RoDLA: Benchmarking the Robustness of Document Layout Analysis Models

Yufan Chen\(^1\), Jiaming Zhang\(^{1,2,*}\), Kunyu Peng\(^1\), Junwei Zheng\(^1\), Ruiping Liu\(^1\), Philip Torr\(^2\), Rainer Stiefelhagen\(^1\)

\(^1\)Karlsruhe Institute of Technology, \(^2\)University of Oxford

https://yufanchen96.github.io/projects/RoDLA/

Abstract

Before developing a Document Layout Analysis (DLA) model in real-world applications, conducting comprehensive robustness testing is essential. However, the robustness of DLA models remains underexplored in the literature. To address this, we are the first to introduce a robustness benchmark for DLA models, which includes 450K document images of three datasets. To cover realistic corruptions, we propose a perturbation taxonomy with 12 types of perturbations (P1–P12) inspired by real-world document processing, as well as 3 severity levels (L1–L3) for each perturbation. Our RoDLA model method obtains higher mean Robust Degradation (mRD) in all 5 groups of perturbations, and shows stable mAP scores across 3 levels of perturbation (e.g., defocus).

1. Introduction

Document Layout Analysis (DLA) is an essential component in document understanding, it indicates a fundamental comprehension of documents. As this field evolves, the shift from electronic to real-world documents presents unique challenges. These challenges are largely due to variable image quality influenced by factors like uneven illumination and human-induced vibrations [24, 50, 56]. These factors introduce additional complexities in DLA, as they can lead to distorted representations of documents, making accurate layout analysis more challenging. According to our observation in Fig. 1c, perturbed document images raise large performance drops of previous state-of-the-art DLA models [4, 5, 54], e.g., 91.0% performance decrease with SwinDocSegmenter [4] from L1 to L3, i.e., increasing perturbation severity in defocus perturbation. It shows a notable weakness of previous models in resisting document

\(115.7, 135.4, \text{ and } 150.4, \text{ respectively. Compared to previous methods, RoDLA achieves notable improvements in mAP of } +3.8\%, +7.1\% \text{ and } +12.1\%, \text{ respectively.}

\(\)
perturbations. However, the robustness of DLA models remains underexplored in the literature.

To fill the gap, we propose an extensive benchmark with almost 450,000 document images from 3 datasets for evaluating robustness of DLA models. The perturbation taxonomy is shown in Fig. 1a. Building on research into document image degradation [15, 16, 24, 27, 29, 50, 56], we categorize all document image perturbations into 5 high-level groups and 12 types of perturbations (P1–P12). Recognizing that perturbations not uniformly impact documents, we include 3 severity levels (L1–L3) for each perturbation.

A suitable metric to evaluate the perturbation effects of document images and model robustness is required for the new benchmark. Previous metrics are constrained by the pre-selected baseline, serving as a reference for perturbation effects [25, 28, 33, 66], which leads to metric uncertainty stemming from inherent model randomness. To address this, we design a perturbation assessment metric, Mean Perturbation Effect (mPE), a combination of traditional image quality assessment methods and model performance. By employing multiple methods to assess the perturbation effects, we mitigate the randomness and inconsistencies in the effect evaluation of different perturbations. Furthermore, we propose Mean Robustness Degradation (mRD), a mPE-based robustness evaluation metric. The mRD results on our benchmark are shown in Fig. 1b. Our metric can minimize the impacts of model randomness and baseline selection, yielding a better measurement of model robustness.

Based on our study of robustness benchmark, we propose a novel robust DLA model, i.e., Robust Document Layout Analyzer (RoDLA). It includes the channel attention to integrate the self-attention. Then, we couple it with average pooling layers to reduce the excessive focus on perturbed tokens. This crucial design enables the model to capture perturbation-insensitive features, thus significantly improving robustness. For instance, Fig. 1c shows the stable performance of RoDLA on the defocus perturbation, while previous methods [4, 5, 54] have large performance drops. Our RoDLA model also achieves 96.0% mAP on the PubLayNet dataset [72]. We obtain 70% in mAP with a +3.2% gain and 116.0 in mRD on the PubLayNet-P benchmark. Besides, our method reaches the state-of-the-art performance on the DocLayNet-P and M6Doc-P datasets, having mAP gains of +7.1% and +12.1%, respectively.

To summarize, we present the following contributions:

- We design a perturbation assessment metric Mean Perturbation Effect (mPE) and a robustness evaluation criteria Mean Robustness Degradation (mRD), which separates the impact of perturbations on images from the intrinsic perturbation robustness of the model, allowing for a more accurate robustness measurement.
- We propose Robust Document Layout Analyzer (RoDLA), which achieves state-of-the-art performance on clean datasets and the robustness benchmarks (PubLayNet-P, DocLayNet-P and M6Doc-P).

2. Related Work

2.1. Document Layout Analysis

Document Layout Analysis is a fundamental document understanding task, which extracts the structure and content layout of documents. Thanks to the diverse datasets and benchmarks [32, 42, 53, 55, 72], machine learning-based approaches [11, 17] and deep learning-based approaches [9, 18, 36, 45, 67, 74] have made progress. Both single-modal methods, e.g., Faster R-CNN [48], Mask R-CNN [23], and DocSegTr [5], and multi-modal methods, e.g., DiT [31] and LayoutLMv3 [26], are well explored in DLA. Besides, text grid-based methods [68, 71] deliver the combination capability of the text grid with the visual features. Recently, transformer models [2, 8, 9, 31, 57, 59, 67] have been explored in DLA. Self-supervised pretraining strategies [1, 26, 34, 37, 38, 64, 65] have also drawn considerable attention in DLA, e.g., DocFormer [1] and LayoutLMv3 [26]. However, existing works in DLA focuses on clean document data, overlooking real-world issues like noise and disturbances. In this work, we aim to fill this gap by benchmarking the robustness of DLA models.

2.2. Robustness of Document Understanding

As a related task, document restoration and rectification is the task of improving the image quality of documents by correcting distortions. DocTr++ [15] explores unrestricted document image rectification. In [16], robustness against adversarial attack is investigated for document image classification. Auer et al. [3] propose a challenge for robust document layout segmentation. To address this, Zhang et al. [70] propose a wechat layout analysis system. The robustness evaluation on the RVL-CDIP dataset [21] is for document classification. Tran et al. [58] propose a robust DLA system by using multilevel homogeneity structure. However, these works are oriented towards optimizing performance on clean documents. Our research is the first to systematically study real-world challenges of DLA. We benchmark DLA methods with extensive perturbation types, encompassing 3 datasets, 5 perturbation groups, 12 distinct types, and 3 severity levels for each type.
2.3. Robust Visual Architectures

A robust visual architecture is required to maintain reliable visual analysis. Some researches are established in object detection [14, 20, 39, 40, 47, 52, 60] and image classification [13, 41, 43, 44]. Modas et al. [41] introduce few primitives which can boost the robustness in image classification field. R-YOLO [60] propose a robust object detector under adverse weathers. FAN [73] proposes fully attention networks to strengthen the robust representations. A Token-aware Average Pooling (TAP) [19] module is proposed to encourage the local neighborhood of tokens to participate in the self-attention mechanism. However, directly applying existing robust methods to domain-specific tasks like DLA cannot yield optimal performance due to the unique challenges. To address this, a synergistic attention-integrated model is designed to enhance the robustness of DLA by concentrating attention on key tokens on multi-scale features and enhancing attention interrelations.

3. Perturbation Taxonomy

3.1. Hierarchical Perturbations

In the field of Document Layout Analysis (DLA), it is essential to grasp the structure and content arrangement. This is particularly challenging with document images from scans or photos, where processing errors and disturbances can degrade DLA performance. Previous benchmarks [32, 46, 72] have not fully addressed these challenges, as they simply include a collection of real documents images without analysis of perturbations. Therefore, we introduce a robustness benchmark with 450,000 document images, incorporating 12 perturbations with 3 level of severity to systematically evaluate the robustness of DLA models.

Given a DLA model $g:X \rightarrow L$, trained on a digital documents dataset $X$, the model in traditional benchmark is tested by the probability $P_{x,l}(g(x) = l)$. However, real-world documents images often experience various disturbances [24, 27, 50], which are not represented in a digital dataset. We consider a set of document-specific perturbation functions $O$, and approximate the perturbation distribution in real-world with $P_{O}(o)$. To bridge gap from digital to real, our approach evaluates the DLA model’s effectiveness against common perturbations in documents by the expectation $E_{O}P_{x,l}(g(\hat{o}(x))=o(l))$. Through expectation, we can effectively measure the average model performance impact under same perturbation type, rather than merely assessing performance under worst-case perturbation.

As shown in Fig. 1a, we divide the document perturbations into 5 groups, i.e., spatial transformation, content interference, inconsistency distortion, noise and blur. In these 5 groups, there are totally 12 types of document perturbation with 36 severity levels. The visualization of 12 perturbations is shown in Fig. 2. These document perturbation settings are derived from common perturbation observed in scanned documents and photographic document images. They represent the typical challenges and variations encountered in real-world scenarios, ensuring that our approach is both comprehensive and practical. The detailed settings of each perturbation are described as the following.

3.2. Perturbation Description

In Table 1, 5 groups and 12 types of perturbations are mathematically defined and divided into 3 levels of severity.

1 Spatial Transformation. This group involves three perturbations: (P1) Rotation around the center of document without preserving the ratio. The severity is determined by the angle $\theta$, which is randomly chosen from a uniform distribution within predefined ranges. (P2) Warping, which is enacted by generating two deformation fields, $(\Delta x, \Delta y) = \alpha \cdot G(U(x,y), \sigma)$. $U(x,y)$ obeys the uniform distribution, $G$ denotes Gaussian filter, $\sigma$ and $\alpha$ are the smoothness and magnitude controlled by scaling factor $R_{\alpha}$ and $R_{\sigma}$, respectively. (P3) Keystoning, simulated by perspective transformation using homograph matrix $H$, maps the original coordinates to new ones. The corner points are adjusted by offsets drawn from a normal distribution, with standard deviations scaled by factor $R_{\sigma}$. Annotation transformations are adjusted accordingly to alignment.

2 Content Interference. Document content interference stems from two primary sources: (P4) Text Watermark is commonly used as an anti-piracy measure. To simulate the text watermark, we overlay uniformly sized characters at random document positions with random rotations. The interference strength is controlled by adjusting the zoom ratio

![Figure 2. Visualization of document perturbations.](image-url)
R_z and alpha channel opacity α_w. (P5) Complex Background, which is often leveraged in printed media to enrich the document’s visual content. Images from ILSVRC dataset [49], are randomly embedded within the document images for the simulation of background. The severity level is determined by the number of embedded images N_f.

3 Inconsistency Distortion, which addresses distortions, e.g., (P6) Non-uniform Illumination, simulated through a 50% probability of glare or shadow, with levels determined by the brightness V_l or shadow V_s. (P7) Ink-bleeding, achieved by using erosion process within the document image, and (P8) Ink-holdout, which are accomplished by using morphological dilation operations with elliptical kernel K_e and K_d, respectively.

4 Blur. Blur effects arise from (P9) Defocus, which is approximated with Gaussian kernel function, i.e., the point spread function (PSF) with severity levels controlled by kernel size K_g, and (P10) Vibration, which is achieved by using a convolution kernel rotated by a random angle to replicate directional motion in size K_m. This perturbation can be applied via a Gaussian filtering operation.

5 Noise. Two unique noise types in document images are included. (P11) Speckle caused by ink clumping, can be achieved by using Gaussian noise modulated by blob density D_b. It reproduces both foreground and background noise, reflecting the complex stochastic nature of noise artifacts. (P12) Texture, resembling the fibers in paper, is added to simulate the natural fiber structures of archival documents. In each fiber segment, curvature and length are independently simulated with trigonometric functions and Cauchy distribution, forming the complete fiber through segment assembly. The number of fibers N_f varies across noise levels to represent different paper qualities.

The principles of design and division of these perturbations are three-fold: (1) All perturbations are realistic and inspired by the real-world disturbance or document layout analysis. (2) The leveraged perturbations are comprehensive and occur in document from top to bottom, from global to local, from dense to sparse, and from content-wise to pixel-wise. (3) All perturbation levels are reasonably determined by Image Quality Assessment (IQA) metrics, e.g., MS-SSIM [62] and CW-SSIM [51]. Degradation and our proposed metrics which are detailed as the following.

3.3. Perturbation Evaluation Metrics

Quantifying the document perturbations is a crucial task, which lacks straightforward metrics. Previous methods [25, 28, 33, 66] are constrained by using a pre-selected baseline model performance as a reference to measure the perturbation effect, which will confute the perturbation effect with the model robustness due to the inherent variability.

To design an effective perturbation metric, our considerations are two-fold: (1) Evaluation should be relatively independent to any specific model performance. (2) Evaluation should be inclusive and sensitive to all perturbations. Inspired by the IQA methods [12], we analyse all 12 perturbations through 4 metrics, as presented in Fig. 3.

MS-SSIM (Multi-Scale Structural Similarity Index) [62] evaluates image quality by considering difference in structural information across different resolutions, as:

\[
f_{\text{MS-SSIM}}(x,y) = |l_M(x,y)|^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\gamma_j},
\]

where l_M(x,y) represents the luminance difference at the M-th scale, and c_j(x,y), s_j(x,y) are the contrast and structure comparison at the j-th scale. It measures the impact of most perturbations effectively but lacks sensitivity to watermark and warping (Fig. 3b), and is overly sensitive to rotation and keystoning, as shown in Fig. 3a.
Comparison of the sensitivity of metrics to different perturbation levels.

Figure 3. Analysis of perturbation evaluation metrics. (a) Comparison of perturbation metrics, including MS-SSIM, CW-SSIM, Degradation w.r.t a baseline, and the proposed mean Perturbation Effect (mPE). mPE is more balanced and inclusive to different perturbations. (b) Six documents perturbed by warping and defocus and their scores indicate that mPE is more sensitive to measure different levels.

CW-SSIM (Complex Wavelet Structural Similarity Index) [51] assesses the similarity of local patterns of pixel intensities transformed into the wavelet domain, which is robust to spatial distortions, e.g., keystoning and rotation, formulated by the following equation:

\[ f_{\text{CW-SSIM}}(x, y) = \frac{2|\sum_{l=1}^{L} w_l x_l y_l^*| + K}{\sum_{l=1}^{L} |x_l|^2 + \sum_{l=1}^{L} |y_l|^2 + K}, \tag{2} \]

where \( x_l \) and \( y_l \) denote the wavelet coefficients, \( w_l \) is the \( l \)-th coefficient weight, \( L \) indicates the local region number, \( K \) is a tiny value, and \( * \) denotes the complex conjugation. As shown in Fig. 3a, it reflects the perturbations impact but fails for quantitative evaluation at different levels for some perturbations, e.g., defocus, demonstrated in Fig. 3b.

Degradation across severity levels, as indicated by ImageNet-C [25], is a common practice for various robustness benchmark [25, 28, 33, 66]. Degradation \( D \), where \( D=1-mAP \) similar with ImageNet-C, illustrates the performance impact of perturbation on a baseline model. From Fig. 3a, we notice that the Degradation metric is highly dependent on the choice of the baseline and tends to be overly sensitive to perturbations, e.g., background and texture.

Mean Perturbation Effect (mPE) is a new metric proposed to assess the compound effects of document perturbations. For specific perturbation \( p \) at each severity level \( s \), degradation of model \( g \) is written as \( D_{s,p}^g \), and IQA metrics is written as \( f_{s,p}^g \), i.e., from Eq. (1) and Eq. (2). The mPE to measure the effect of perturbation \( p \) is calculated as:

\[ \text{mPE}_p = \frac{1}{NMK} \sum_{s=1}^{N} \left( \sum_{i=1}^{M} f_{s,p}^i + \sum_{g=1}^{K} D_{s,p}^g \right). \tag{3} \]

mPE aims to isolate the model’s robustness from the image alterations impact. We employ MS-SSIM and CW-SSIM as IQA metrics, and we select Faster R-CNN [48] as a baseline in Degradation metric. As evident from Fig. 3b, our mPE metric effectively quantifies the impact of each perturbation and severity level, while other metrics are less sensitive.

3.4. Perturbation Robustness Benchmarks

To systematically benchmark the robustness of DLA models, we apply the mentioned 12 perturbations to 3 datasets. PubLayNet-P [72] is a commonly used large-scale DLA dataset. It contains over 360,000 document images sourced from PubMed Central, and focuses on 5 principal types of layout elements, i.e., text, title, list, table, and figure.

DocLayoutNet-P [46] includes academic papers, brochures, business letters, technical papers, and more with 80,863 document pages. It contains annotations for 11 layout components, aiming at enabling generalizable DLA models.

M6Doc-P [8] provides more nuanced annotations (74 classes), capturing finer aspects of document layouts with 9,080 document images. This dataset is tailored to advance research in sophisticated document understanding tasks.

While PubLayNet provides vast array of annotated scientific documents, DocLayoutNet broadens the variety of document types, and M6Doc delves into more detailed and complex annotations, catering to advanced document analysis.

4. Robust Document Layout Analyzer

4.1. Framework Overview

To boost the robustness performance on perturbed documents, we propose a Robust Document Layout Analyzer (RoDLA) model. As shown in Fig. 4, our proposed model is inspired by the architecture of DINO [69] but introduces channel attention blocks and average pooling blocks in the Encoder, allowing for more robust feature extraction. Additionally, we incorporate InternImage[61] as a backbone, which is pre-trained on ImageNet22K [10] dataset. This pretraining setting enhances multi-scale feature extraction for more stable performance. The overall architecture and robustness enhancement design is depicted in Fig. 4.

4.2. Robustness Enhancement Design

As identified in [19], self-attention tends to overemphasize focus on irrelevant tokens, that leads to attention shifting in...
visual document analysis, where globally accumulated perturbations may interfere with the model performance. To address this, we integrate spatial-wise average pooling layers with varying dilation into the Robust Encoder, accompanied by a Dilatation Predictor employing two $3 \times 3$ convolutional layers to predict the utilization of dilated average pooling layers. This approach leverages the attention of neighboring tokens to mitigate the overemphasis on tokens, and dynamically constrains long-distance perception based on context. By diminishing attention on irrelevant tokens, this approach effectively captures robust feature.

As outlined in [73], self-attention inherently facilitates visual grouping, filtering out irrelevant information. Thus, we have incorporated a Channel-wise Attention (CA) module within the Encoder. This module implements a self-attention operation in channel dimension $d$, as: $y = \sigma \left( \frac{1}{d} \sum_{s=1}^{N} \text{Softmax}(Q^T V) \right)$, where $Q, V \in \mathbb{R}^{d \times n}$. This mechanism not only simplifies computational complexity but also aggregates local features in channel-wise. The encoder module synergies channel-wise and spatial-wise self-attention, aggregating strongly correlated feature channels while directing attention towards spatially relevant key tokens. This enhancement of global attention minimizes the shifting of attention tokens, leading to robust feature extraction for complex analysis task.

5. Benchmarking and Analysis

5.1. Experiment Settings

**Selected Baselines.** To establish a comprehensive robustness benchmark for DLA, we not only include three distinct datasets but also reproduce a variety of state-of-the-art methods and backbones, including CNN-based backbones (ResNet [22], LayoutParser [54], InternImage [61]), Transformer-based structures (SwinDocSegmenter [4], DiT [31], and LayoutLMv3 [26]), and methods from CNN-based structures (Faster R-CNN [48], Mask R-CNN [23], Cascade R-CNN [6]), to Transformer-based frameworks (DocSegTr [5], DINO [69], Co-DINO [75], and our RoDLA). Note that all the methods are pre-trained on ImageNet dataset, while DiT and LayoutLMv3 are pre-trained on the IIT-CDIP Test Collection 1.0 document dataset [30], which gives a comparison of the pre-training data.

**Implementation Details.** For fair comparison, we reproduce all methods in the MM Detection [7] framework. Hyperparameters are from their original settings or papers. All models are trained and validated exclusively on clean data. Only during the testing phase, the impact of perturbations is evaluated on all methods. This approach ensures that our robustness benchmark is conducted under controlled and unbiased conditions. More details are in supplementary.

5.2. Robustness Evaluation Metrics

From our analysis in Fig. 2 and Table 1, it is evident that various types of perturbations impact the same document images differently, leading to a range of effects on the models’ performance. Previous robustness benchmarks [25, 33, 39, 66] measure the perturbations impact solely relying on a pre-selected baseline model. This approach tends to overlook the varying degrees of robustness different models exhibit towards the perturbations impact, overly relying on the benchmark performance of a particular model. To address this limitation, we propose **Mean Robustness Degradation (mRD)** for evaluating model robustness. mRD is a comprehensive measure of the average performance of models under different levels and types of perturbations. Additionally, inspired by ImageNet-C, we opt to incorporate Degradation $D$ into mRD because we found that different perturbations pose varying levels of difficulty. Thus, the RD of a perturbation $p$ is calculated as:

$$RD_p = \frac{1}{N} \sum_{s=1}^{N} \frac{D_{s,p}}{mPE_{s,p}},$$

where $mPE$ is from Eq. (3). The overall robustness score $mRD$ is the average score of $RD_p$ across perturbations $p \in [1, 12]$. Exceeding 100 indicates the model’s performance degrades more than expected due to perturbations, while falling below 100 suggests performance improvements despite these impacts. The lower the better. This metric offers a comprehensive assessment of model robustness, enabling evaluation under diverse and realistic conditions. To ensure a more comprehensive evaluation, we test DLA models on clean datasets with mAP scores, as well as the average mAP performance on 12 perturbations (i.e., $P$-Avg).
Table 2. The robustness benchmark on PubLayNet-P dataset. V, L, and T stand for Visual, Layout, and Textual modality. ‘Ext.’ means using extra pre-training data. mAP scores are evaluated on the clean data, the 12 perturbation types (P1–P12), and the perturbation average (P-Avg). mRD is the proposed mean Robustness Degradation, the smaller the better.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Method</th>
<th>Modality V</th>
<th>L</th>
<th>T</th>
<th>Clean P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>Content P6</th>
<th>P7</th>
<th>P8</th>
<th>Inconsistency P9</th>
<th>P10</th>
<th>P11</th>
<th>Noise P12</th>
<th>P-Avg</th>
<th>mRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet [63]</td>
<td>LayoutParser [54]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>89.0</td>
<td>35.8</td>
<td>78.1</td>
<td>68.0</td>
<td>84.8</td>
<td>45.9</td>
<td>79.0</td>
<td>85.8</td>
<td>82.6</td>
<td>51.4</td>
<td>31.1</td>
<td>53.6</td>
<td>0.03</td>
<td>58.0</td>
</tr>
<tr>
<td>ResNet [22]</td>
<td>Faster R-CNN [48]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>90.2</td>
<td>44.2</td>
<td>75.3</td>
<td>74.3</td>
<td>78.9</td>
<td>53.2</td>
<td>81.4</td>
<td>82.7</td>
<td>80.2</td>
<td>80.6</td>
<td>63.2</td>
<td>55.7</td>
<td>24.3</td>
<td>66.2</td>
</tr>
<tr>
<td>ResNet [22]</td>
<td>DocSegTr [5]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>90.4</td>
<td>28.3</td>
<td>76.0</td>
<td>71.2</td>
<td>85.2</td>
<td>46.2</td>
<td>69.3</td>
<td>86.2</td>
<td>86.9</td>
<td>50.8</td>
<td>19.8</td>
<td>60.2</td>
<td>0.04</td>
<td>55.2</td>
</tr>
<tr>
<td>ResNet [22]</td>
<td>Mask R-CNN [23]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>91.0</td>
<td>40.0</td>
<td>74.0</td>
<td>71.8</td>
<td>76.9</td>
<td>48.7</td>
<td>79.0</td>
<td>80.6</td>
<td>77.4</td>
<td>78.2</td>
<td>54.2</td>
<td>55.1</td>
<td>31.9</td>
<td>64.0</td>
</tr>
<tr>
<td>Swin [35]</td>
<td>SwinDocSegmenter [4]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>93.7</td>
<td>39.0</td>
<td>83.3</td>
<td>61.3</td>
<td>88.6</td>
<td>46.6</td>
<td>87.9</td>
<td>88.1</td>
<td>86.3</td>
<td>37.3</td>
<td>29.2</td>
<td>36.1</td>
<td>0.05</td>
<td>57.1</td>
</tr>
<tr>
<td>DiT [31]</td>
<td>Cascade R-CNN (6)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>94.5</td>
<td>31.9</td>
<td>86.5</td>
<td>82.2</td>
<td>92.1</td>
<td>79.6</td>
<td>87.2</td>
<td>92.0</td>
<td>91.6</td>
<td>93.8</td>
<td>71.3</td>
<td>69.9</td>
<td>41.8</td>
<td>76.7</td>
</tr>
<tr>
<td>LayoutMv3 (26)</td>
<td>Cascade R-CNN (6)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>95.0</td>
<td>32.7</td>
<td>85.8</td>
<td>79.8</td>
<td>92.3</td>
<td>68.5</td>
<td>86.5</td>
<td>89.3</td>
<td>86.7</td>
<td>82.9</td>
<td>47.0</td>
<td>82.1</td>
<td>45.3</td>
<td>73.6</td>
</tr>
</tbody>
</table>

Swin [35]     | Co-DINO [75]           | ✓          | ✓ | ✓ | 94.2    | 22.4 | 43.0 | 24.7 | 92.6 | 56.8 | 86.4 | 72.6 | 75.8 | 35.0 | 20.2 | 55.4 | 14.9 | 50.0 | 254.1 |
| InternImage [61] | Cascade R-CNN (6)   | ✓          | ✓ | ✓ | 94.1    | 27.7 | 80.2 | 79.7 | 89.6 | 56.6 | 83.8 | 91.6 | 86.9 | 84.9 | 54.0 | 55.2 | 0.02 | 66.1 | 141.9 |
| InternImage [61] | DINO [69]             | ✓          | ✓ | ✓ | 95.4    | 34.4 | 82.2 | 82.2 | 92.3 | 57.7 | 91.7 | 92.8 | 92.1 | 90.9 | 63.4 | 47.3 | 0.13 | 69.0 | 120.7 |
| InternImage [61] | Co-DINO [75]           | ✓          | ✓ | ✓ | 94.2    | 19.1 | 48.3 | 35.7 | 91.3 | 60.6 | 87.9 | 81.4 | 87.6 | 37.9 | 24.7 | 49.7 | 11.2 | 53.2 | 230.3 |
| InternImage [61] | RoDLA (Ours)          | ✓          | ✓ | ✓ | 96.0    | 31.6 | 79.3 | 80.6 | 92.9 | 61.6 | 91.6 | 92.6 | 91.6 | 91.3 | 67.7 | 58.8 | 0.03 | 70.0 | 116.0 |

5.3. Results on PubLayNet-P

Our analysis of robustness results on PubLayNet-P and test results on PubLayNet [72], as shown in Table 2, reveals a noteworthy aspect of our RoDLA model and robustness benchmark. Initially designed to enhance robustness, RoDLA surprisingly achieves state-of-the-art performance on the clean data with 96.0% in mAP. Even without additional pre-training on document-specific datasets, RoDLA maintained high performance under various perturbations, achieving 70.0% in P-Avg and 116.0 in mRD. Compared to the previous model Faster R-CNN [48], it realizes a +3.8% improvement in P-Avg. This suggests that a robustness-focused architecture can inherently boost performance in diverse conditions. We observe significant performance variability among models under different perturbations like blur and noise. For example, SwinDocSegmenter [4] drastically drops from 93.7% to 37.3% and 29.2% under blur in mAP, whereas RoDLA shows exceptional robustness, maintaining higher mAP of 91.3% and 67.7%. These significant results prove the effectiveness of RoDLA’s robust design. Moreover, our findings indicate that extra pre-training on IIT-CDIP Test Collection 1.0 document dataset of DiT [31] and LayoutMv3 [26] can better address unique document perturbations, e.g., texture and speckle. The multimodal LayoutMv3 [26] does not exhibit better robustness compared to the unimodal DiT [31].

5.4. Results on DocLayNet-P

The results in Table 3 on DocLayNet [46] and DocLayNet-P highlights a noteworthy insight. Training on DocLayNet, which includes electronic and printed document images, enables models to learn features that are nonexistent in electronic documents, e.g., PubLayNet. Our RoDLA achieves state-of-the-art performance on both clean (80.5% in mAP) and perturbed dataset (65.6% in P-Avg and 135.7 in mRD), which is 3.6% higher in mAP, 7.1% higher in P-Avg and 147.0 lower in mRD than previous state-of-the-art model [4]. Besides, the multi-modal LayoutLMv3 [26] shows notable disparities, i.e., 5.4% less than our RoDLA in clean mAP, 3.5% less in P-Avg, 37.1 higher in mRD.

5.5. Results on M0Doc-P

The results on M0Doc [8] are detailed in Table 4. RoDLA and DiT [31] exhibit notable performances. RoDLA achieves a balanced profile with 70.0% in mAP on clean data, the highest P-Avg of 61.7%, and the lowest mRD of 147.6. DiT [31], while scoring high in clean accuracy (70.2%) and P-Avg (60.6%), achieves 164.8 in mRD. Other models like Faster R-CNN [48] and LayoutMv3 [26] show moderate performance but lower performance in robustness. Overall, our RoDLA is 8% better in mAP, 12.1% better in P-Avg and 91.6 lower in mRD compared to other state-of-the-art models under the same pre-training.
5.6. Ablation Study

An ablation study on the RoDLA model was conducted to assess the impact of its components on performance in robustness benchmarks and on clean datasets.

**Effect of Backbone.** The comparison in Table 5 is based on Swin [35] and InternImage [61] as the backbone. With InternImage [61], the model provides an advance of 0.4% in mAP on clean dataset and a 9% gain in P-Avg, a 72.5 lower mRD compared to Swin [35], which reveals the exceptional performance in feature extraction by InternImage [61]. However, when employing our RoDLA method, the performance gap narrows: Swin [35] trails by only 0.4% in mAP on clean dataset, 0.1% in P-Avg, and 8.2 higher in mRD relative to InternImage [61] as the backbone. This suggests that InternImage [61] as a backbone outperforms Swin [35] in both performance and robustness under the same pre-training and method conditions. Our RoDLA method effectively harnesses robust features and reduces the performance gap.

**Effect of Method.** When using InternImage as the backbone and comparing different methods, e.g., Co-DINO [75], despite a slight improvement in clean dataset performance compared to Cascade R-CNN, shows a significant decrease in P-Avg by 12.9% and an increase in mRD by 88.3, indicating a drop in robustness. Our RoDLA method shows an improvement over the Cascade R-CNN method. It achieves a 1.9% gain in mAP, a 3.9% rise in P-Avg, and a substantial 25.9 reduction in mRD, indicating enhanced robustness.

**Analysis of Robustness Design.** To explore the robustness performance related to model structural designs, we conduct a detailed ablation study of RoDLA in Table 6. By integrating Channel Attention (CA) into the DINO [69] structure, we found that regardless of placement, adding CA can maintain 95.7% mAP with clean data. CA in different positions led to P-Avg decreases of 1.2%, 3.3%, and 2.5%, and mRD increases of 6.9, 18.6, and 13.4. These results suggest strategic placement of CA is crucial for balancing detail sensitivity and robustness of model. Besides, introducing Average Pooling Layers (APL) to mitigate irrelative attention on damaged tokens, we found the optimal way to be within the Encoder. This adjustment improved performance across all metrics, get benefits of 0.9% in P-Avg and 5.8 in mRD compare to intersecting in Decoder and 3.0% in P-Avg and 16.6 in mRD compare to placement in both. Besides, our robustness design has fewer parameters.

**Analysis of Multiple Perturbations.** To analyze multiple perturbations, we conduct an experiment based on the superposition of five perturbations. The results are illustrated in Table 7. As the number of perturbations increases, i.e., from simple to complex perturbations, the mPE score increases and the model performance (P-Avg) decreases. This trend indicates that the model robustness is more susceptible to degradation with more perturbations. Nonetheless, our RoDLA method can obtain robust performance.

### 6. Conclusion

In this work, we introduce the first robustness benchmark for Document Layout Analysis (DLA) models. Inspired by real-world document processing, we create a taxonomy with 12 common perturbations with 3 levels of severity in document images. The benchmark includes almost 450,000 documents from 3 datasets. To evaluate the impact of document perturbations, we propose two metrics, i.e., mean Perturbation Effect (mPE) and mean Robustness Degradation (mRD). We hope the benchmark and these metrics can enable the future work for robustness analysis of DLA models. Besides, we propose a novel model, Robust Document Layout Analyzer (RoDLA). Extensive experiments on three datasets prove the effectiveness of our methods. RoDLA can obtain state-of-the-art performance on the perturbed and the clean data. We hope our work will establish a solid benchmark for evaluating the robustness of DLA models and foster the advancement of document understanding.

**Limitations.** The proposed robustness benchmark is constrained on the vision-based single-modal DLA models. The benchmark can be extended to cover multi-modal DLA models. Besides, performing human-in-the-loop testing to evaluate the robustness of DLA models would be a crucial step before deploying in real-world applications.
References


[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. IJCV, 2021. 4


[23] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In ICCV, 2017. 2, 6, 7


[34] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. StructTextT: Structured Text Understanding with Multi-Modal Transformers. In ACMMM, 2021. 2


[54] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutoutparser: A unified toolkit for deep learning based document image analysis. In ICDAR, 2021. 1, 2, 6, 7


