Document Analysis and Recognition. pp. 296–300. Barcelona, Spain (Jul 2009)
2. Banerjee, A., Biswas, S., Lladós, J., Pal, U.: SemiDocSeg: Harnessing semi-supervised learning for document layout analysis. *Int. J. on Document Analysis and Recognition* **27**(3), 317–334 (Jun 2024)
3. Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., Pavetic, F.: Flexivit: One model for all patch sizes. In: Proc. 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 14496–14506. Vancouver, BC, Canada (Jun 2023)
4. Binmakhshen, G.M., Mahmoud, S.A.: Document layout analysis: A comprehensive survey. *ACM Comput. Surv.* **52**(6), article no. 109 (Oct 2019)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *Computer Vision - ECCV 2020, Lecture Notes in Computer Science*, vol. 12346, pp. 213–229. Springer, Cham (2020)
6. Chapelle, O., Schölkopf, B., Zien, A.: *Semi-Supervised Learning*. The MIT Press (2006)
7. Da, C., Luo, C., Zheng, Q., Yao, C.: Vision grid transformer for document layout analysis. In: Proc. 2023 IEEE/CVF Int. Conf. on Computer Vision. pp. 19405–19415. Paris, France (Oct 2023)
8. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., Oliver, A., Padlewski, P., Gritsenko, A.A., Lucic, M., Houlsby, N.: Patch n’ pack: Navit, a vision transformer for any aspect ratio and resolution. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Proc. 37th Int. Conf. on Neural Information Processing Systems. Advances in Neural Information Processing Systems*, vol. 36. New Orleans, LA (Dec 2023)
9. 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: Proc. 9th Int. Conf. on Learning Representations. Vienna, Austria (May 2021)
10. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. *Mach. Learn.* **109**(2), 373–440 (Nov 2019)
11. Gemelli, A., Marinai, S., Pisaneschi, L., Santoni, F.: Datasets and annotations for layout analysis of scientific articles. *Int. J. on Document Analysis and Recognition* **27**(4), 683–705 (Nov 2024)
12. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 580–587. Columbus, OH (Jun 2014)
13. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proc. 2006 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1735–1742. New York, NY (Jun 2006)
14. Huang, Q., He, H., Singh, A., Lim, S.N., Benson, A.R.: Combining label propagation and simple models out-performs graph neural networks. In: Proc. 9th Int. Conf. on Learning Representations. Vienna, Austria (May 2021)

15. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-training for document AI with unified text and image masking. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) Proc. 30th ACM Int. Conf. on Multimedia. pp. 4083–4091. Lisbon, Portugal (Oct 2022)
16. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 5070–5079. Long Beach, CA (Jun 2019)
17. Kallemputti, G., Hashmi, K.A., Pagani, A., Afzal, M.Z., Stricker, D.: Toward semi-supervised graphical object detection in document images. *Future Internet* **14**(6), article no. 176 (june 2022)
18. Li, G., Li, X., Wang, Y., Wu, Y., Liang, D., Zhang, S.: PseCo: Pseudo labeling and consistency training for semi-supervised object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *Computer Vision - ECCV 2022, Lecture Notes in Computer Science*, vol. 13669, pp. 457–472. Springer, Cham (2022)
19. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: DiT: Self-supervised pre-training for document image transformer. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) Proc. 30th ACM Int. Conf. on Multimedia. pp. 3530–3539. Lisbon, Portugal (Oct 2022)
20. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank: A benchmark dataset for document layout analysis. In: Scott, D., Bel, N., Zong, C. (eds.) Proc. 28th Int. Conf. on Computational Linguistics. pp. 949–960. Barcelona, Spain (Dec 2020)
21. Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision - ECCV 2014, Lecture Notes in Computer Science*, vol. 8693, pp. 740–755. Springer, Cham (2014)
22. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(2), 318–327 (Feb 2020)
23. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: Proc. 9th Int. Conf. on Learning Representations. Vienna, Austria (May 2021)
24. Nassar, A., Livanthinos, N., Lysak, M., Staar, P.: TableFormer: Table structure understanding with transformers. In: Proc. 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 4604–4613. New Orleans, LA (Jun 2022)
25. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: DocLayNet: A large human-annotated dataset for document-layout segmentation. In: Zhang, A., Rangwala, H. (eds.) Proc. 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. pp. 3743–3751. Washington, DC (Aug 2022)
26. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 779–788. Las Vegas, NV (Jun 2016)
27. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. 29th Int. Conf. on Neural Information Processing Systems. *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99. Montreal, QC, Canada (Dec 2015)
28. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: Proc. 9th Int. Conf. on Learning Representations. Vienna, Austria (May 2021)
29. Smith, R.: An overview of the tesseract OCR engine. In: Proc. 9th Int. Conf. on Document Analysis and Recognition. pp. 629–633. Curitiba, Brazil (Sep 2007)

30. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv:2005.04757v2 [cs.CV] (2020)
31. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proc. 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 3132–3141. virtual (Jun 2021)
32. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. In: Proc. 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 19254–19264. Vancouver, BC, Canada (Jun 2023)
33. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Proc. 31st Int. Conf. on Neural Information Processing Systems. Advances in Neural Information Processing Systems, vol. 30, pp. 1195–1204. Long Beach, CA (Dec 2017)
34. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786 [cs.CV] (Feb 2025)
35. Wang, J., Hu, K., Huo, Q.: DLAFormer: An end-to-end transformer for document layout analysis. In: Smith, E.H.B., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024, Lecture Notes in Computer Science, vol. 14807, pp. 40–57. Springer, Cham (2024)
36. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual E5 text embeddings: A technical report. arXiv:2402.05672v1 [cs.CL] (Feb 2024)
37. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proc. 2021 IEEE/CVF Int. Conf. on Computer Vision. pp. 3040–3049. Montreal, Canada (Oct 2021)
38. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of text and layout for document image understanding. In: Proc. 26th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. pp. 1192–1200. virtual (Aug 2020)
39. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proc. 11th Int. Conf. on Learning Representations. Kigali, Rwanda (May 2023)
40. Zhong, X., Tang, J., Jimeno-Yepes, A.: PubLayNet: Largest dataset ever for document layout analysis. In: Proc. 2019 Int. Conf. on Document Analysis and Recognition. pp. 1015–1022. Sydney, Australia (Sep 2019)
41. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) Proc. 17th Int. Conf. on Neural Information Processing Systems. Advances in Neural Information Processing Systems, vol. 16, pp. 321–328. Vancouver and Whistler, Canada (Dec 2003)
42. Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., Sun, J.: Dense teacher: Dense pseudo-labels for semi-supervised object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022, Lecture Notes in Computer Science, vol. 13669, pp. 35–50. Springer, Cham (2022)
43. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical report (2002)