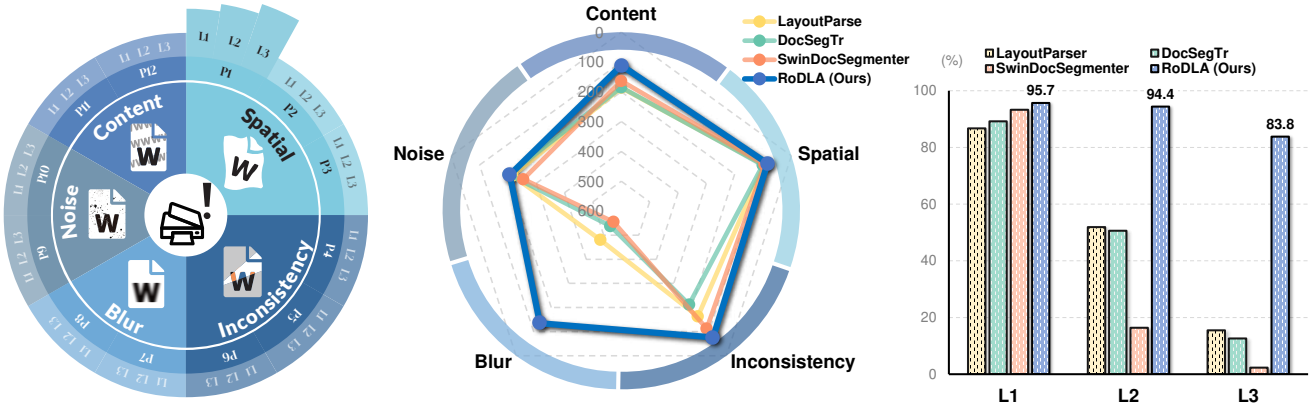
RoDLA: Benchmarking the Robustness of Document Layout Analysis Models

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/



(a) Taxonomy of document perturbations (b) mRD results in five perturbation groups. (c) mAP across severity levels, *e.g.*, *defocus*.

Figure 1. **Robust Document Layout Analysis (RoDLA)** with hierarchical perturbations. (a) For benchmarking, we propose 5 groups (*i.e.*, *spatial transformation*, *content interference*, *inconsistency distortion*, *blur*, and *noise*) and 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 method obtains (b) higher mean Robust Degradation (mRD) in all 5 groups of perturbations, and (c) stable mAP scores across 3 levels of perturbation (*e.g.*, *defocus*).

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 *common document perturbations with 3 severity levels inspired by real-world document processing. Additionally, to better understand document perturbation impacts, we propose two metrics, Mean Perturbation Effect (mPE) for perturbation assessment and Mean Robustness Degradation (mRD) for robustness evaluation. Furthermore, we introduce a self-titled model,* i.e., *Robust Document Layout Analyzer (RoDLA), which improves attention mechanisms to boost extraction of robust features. Experiments on the proposed benchmarks (PubLayNet-P, DocLayNet-P, and M*^{6}*Doc-P) demonstrate that RoDLA obtains state-of-the-art mRD scores of* 115.7, 135.4, *and* 150.4, *respectively. Compared to previous methods, RoDLA achieves notable improvements in mAP of* +3.8%, +7.1% *and* +12.1%, *respectively.*

*Corresponding author (e-mail: jiaming.zhang@kit.edu).

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

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 M^{6}Doc-P datasets, having mAP gains of $+7.1\%$ and $+12.1\%$, respectively.

To summarize, we present the following contributions:

- We are the first to benchmark the robustness of Document Layout Analysis (DLA) models. We benchmark over 10 single- and multi-modal DLA methods by utilizing almost $450{,}000$ documents.
- We introduce a comprehensive taxonomy with common document image perturbations, which includes 5 groups of high-level perturbations, comprising 12 distinct types, each with 3 levels of severity.
- 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 M^{6}Doc-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 systemically evaluate the robustness of DLA models.

Given a DLA model $g{:}X{\rightarrow}L$, trained on a digital documents dataset $\mathcal{N}$, the model in traditional benchmark is tested by the probability $P_{(x,l)\sim\mathcal{N}}(g(x) = l)$. 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 $E_{o\sim O}[P_{(x,l)\sim\mathcal{N}}(g(o(x)){=}o(l))]$. 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.

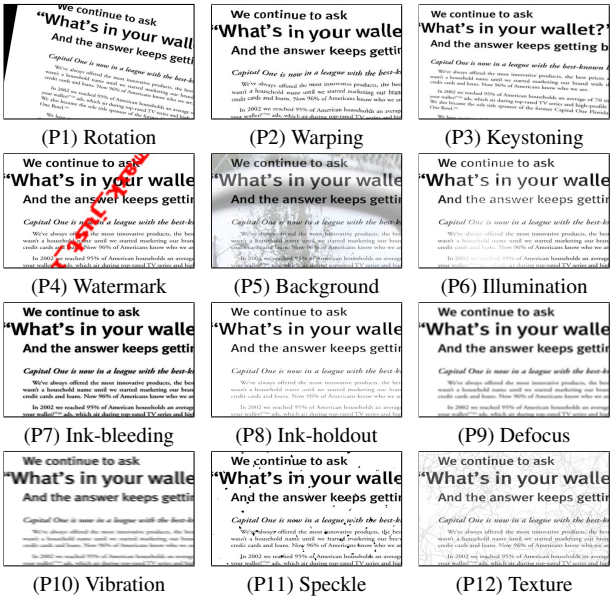


Figure 2. **Visualization of document perturbations.**

3.2. Perturbation Description

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

❶ **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 $\boldsymbol{\theta}$, which is randomly chosen from a uniform distribution within predefined ranges. (P2) **Warping**, which is enacted by generating two deformation fields, $(\Delta x, \Delta y){=}\alpha{\cdot}\mathcal{G}(U_{(x,y)}, \sigma)$. $U_{(x,y)}$ obeys the uniform distribution, $\mathcal{G}$ denotes Gaussian filter, σ and α are the smoothness and magnitude controlled by scaling factor $\boldsymbol{R_\sigma}$ and $\boldsymbol{R_\alpha}$, 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 $\boldsymbol{R_k}$. Annotation transformations are adjusted accordingly to alignment.

❷ **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

Table 1. **Document perturbations** on PubLayNet-P, DocLayNet-P, and M^{6}Doc-P datasets. mPE: mean Perturbation Effect.

Document Perturbation (P)		Description	mPE↓
None		The original clean data	0
❶ Spatial	**(P1) Rotation**	Rotation simulation with θ=[5°, 10°, 15°]	58.30
	(P2) Warping	Elastic transformation simulation with R_σ=[0.2, 0.06, 0.04] and R_α=[2, 0.6, 0.4]	22.00
	(P3) Keystoning	Perspective transformation simulation with R_k=[0.02, 0.06, 0.1]	34.49
❷ Content	**(P4) Watermark**	Text overlay with α_w=[51, 153, 255] and R_z=[2, 4, 6]	09.05
	(P5) Background	Random image inserting in background with N_i=[1, 3, 5]	26.70
❸ Inconsistency	**(P6) Illumination**	Region illumination change with V_l=[51, 102, 153] or V_s=[0.5, 0.25, 0.17]	10.67
	(P7) Ink-bleeding	Erosion simulation with kernel size K_e=[3, 7, 11]	08.91
	(P8) Ink-holdout	Dilation simulation with kernel size K_d=[3, 7, 11]	15.64
❹ Blur	**(P9) Defocus**	Gaussian blur simulation with kernel size K_g=[1, 3, 5]	08.10
	(P10) Vibration	Motion blur simulation with kernel size K_m=[3, 9, 15]	16.29
❺ Noise	**(P11) Speckle**	Blotches noise simulation with D_b=[1×10^{-4},3×10^{-4},5×10^{-4}]	24.31
	(P12) Texture	Fibrous noise simulation with N_f=[300, 900, 1500]	40.57
Overall		The average of all perturbations	22.90

$\boldsymbol{R_z}$ and alpha channel opacity $\boldsymbol{\alpha_w}$. (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 $\boldsymbol{N_i}$.

❸ **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 $\boldsymbol{V_l}$ or shadow $\boldsymbol{V_s}$, (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 $\boldsymbol{K_e}$ and $\boldsymbol{K_d}$, respectively.

❹ **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 $\boldsymbol{K_g}$, and (P10) **Vibration**, which is achieved by using a convolution kernel rotated by a random angle to replicate directional motion in size $\boldsymbol{K_m}$. This perturbation can be applied via a Gaussian filtering operation.

❺ **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 $\boldsymbol{D_b}$. 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 $\boldsymbol{N_f}$ 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

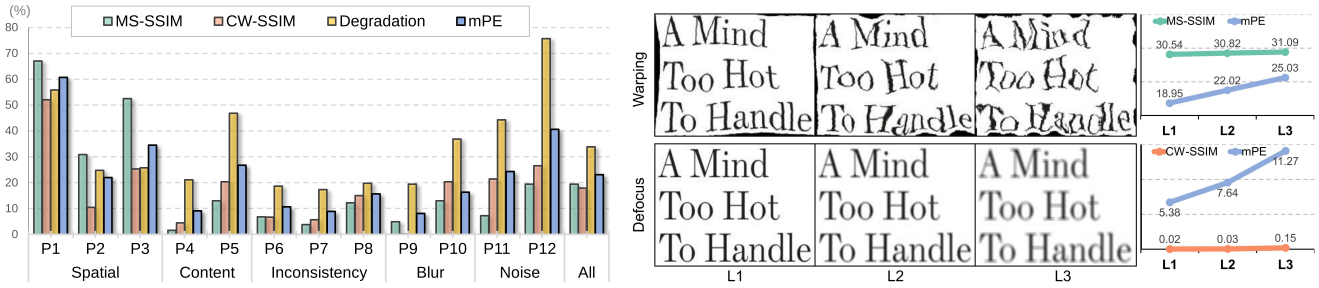
Quantifying the document perturbations is a crucial task, which lacks straightforward metrics. Previous methods [25, 28, 33, 66] are constrained by using a pre-selected baseline model performance as a reference to measure the perturbation effect, which will conflate the perturbation effect with the model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

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

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.



(a) Comparison of four evaluation metrics on 12 perturbations

(b) 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:

$$f^{\text{CW-SSIM}}(x,y) = \frac{2|\sum_{l=1}^{L} w_l x_l y_l^*| + K}{\sum_{l=1}^{L} |x_l|^2 + \sum_{l=1}^{L} |y_l|^2 + K}, \quad (2)$$

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 tiny value, and $*$ denotes the complex conjugation. As shown in Fig. 3a, it reflects the perturbations impact but fails for quantitive 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 $D^g_{s,p}$, and IQA metrics is written as $f^i_{s,p}$, *i.e.*, from Eq. (1) and Eq. (2). The **mPE** to measure the effect of perturbation p is calculated as:

$$\textbf{mPE}_p = \frac{1}{NMK} \sum_{s=1}^{N} \left(\sum_{i=1}^{M} f^i_{s,p} + \sum_{g=1}^{K} D^g_{s,p} \right). \quad (3)$$

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*.

DocLayNet-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.

M^{6}Doc-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, DocLayNet 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

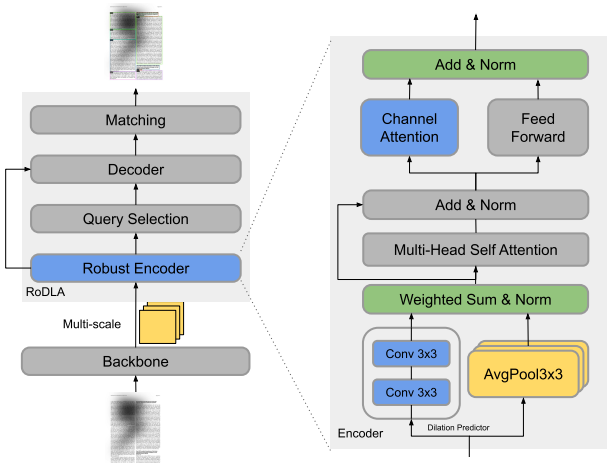


Figure 4. **The architecture of RoDLA model.** RoDLA is comprised of Encoder, Query Selection, Decoder, and Matching components. It optimizes the attention mechanism in Encoder, heightening focus on crucial tokens and reinforcing key token connections in multi-scale features to extract stable features.

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 *Dilation Predictor* employing two 3×3 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: $\mathbf{Y}=\sigma(\frac{\texttt{Softmax}(\mathbf{Q})\cdot\texttt{Softmax}(\mathbf{K}^T)}{\sqrt{n}}\cdot\texttt{MLP}(\mathbf{V}))$, where $\mathbf{Q},\mathbf{K},\mathbf{V}\in\mathbb{R}^{d\times n}$. 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 MMDetection [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:

$$\mathbf{RD}_p = \frac{1}{N}\sum_{s=1}^{N}\frac{D^g_{s,p}}{\text{mPE}_{s,p}}, \tag{4}$$

where mPE is from Eq. (3). The overall robustness score **mRD** is the average score of $\mathbf{RD}_p$ 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.

Backbone	Method	Modality V	L	T	Ext.	Clean	Spatial P1	P2	P3	Content P4	P5	Inconsistency P6	P7	P8	Blur P9	P10	Noise P11	P12	P-Avg↑	mRD↓
ResNeXt [63]	LayoutParser [54]	✓	✗	✗	✗	89.0	35.8	78.1	68.0	84.8	45.9	79.0	85.8	82.6	51.4	31.1	53.6	00.3	58.0	212.7
ResNet [22]	Faster R-CNN [48]	✓	✗	✗	✗	90.2	**44.2**	75.3	74.3	78.9	53.2	81.4	82.7	80.2	80.6	63.2	55.7	24.3	66.2	175.5
ResNet [22]	DocSegTr [5]	✓	✗	✗	✗	90.4	28.3	76.0	71.2	85.2	46.2	69.3	86.2	68.9	50.8	19.8	60.2	00.4	55.2	233.0
ResNet [22]	Mask R-CNN [23]	✓	✗	✗	✗	91.0	40.0	74.0	71.8	76.9	48.7	79.0	80.6	77.4	78.2	54.2	55.1	**31.9**	64.0	192.7
Swin [35]	SwinDocSegmenter [4]	✓	✗	✗	✗	93.7	39.0	83.3	61.3	88.9	46.6	87.9	88.1	86.6	37.3	29.2	36.1	00.5	57.1	214.4
DiT [31]	Cascade R-CNN [6]	✓	✗	✗	✓	94.5	31.9	86.5	82.2	92.1	79.6	87.2	92.0	91.6	93.8	71.3	69.9	41.8	76.7	95.8
LayoutLMv3 [26]	Cascade R-CNN [6]	✓	✓	✓	✓	95.1	32.7	85.9	79.8	92.3	68.5	86.5	93.1	86.7	82.9	47.0	82.1	45.1	73.6	116.2
Swin [35]	Co-DINO [75]	✓	✗	✗	✗	94.3	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	89.6	84.9	54.0	55.2	00.2	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	01.3	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	27.9	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**	00.3	**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 LayoutLMv3 [26] can better address unique document perturbations, *e.g.*, *texture* and *speckle*. The multimodal LayoutLMv3 [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.

Table 3. **The robustness benchmark on DocLayNet-P dataset**.

Backbone	Method	Modality V	L	T	Clean	P-Avg↑	mRD↓
ResNet [22]	DocSegTr [5]	✓	✗	✗	69.3	47.5	234.7
ResNet [22]	Mask R-CNN [23]	✓	✗	✗	73.5	52.7	195.7
ResNet [22]	Faster R-CNN [48]	✓	✗	✗	73.4	53.9	189.1
Swin [35]	SwinDocSegmenter [4]	✓	✗	✗	76.9	58.5	282.7
DiT [31]	Cascade R-CNN [6]	✓	✗	✗	62.1	52.1	216.5
LayoutLMv3 [26]	Cascade R-CNN [6]	✓	✓	✓	75.1	62.1	172.8
InternImage [61]	RoDLA (Ours)	✓	✗	✗	**80.5**	**65.6**	**135.7**

Table 4. **The robustness benchmark on M^{6}Doc-P dataset**.

Backbone	Method	Modality V	L	T	Clean	P-Avg↑	mRD↓
ResNet [22]	DocSegTr [5]	✓	✗	✗	60.3	43.2	212.6
ResNet [22]	Mask R-CNN [23]	✓	✗	✗	61.9	48.9	216.7
ResNet [22]	Faster R-CNN [48]	✓	✗	✗	62.0	49.6	192.2
Swin [35]	SwinDocSegmenter [4]	✓	✗	✗	47.1	39.7	239.2
DiT [31]	Cascade R-CNN [6]	✓	✗	✗	70.2	60.6	164.8
LayoutLMv3 [26]	Cascade R-CNN [6]	✓	✓	✓	64.3	57.0	176.1
InternImage [61]	RoDLA (Ours)	✓	✗	✗	**70.0**	**61.7**	**147.6**

5.5. Results on M^{6}Doc-P

The results on M^{6}Doc [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 LayoutLMv3 [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 compare to other state-of-the-art models under the same pre-training.

Table 5. **Ablation study** about robustness design for layout analysis models on PubLayNet and PubLayNet-P.

Backbone	Method	Clean	P-Avg↑	mRD↓
Swin [35]	Cascade R-CNN	93.7	57.1	214.4
	RoDLA (Ours)	95.6	69.9	124.2
InternImage [61]	Cascade R-CNN	94.1	66.1	141.9
	Co-DINO	94.2	53.2	230.3
	DINO	95.4	69.1	120.7
	RoDLA (Ours)	**96.0**	**70.0**	**116.0**

Table 6. **Analysis of robustness design** for RoDLA on PubLayNet and PubLayNet-P. All methods use InternImage. CA: Channel Attention [73]. APL: Average Pooling Layer [19].

CA Encoder	CA Decoder	APL Encoder	APL Decoder	#Params	Clean	P-Avg↑	mRD↓
				335.3M	95.4	69.0	120.4
✓				335.7M	95.7	67.8	127.3
	✓			335.7M	95.7	65.7	139.0
✓	✓			336.1M	95.7	66.5	133.8
		✓		320.0M	96.0	69.1	120.1
			✓	335.4M	96.0	64.2	127.4
		✓	✓	320.0M	96.0	67.8	126.2
✓		✓		323.2M	96.0	**70.0**	**115.7**
	✓		✓	335.9M	**96.1**	69.1	121.5
✓	✓	✓	✓	323.8M	95.9	67.0	132.3

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.

Table 7. **Analysis of Multiple Perturbations** on M^{6}Doc-P.

Perturbations	mPE↓	RoDLA P-Avg↑
Warping + Watermark	24.1	64.9
Warping + Watermark + Illumination	29.0	62.0
Warping + Watermark + Illumination + Defocus	32.3	50.8
Warping + Watermark + Illumination + Defocus + Speckle	46.2	41.9

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

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-End Transformer for Document Understanding. In *ICCV*, 2021. 2

[2] Diego Martin Arroyo, Janis Postels, and Federico Tombari. Variational transformer networks for layout generation. In *CVPR*, 2021. 2

[3] Christoph Auer, Ahmed Samy Nassar, Maksym Lysak, Michele Dolfi, Nikolaos Livathinos, and Peter W. J. Staar. Icdar 2023 competition on robust layout segmentation in corporate documents. *ArXiv*, 2023. 2

[4] Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. SwinDocSegmenter: An End-to-End Unified Domain Adaptive Transformer for Document Instance Segmentation. *ICDAR*, 2023. 1, 2, 6, 7

[5] Sanket Biswas, Ayan Banerjee, Josep Lladós, and Umapada Pal. DocSegTr: An Instance-Level End-to-End Document Image Segmentation Transformer. *arXiv preprint arXiv:2201.11438*, 2022. 1, 2, 6, 7

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 6, 7

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[8] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis. In *CVPR*, 2023. 2, 5, 7

[9] Denis Coquenet, Clément Chatelain, and Thierry Paquet. Dan: a segmentation-free document attention network for handwritten document recognition. *TPAMI*, 2023. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[11] Markus Diem, Florian Kleber, and Robert Sablatnig. Text classification and document layout analysis of paper fragments. In *ICDAR*, 2011. 2

[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

[13] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, 2020. 3

[14] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *CVPR*, 2022. 3

[15] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. Deep Unrestricted Document Image Rectification. *arXiv preprint arXiv:2304.08796*, 2023. 2

[16] Timothée Fronteau, Arnaud Paran, and Aymen Shabou. Evaluating Adversarial Robustness on Document Image Classification. *arXiv preprint arXiv:2304.12486*, 2023. 2

[17] Angelika Garz, Markus Diem, and Robert Sablatnig. Detecting text areas and decorative elements in ancient manuscripts. In *ICFHR*, 2010. 2

[18] Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. Doc2graph: a task agnostic document understanding framework based on graph neural networks. In *ECCV*, 2022. 2

[19] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *ICCV*, 2023. 3, 5, 8

[20] Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. Lidar snowfall simulation for robust 3d object detection. In *CVPR*, 2022. 3

[21] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In *ICDAR*. IEEE, 2015. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 6, 7

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

[24] Thomas Hegghammer. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5(1):861–882, 2022. 1, 2, 3

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2, 4, 5, 6

[26] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACMMM*, 2022. 2, 6, 7

[27] X. Jiang, R. Long, N. Xue, Z. Yang, C. Yao, and G. Xia. Revisiting document image dewarping by grid regularization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4533–4542, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2, 3

[28] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5

[29] Praveen Krishnan and C. V. Jawahar. Hwnet v2: an efficient word image representation for handwritten documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(4):387–405, 2019. 2

[30] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 665–666, New York, NY, USA, 2006. Association for Computing Machinery. 6

[31] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. DiT: Self-supervised Pre-training for Document Image Transformer. *arXiv preprint arXiv:2203.02378*, 2022. 2, 6, 7

[32] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020. 2, 3

[33] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20371–20381, 2023. 2, 4, 5, 6

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

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 7, 8

[36] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*, 2022. 2

[37] Chuwei Luo, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. Bi-vldoc: Bidirectional vision-language modeling for visually-rich document understanding. *ArXiv*, 2022. 2

[38] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. In *CVPR*, 2023. 2

[39] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3, 6

[40] S Milyaev and I Laptev. Towards reliable object detection in noisy images. *TPAMI*, 2017. 3

[41] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. In *ECCV*, 2022. 3

[42] Omar Moured, Jiaming Zhang, Alina Roitberg, Thorsten Schwarz, and Rainer Stiefelhagen. Line graphics digitization: A step towards full automation. In *ICDAR*, pages 438–453. Springer, 2023. 2

[43] Zoe Papakipos and Joanna Bitton. Augly: Data augmentations for robustness. *arXiv preprint arXiv:2201.06494*, 2022. 3

[44] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang. Effects of image degradation and degradation removal to cnn-based image classification. *TPAMI*, 2021. 3

[45] Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*, 2022. 2

[46] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *SIGKDD*, 2022. 3, 5, 7

[47] Yongri Piao, Wei Wu, Miao Zhang, Yongyao Jiang, and Huchuan Lu. Noise-sensitive adversarial learning for weakly supervised salient object detection. *TMM*, 2022. 3

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI*, 2017. 2, 5, 6, 7

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[50] Saifullah, Shoaib Ahmed Siddiqui, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. Are deep models robust against real distortions? a case study on document image classification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1628–1635, 2022. 1, 2, 3

[51] Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11):2385–2401, 2009. 4, 5

[52] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, and Qi Tian. Noise-aware fully webly supervised object detection. In *CVPR*, 2020. 3

[53] Zejiang Shen, Kaixuan Zhang, and Melissa Dell. A large dataset of historical japanese documents with complex layouts. In *CVPR*, 2020. 2

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

[55] Md Istiak Hossain Shihab, Md Rakibul Hasan, Mahfuzur Rahman Emon, Syed Mobassir Hossen, Md Nazmuddoha Ansary, Intesur Ahmed, Fazle Rabbi Rakib, Shahriar Elahi Dhruvo, Souhardya Saha Dip, Akib Hasan Pavel, et al. Badlad: A large multi-domain bengali document layout analysis dataset. In *ICDAR*, 2023. 2

[56] Alaa Sulaiman, Khairuddin Omar, and Mohammad F. Nasrudin. Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *Journal of Imaging*, 5(4), 2019. 1, 2

[57] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *CVPR*, 2023. 2

[58] Tuan Anh Tran, Kang Han Oh, In Seop Na, Gueesang Lee, Hyung-Jeong Yang, and Soohyung Kim. A robust system for document layout analysis using multilevel homogeneity structure. *Expert Syst. Appl.*, 2017. 2

[59] Jiapeng Wang, Lianwen Jin, and Kai Ding. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *ACL*, 2022. 2

[60] Lucai Wang, Hongda Qin, Xuanyu Zhou, Xiao Lu, and Fengting Zhang. R-YOLO: A Robust Object Detector in Adverse Weather. *TIM*, 2022. 3

[61] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, 2023. 5, 6, 7, 8

[62] Z. Wang, Eero Simoncelli, and Alan Bovik. Multiscale structural similarity for image quality assessment. pages 1398 – 1402 Vol.2, 2003. 4

[63] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. 7

[64] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*, 2020. 2

[65] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL*, 2021. 2

[66] Xu Yan, Chaoda Zheng, Zhen Li, Shuguang Cui, and Dengxin Dai. Benchmarking the robustness of lidar semantic segmentation models. *arXiv preprint arXiv:2301.00970*, 2023. 2, 4, 5, 6

[67] Huichen Yang and William Hsu. Transformer-based approach for document layout understanding. In *ICIP*, 2022. 2

[68] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *CVPR*, 2017. 2

[69] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 5, 6, 7, 8

[70] Mingliang Zhang, Zhen Cao, Juntao Liu, Liqiang Niu, Fandong Meng, and Jie Zhou. Welayout: Wechat layout analysis system for the icdar 2023 competition on robust layout segmentation in corporate documents. *ArXiv*, 2023. 2

[71] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. VSR: a unified framework for document layout analysis combining vision, semantics and relations. In *ICDAR*, 2021. 2

[72] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: largest dataset ever for document layout analysis. In *ICDAR*, 2019. 2, 3, 5, 7

[73] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding The Robustness in Vision Transformers. In *ICML*, 2022. 3, 6, 8

[74] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *ACMMM*, 2022. 2

[75] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training. In *ICCV*, 2023. 6, 7, 8