RoDLA: Benchmarking the Robustness of Document Layout Analysis Models (Supplementary Materials)

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

1. Implementation Details

Hardware Setup. In this work, we have trained all models (including reproduced models) on machines equipped with four A100, each having 40 GB of memory. Each node would also with 300 GB CPU memory.

Training Settings. After input batch normalization, we have applied flip with a 0.5 filp ratio and crop with a crop size $(384, 600)$ as data augmentation method. For a fair comparison in robustness benchmark for DLA models, we have trained all models (including reproduced models) using the same training strategy as in Table 1.

Table 1. Training settings.

Configurations	Parameter
Optimizer	AdamW
Learning Rate	$2e^{-4}$
Weight Decay	$1e^{-4}$
Scheduler	step-base
Training Epochs	24
Warm-up Step	{16, 22}
Warm-up Ratio	$1e^{-3}$
Batch-size per GPU	2

To create the benchmark, we have re-trained 38 models for this robustness benchmark for DLA models: On PubLayNet [17] dataset, we have re-trained 24 models (including ablation study). On DocLayNet [11] and M^{6}Doc [4] datasets, we have re-trained 7 models each, as we have only re-trained the models with representative performance, *i.e.*, high **mRD** or **mAP** for specific perturbation, on the robustness benchmark for PubLayNet [17] dataset.

2. Detail of Perturbation Taxonomy

In this section, we provide more details about our 12 document image perturbations in 3 severity levels.

(P1) Rotation. We apply a random rotation to document images, along with corresponding annotations. The rotation operation on an image of a document is an affine transformation, mathematically described by a rotation matrix. If θ is the angle of rotation, the transformation for rotating a point (x,y) around the origin is given by:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{1}$$

Here, (x', y') are the coordinates of the point after rotation. For L1, the θ is selected randomly from the range $[-5, 5]$. For L2, the θ is chosen randomly from $[-10, -5]$ or $[5, 10]$, each with 50% probability. For L3, the θ is taken randomly from $[-15, -10]$ or $[10, 15]$ for simulating real-world scenarios where object orientations vary.

(P2) Warping. We apply a pixel-wise displacement defined by a displacement field D. This field is typically generated using the Gaussian smoothing of random noise to simulate elastic deformation on document paper. The warping operation is as follows:

$$D(x, y) = \alpha \cdot G_\sigma(R(x, y)), \tag{2}$$

$$\begin{cases} x' = x + D_x(x, y) \\ y' = y + D_y(x, y), \end{cases} \tag{3}$$

where $R(x,y)$ is a random field for displacement in both the x and y directions. G_σ is a Gaussian function with standard deviation σ; the intensity or amplitude of the displacement is controlled by a factor α. D_x and D_y are the x and y components of the displacement field D.

(P3) Keystoning. We apply a 3D transformation to a 2D plane through a 3×3 matrix H, preserving lines but not necessarily the actual angles or lengths. This operation maps the homogeneous coordinates of a point in the source image to its new coordinates in the destination image:

$$\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = H \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \tag{4}$$

Then the actual position in the transformed image is given by normalizing with w' by $(\frac{x'}{w'}, \frac{y'}{w'})$. The elements of H are typically derived from corner-point correspondences between the source and destination images. The coordinates of destination images are selected randomly from a Gaussian distribution centered around the original coordinates.

The standard deviation of the Gaussian distribution is determined by the level.

(P4) Watermark. The process of adding a watermark involves several steps, primarily dealing with image composition and potential rotation. The watermark image is rotated by a random angle θ from range $[0°, 360°]$ before the blending process. Then the watermark W is blended onto the original image O using a technique called alpha blending. The resulting pixel value I is calculated as:

$$I = \alpha_w \cdot W + (1 - \alpha_w) \cdot O, \tag{5}$$

where α_w is the transparency level of the watermark, which allows the original image to show through to varying severity levels.

(P5) Background. For complex background simulation, we overlay multiple images onto the original image. Before background alpha composition, multiple background images are resized and placed on a copy image B of the original image A. The placement is defined by the position (x_{pos}, y_{pos}), which is randomly generated. The alpha composition can be described as:

$$I = \alpha_A \cdot A + (1 - \alpha_A) \cdot \alpha_B \cdot B, \tag{6}$$

where α_A and α_B are the alpha values of the original image and the background image, respectively.

(P6) Illumination. We introduce non-uniform illumination into document images, simulating effects such as shadows or glare. Mask M is created with random polygons filled with black on a white canvas, which is then blurred using a Gaussian filter. The illumination adjustment can be described mathematically as a pixel-wise multiplication of the image I with mask M :

$$I'(x, y) = V \cdot I(x, y) \cdot M(x, y), \tag{7}$$

where V is the illumination scaling factor, determined by the severity levels and type of illumination adjustment, *i.e.*, shadow with V_s and glare with V_l.

(P7) Ink-Bleeding. We apply an erosion operation for ink-bleeding simulation with an elliptical structuring element. The kernel size K_e determines the extent of erosion, depending on severity levels. The basic mathematical formula for erosion $\ominus$ of an image A by a structuring element B is:

$$(A \ominus B)(x, y) = \min_{(b_x, b_y) \in B} \{A(x + b_x, y + b_y)\}. \tag{8}$$

To improve image quality during erosion, we upscale the image tenfold in both dimensions before applying the erosion. This is followed by downscaling to the original size, ensuring enhanced detail and quality in the final image.

(P8) Ink-Holdout. To simulate Ink-Holdout, which is the opposite of Ink-Bleeding, we use the dilation operation, the inverse of erosion. The parameters for the dilation process, including the kernel size and the number of iterations, remain the same as those used for the erosion operation to maintain consistency in simulating these opposing ink behaviors. The mathematical formula for dilation $\oplus$ of an document image A by a elliptical structuring element B is:

$$(A \oplus B)(x, y) = \max_{(b_x, b_y) \in B} \{A(x - b_x, y - b_y)\}. \tag{9}$$

(P9) Defocus. The simulation of defocus blur is inherently complex due to the variability of point spread functions (PSFs) within diverse photographic conditions. Nevertheless, given that document images are frequently captured at close quarters, it is feasible to approximate the PSF with a Gaussian kernel function for simulating defocus blur which demonstrated as follows:

$$I_{\text{defocus}}(x, y) = (I * G)(x, y), \tag{10}$$

with

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{11}$$

where parameters of Gaussian kernel G are correspond to the level of severity which calibrated to manipulate the scope and the depth of field of the blur.

(P10) Vibration. Document vibration is simulated by motion blur. The kernel for motion blur is a matrix with non-zero values along a line. This line simulates the path of motion. The kernel for a horizontal motion blur is:

$$K = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{12}$$

where n is the number of non-zero elements in the kernel. This kernel K is then rotated through a random angle θ within a predetermined range, which emulates the directional motion effect. Similar to defocus blur, the motion blur effect is applied using a convolution operation between the image and the rotated kernel:

$$B(x, y) = (I * K_{\text{rotated}})(x, y), \tag{13}$$

where I is the original image, K is the motion blur kernel, and B is the blurred image.

(P11) Speckle. Document speckle is generated by superimposing random light (background) and dark (foreground) blobs onto a document image. We generate random blobs based on density, size, and roughness through randomly placed points and Gaussian smoothing. These foreground and background blobs are combined with the original image I as:

$$I_{\text{modified}} = \min\left(\max\left(I_{\text{original}}, N_{\text{fg}}\right), 1 - N_{\text{bg}}\right), \tag{14}$$

where N_{fg}, N_{bg} represent the foreground and background blob noise intensity. In the mathematical simulation of speckle and blotch noise on document images, Gaussian distributed blob noise are generated within the image domain, modulated by a blob density factor D_b , which is parametrically governed by designated severity levels.

(P12) Texture. We have endeavored to replicate the texture interference patterns characteristic of document imagery. This approach aims to emulate texture interference by simulating the complex plant fiber structures historically present in archival documents. We have modeled the random walk of fiber paths as follows:

$$\text{FiberPath} = \left[\sum_{k=1}^{n} \cos(\theta_k) \cdot \delta, \sum_{k=1}^{n} \sin(\theta_k) \cdot \delta \right], \quad (15)$$

where θ_k are angles drawn from a Cauchy distribution, δ is the step length, and k is the step number. The final fibrous image is obtained by blending fibrous textures:

$$I' = (\text{M} \cdot I_{\text{ink}}) + ((1 - \text{M}) \cdot (1 - I_{\text{paper}})) \times 255, \quad (16)$$

where mask M determines the application of ink and paper textures to the original image. Within this simulation, the spatial distribution of the fibers predominantly conforms to a Gaussian distribution, thereby reflecting the randomness inherent in the physical composition of paper. To impart authenticity to the fiber noise and facilitate a more accurate representation of document wear and quality, we have modulated the fiber density across varying noise levels.

3. Evaluation Metrics

In this work, there are two types of evaluation metrics, including: (1) the metrics for quantifying the impacts of perturbations will be presented in Sec. 3.1, such as MS-SSIM, CW-SSIM, and our proposed mPE. (2) the metrics for assessing the robustness of models will be detailed in Sec. 3.2, such as mAP and our proposed mRD.

3.1. Details of Perturbation Evaluation Metrics

To elucidate the effects of different perturbations and compare perturbation evaluation metrics, we present detailed analyses in Fig. 1, showcasing the impact of various perturbation categories and levels on document images.

MS-SSIM & CW-SSIM. In our robustness benchmark, we utilize MS-SSIM (Multi-Scale Structural Similarity Index) and CW-SSIM (Complex Wavelet Structural Similarity Index) metrics, both widely recognized for assessing the similarity between two images and pertinent for evaluating the extent of information loss caused by such perturbations. These indices exhibit varying sensitivity to image perturbations, as in Fig. 1. However, in this study, we deviate from the conventional usage of MS-SSIM and CW-SSIM as mere similarity measures. Given that these metrics yield a value of 100 for identical images, we propose using their complements relative to 100 to represent the loss of information, *i.e.*, $100-f^{\text{MS-SSIM}}$ and $100-f^{\text{CW-SSIM}}$. This approach enables a nuanced assessment of the impact of perturbations on document images, thereby enhancing the evaluation of model robustness in handling document perturbations.

mPE. The Mean Perturbation Effect (mPE) metric integrates the effects of image quality degradation and model performance reduction under various perturbations in DLA. Our mPE metric reveals a consistent trend, with an escalation in values corresponding to increased severity, particularly evident in Keystoning and Texture perturbations, as shown in Fig. 1. It highlights the compounded effects of perturbations, underscoring the importance of robustness in document analysis models. While all metrics show heightened impact with more severe perturbations, mPE uniquely captures the overall impact, serving as a dependable measure of model robustness against document perturbations and offering a comprehensive view of model robustness.

3.2. Details of Robustness Evaluation Metrics

mAP. The mean Average Precision (mAP) is a crucial metric in object detection, assessing a model's performance across various classes. It is calculated as the mean of the Average Precision (AP) for each category, where AP is the area under the Precision-Recall curve.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \, . \quad (17)$$

Here, AP_i is the Average Precision for the i^{th} class. mAP is especially important in multi-class detection tasks with varying Intersection over Union (IoU) thresholds.

P-Avg. We introduce P-Avg (Perturbation Average), a novel metric based on the mAP framework, designed to evaluate a model's robustness in document layout recognition across various levels and types of perturbations. P-Avg extends mAP to quantify a model's ability to maintain recognition accuracy under diverse perturbation scenarios. Based on Eq. (17), the P-Avg can be mathematically expressed as:

$$\text{P-Avg} = \frac{1}{MN} \sum_{s=1}^{M} \sum_{p=1}^{N} \text{mAP}_{s,p} \, . \quad (18)$$

In this formula, s represents perturbation level, p represents perturbation categories, and $\text{mAP}_{s,p}$ is the mAP calculated for the s^{th} level of perturbation in the p^{th} category. This metric provides insights into the model's adaptability and consistency in recognizing document layouts despite the presence of diverse and challenging distortions.

mRD. The mathematical underpinning of mRD pivots on the interplay between degradation D and the Mean Perturbation Effect (mPE). The metric is designed to normalize

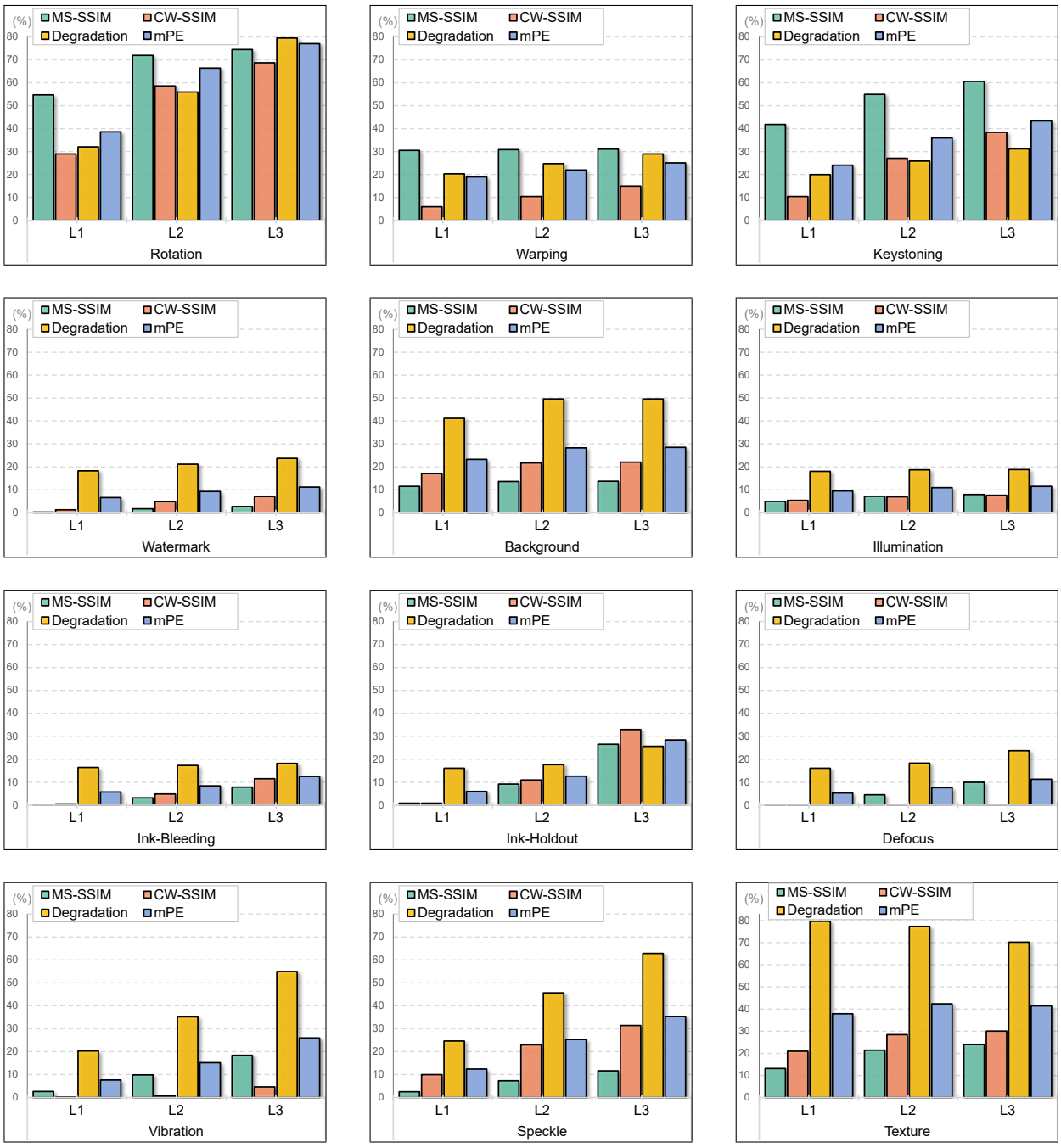


Figure 1. **Comparison between perturbation evaluation metrics** on 12 perturbation categories and 3 severity levels, including Image Quality Assessment methods (MS-SSIM and CW-SSIM), Degradation *w.r.t* a baseline, and the proposed mean Perturbation Effect (mPE). Other metrics cannot assess specific perturbations, for example MS-SSIM is insensitive to *warping* perturbation, and Degradation inversely measures *texture* perturbation across levels. In contrast, mPE is more balanced and inclusive to all perturbations and severity levels.

the degradation observed for a given perturbation by the perturbation's inherent difficulty as captured by mPE. This normalization is crucial as it accounts for the perturbation's baseline impact on the images, thus offering a relativized robustness measure. The degradation D represents how much a model's performance deviates from its unperturbed state when subjected to a specific perturbation and severity level.

Table 2. The robustness benchmark with the **best-case** result 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. **Best-case** scores are measured to show the upper-bound robustness performance that the DLA model could achieve at each perturbation in our robustness benchmark.

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 [15]	LayoutParser [13]	✓	✗	✗	✗	89.0	74.8	83.3	76.8	87.5	50.9	80.5	89.1	89.2	86.7	74.8	78.3	00.5	72.7	169.0
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	90.2	67.9	79.7	80.0	81.7	58.8	81.9	83.6	83.9	83.9	79.7	75.4	29.8	73.9	151.4
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	90.4	72.0	84.1	79.1	87.1	49.5	72.7	88.3	88.2	89.2	55.0	81.3	00.9	70.6	173.7
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	91.0	65.2	77.9	78.0	79.4	54.7	79.4	81.3	81.6	81.4	77.0	72.1	36.4	72.0	165.0
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	93.7	75.6	88.4	66.3	90.2	49.5	88.3	92.6	92.9	93.3	81.6	68.5	01.1	74.0	139.8
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	94.5	70.7	90.0	90.1	93.2	80.1	88.1	94.2	94.4	94.8	94.6	85.8	43.4	85.0	80.0
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	95.1	70.7	89.2	88.1	93.6	75.1	90.1	94.6	94.2	95.1	92.2	89.9	57.9	85.9	81.4
Swin [10]	Co-DINO [18]	✓	✗	✗	✗	94.3	57.0	71.5	29.0	92.9	61.3	87.2	92.5	92.1	93.6	54.5	71.8	22.0	68.8	153.5
InternImage [14]	Cascade R-CNN [3]	✓	✗	✗	✗	94.1	71.5	88.8	88.6	91.8	59.1	84.7	93.4	92.8	93.6	91.6	81.2	00.5	78.1	110.3
InternImage [14]	DINO [16]	✓	✗	✗	✗	95.4	74.6	89.5	89.2	93.9	59.8	92.0	94.4	94.2	95.1	93.5	80.4	01.6	79.9	94.4
InternImage [14]	Co-DINO [18]	✓	✗	✗	✗	94.2	49.8	73.0	43.3	92.5	63.6	88.5	91.9	92.2	93.5	72.4	75.7	13.8	70.9	153.5
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	96.0	71.9	89.0	88.4	94.2	64.4	92.0	94.6	94.6	95.7	93.3	81.6	00.5	80.0	93.8

Table 3. The robustness benchmark with the **worst-case** result on PubLayNet-P dataset. **Worst-case** scores are measured to show the lower-bound robustness performance that the DLA model could achieve at each perturbation in our robustness benchmark.

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 [15]	LayoutParser [13]	✓	✗	✗	✗	89.0	03.7	73.5	59.2	81.9	42.9	78.1	81.2	71.8	15.5	03.6	28.9	00.1	45.0	256.4
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	90.2	20.6	71.0	68.8	76.3	50.4	81.1	81.8	74.4	76.3	45.1	37.2	20.3	58.6	206.3
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	90.4	00.7	68.1	61.5	82.9	42.8	67.2	83.5	34.5	12.6	1.5	39.1	00.1	41.2	278.3
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	91.0	15.4	70.4	65.8	74.5	45.6	78.8	79.9	70.2	73.9	30.1	39.5	28.2	56.0	228.2
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	93.7	03.6	78.1	55.3	87.7	42.8	87.7	82.3	76.2	02.3	02.0	12.1	00.2	44.2	281.2
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	94.5	01.5	83.7	73.9	91.0	78.7	86.4	89.7	87.2	91.9	33.6	55.8	41.0	67.9	116.0
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	95.1	03.1	83.2	71.2	91.4	65.2	84.2	91.5	74.0	63.0	06.3	74.8	36.3	62.0	146.0
Swin [10]	Co-DINO [18]	✓	✗	✗	✗	94.3	00.9	17.5	20.2	92.3	53.9	86.0	42.3	45.7	01.4	02.0	45.2	10.5	34.8	342.8
InternImage [14]	Cascade R-CNN [3]	✓	✗	✗	✗	94.1	00.3	70.1	69.4	87.5	53.6	83.3	89.8	84.4	69.0	12.8	31.5	00.0	54.3	174.6
InternImage [14]	DINO [16]	✓	✗	✗	✗	95.4	02.2	74.3	74.9	90.5	54.2	91.6	91.5	88.4	83.8	25.8	22.0	00.9	58.3	147.0
InternImage [14]	Co-DINO [18]	✓	✗	✗	✗	94.2	00.4	26.8	28.1	90.1	58.6	87.5	66.2	80.7	02.3	03.2	30.8	09.9	40.4	301.5
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	96.0	03.2	57.8	72.3	88.0	48.2	89.6	89.6	87.3	72.5	19.8	16.2	00.2	53.7	138.7

Table 4. The detailed per-level **P-Avg**↑ results on PubLayNet-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Clean	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNeXt [15]	LayoutParser [13]	✓	✗	✗	✗	89.0	74.8	29.0	03.7	83.3	77.6	73.5	76.8	68.0	59.2	87.5	85.0	81.9	50.9	42.9	44.0	80.5	78.4	78.1	89.1	87.2	81.2	89.2	86.9	71.8	86.7	51.9	15.5	74.8	14.9	3.6	78.3	53.5	28.9	00.5	00.2	00.1
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	90.2	67.9	44.1	20.6	79.7	75.2	71.0	80.0	74.1	68.8	81.7	78.8	76.3	58.8	50.4	50.4	81.9	81.3	81.1	83.6	82.7	81.8	83.9	82.4	74.4	83.9	81.7	76.3	79.7	64.8	45.1	75.4	54.4	37.2	20.3	22.7	29.8
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	90.4	72.0	12.2	00.7	84.1	75.7	68.1	79.1	72.9	61.5	87.1	85.5	82.9	49.5	42.8	46.2	72.7	67.9	67.2	88.3	86.7	83.5	88.2	83.9	34.5	89.2	50.6	12.6	55.0	02.9	01.5	81.3	60.3	39.1	00.9	00.1	00.1
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	91.0	65.2	39.3	15.4	77.9	73.8	70.4	78.0	71.5	65.8	79.4	76.9	74.5	54.7	45.9	45.6	79.4	78.9	78.8	81.3	80.7	79.9	81.6	80.5	70.2	81.4	79.3	73.9	77.0	55.6	30.1	72.1	53.6	39.5	28.2	31.1	36.4
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	93.7	75.6	37.9	03.6	88.4	83.5	78.1	66.3	62.3	55.3	90.2	88.8	87.7	49.5	42.8	47.4	88.3	87.7	87.7	92.6	89.4	82.3	92.9	90.7	76.2	93.3	16.4	02.3	81.6	04.1	02.0	68.5	27.8	12.1	01.1	00.2	00.2
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	94.5	70.7	23.5	01.5	90.0	85.9	83.7	90.1	82.5	73.9	93.2	92.0	91.0	80.1	78.7	79.9	88.1	87.0	86.4	94.2	92.2	89.7	94.4	93.1	87.2	94.8	94.7	91.9	94.6	85.7	33.6	85.8	68.2	55.8	43.4	41.0	41.0
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	95.1	70.7	24.3	03.1	89.2	85.2	83.2	88.1	80.2	71.2	93.6	92.0	91.4	75.1	65.2	65.2	90.1	85.3	84.2	94.6	93.1	91.5	94.2	91.9	74.0	95.1	90.7	63.0	92.2	42.5	06.3	89.9	81.5	74.8	57.9	41.0	36.3
Swin [10]	Co-DINO [18]	✓	✗	✗	✗	94.3	57.0	09.3	00.9	71.5	40.0	17.5	29.0	24.8	20.2	92.9	92.7	92.3	53.9	55.3	61.3	87.2	86.1	86.0	92.5	83.0	42.3	92.1	89.5	45.7	93.6	09.9	01.4	54.5	04.1	02.0	71.8	49.1	45.2	22.0	12.2	10.5
InternImage [14]	Cascade R-CNN [3]	✓	✗	✗	✗	94.1	71.5	11.2	00.3	88.8	81.7	70.1	88.6	81.2	69.4	91.8	89.5	87.5	59.1	53.6	57.1	84.7	83.3	83.3	93.4	91.5	89.8	92.8	91.5	84.4	93.6	92.2	69.0	91.6	57.6	12.8	81.2	52.8	31.5	00.5	00.1	00.0
InternImage [14]	DINO [16]	✓	✗	✗	✗	95.4	74.6	26.5	02.2	89.5	82.7	74.3	89.2	82.6	74.9	93.9	92.4	90.5	59.8	54.2	59.2	92.0	91.6	91.6	94.4	92.4	91.5	94.2	93.6	88.4	95.1	93.7	83.8	93.5	70.9	25.8	80.4	39.5	22.0	01.3	00.9	01.6
InternImage [14]	Co-DINO [18]	✓	✗	✗	✗	94.2	49.8	07.2	00.4	73.0	45.0	26.8	43.3	35.8	28.1	92.5	91.3	90.1	59.7	58.6	63.6	88.5	87.8	87.5	91.9	86.0	66.2	92.2	89.9	80.7	93.5	17.8	02.3	72.4	08.0	03.2	75.7	42.6	30.8	13.8	09.9	10.0
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	96.0	71.9	19.9	02.9	89.0	80.4	68.5	88.4	81.2	72.1	94.2	93.0	91.6	64.4	58.4	61.9	92.0	91.4	91.5	94.6	92.4	90.8	94.6	92.9	87.3	95.7	94.4	83.8	93.3	71.2	38.5	81.6	51.2	43.6	00.5	00.2	00.1

Table 5. The detailed per-level **RD**↓ results on PubLayNet-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNeXt [15]	LayoutParser [13]	✓	✗	✗	✗	065.3	114.3	129.8	088.1	101.7	105.9	096.3	089.0	094.0	187.3	161.6	161.6	210.8	201.5	196.7	204.6	196.6	190.6	189.1	151.1	150.3	181.6	103.8	099.4	247.1	629.6	749.6	329.3	558.3	371.0	175.3	183.9	201.5	262.3	235.5	241.3
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	083.2	090.0	107.0	107.1	112.6	115.9	083.0	072.0	071.9	274.2	228.4	211.5	176.8	175.0	174.2	189.9	170.2	164.5	284.6	204.3	145.5	270.7	139.5	090.2	299.1	239.5	210.2	265.2	230.9	211.3	198.7	180.4	178.0	210.1	182.4	169.5
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	072.6	141.4	133.9	083.9	110.4	127.5	086.8	075.3	088.7	193.3	156.2	152.6	216.8	201.8	188.9	286.5	292.2	285.5	203.0	157.0	131.9	198.4	127.6	230.9	200.6	646.6	775.3	588.0	637.0	379.1	151.1	157.0	172.6	261.3	235.8	241.3
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	090.2	097.7	114.1	116.6	119.0	118.3	091.3	079.2	078.8	308.7	248.8	227.6	194.4	190.9	191.0	216.2	192.1	184.5	324.5	227.9	160.7	309.4	154.6	105.0	345.5	270.9	231.5	300.5	291.3	269.0	225.4	183.5	171.5	189.3	162.6	153.6
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	063.3	100.0	130.0	061.2	074.9	087.5	139.9	104.8	103.0	146.9	120.6	109.8	216.8	201.8	184.7	122.8	112.0	107.0	128.4	125.1	141.5	119.4	073.7	083.9	124.5	1094.2	866.6	240.4	629.1	377.2	254.4	285.6	249.1	260.7	235.5	241.0
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	076.0	121.2	152.4	052.2	064.0	065.1	111.1	048.3	060.1	101.9	080.2	080.2	063.4	073.2	070.6	114.9	116.1	118.4	100.6	092.1	082.5	094.2	054.7	043.1	091.6	069.4	071.4	101.6	091.5	253.7	114.7	125.8	152.1	140.2	79.2	142.9
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	076.0	121.9	130.6	057.0	067.2	067.1	049.4	055.0	066.4	095.9	086.2	076.8	106.9	122.8	122.2	103.9	133.8	137.5	093.7	081.5	067.9	097.5	064.2	091.6	091.0	121.7	328.2	101.9	377.2	360.6	081.6	073.2	071.4	111.0	139.2	153.8
Swin [10]	Co-DINO [18]	✓	✗	✗	✗	065.8	118.4	131.8	055.4	078.6	102.7	044.8	048.4	057.8	091.4	081.9	084.8	172.6	161.6	143.3	083.9	076.5	073.1	097.2	089.7	067.9	097.5	050.7	040.9	091.0	082.5	143.7	084.9	190.9	285.6	158.3	239.3	221.1	260.2	233.9	237.6
InternImage [14]	Cascade R-CNN [3]	✓	✗	✗	✗	073.9	143.0	134.4	059.1	083.1	119.5	047.3	052.3	070.5	122.9	113.1	111.6	175.6	163.7	150.6	160.5	152.0	145.3	114.5	100.4	081.5	121.1	067.4	055.0	118.9	102.1	275.0	109.8	278.2	335.6	151.9	186.7	194.1	262.3	235.8	241.5
InternImage [14]	DINO [16]	✓	✗	✗	✗	065.8	118.4	131.8	055.4	078.6	102.7	044.8	048.4	057.8	091.4	081.9	084.8	172.6	161.6	143.3	083.9	076.5	073.1	097.2	089.7	067.9	097.5	050.7	040.9	091.0	082.5	143.7	084.9	190.9	285.6	158.3	239.3	221.1	260.2	233.9	237.6
InternImage [14]	Co-DINO [18]	✓	✗	✗	✗	130.1	149.4	134.3	142.5	249.8	292.5	235.4	178.5	165.7	112.4	093.7	088.4	173.0	146.1	127.8	120.7	111.0	108.8	140.5	165.3	270.2	131.2	080.1	068.0	120.7	1075.9	866.6	360.6	603.5	372.5	196.3	227.1	196.1	227.3	212.7	217.4
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	072.8	129.0	130.9	058.0	089.0	125.9	048.2	052.3	064.3	086.9	075.4	075.0	152.8	146.8	133.8	083.9	078.3	074.0	093.7	089.7	073.5	090.8	056.3	044.8	079.9	073.3	143.7	087.5	188.9	236.7	148.6	193.0	159.8	262.3	235.5	241.3

4. More Results

4.1. Detailed Results on PubLayNet-P

To comprehensively evaluate methods on PubLayNet-P, we report the best and the worst results in Table 2 and Table 3 which according to the severity levels. The best-case metric in our robustness benchmark is calculated using the best outcomes under each perturbation according to result in severity level, while the worst-case uses the poorest outcomes which selected from result of severity levels. These metrics reveal the range of a model's robustness to the perturbations we have established, showing the potential fluctuation in robustness performance across different models. Besides, exhaustive results for the **P-Avg** metric are presented in Table 4, while the ones for the **RD** metric are detailed in Table 5.

Table 6. The robustness benchmark with the **best-case** result on DocLayNet-P dataset.

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-Avgt↑	mRD↓
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	73.4	37.6	65.2	62.1	70.2	65.8	70.6	72.1	72.5	72.3	69.4	48.3	35.5	61.8	178.1
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	69.3	40.3	60.5	62.0	69.9	65.3	67.6	46.3	56.9	61.7	26.5	49.7	35.1	53.5	190.7
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	73.5	44.2	65.2	62.6	70.4	65.3	71.1	72.6	72.7	72.9	69.3	44.7	35.2	62.2	176.1
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	76.9	48.6	67.4	68.6	69.9	67.8	75.1	75.2	76.7	75.2	70.6	47.6	39.1	65.2	161.7
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	62.1	35.2	58.5	58.3	61.1	52.5	60.2	61.4	62.3	63.0	63.0	52.3	56.3	57.0	170.9
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	75.1	56.7	75.1	65.8	75.2	67.5	75.2	69.1	75.5	62.6	39.6	72.3	73.7	67.4	151.8
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	80.5	49.6	72.6	73.2	79.0	74.2	80.3	80.7	81.6	82.0	81.8	58.9	59.7	72.8	117.9

Table 7. The robustness benchmark with the **worst-case** result on DocLayNet-P dataset.

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-Avgt↑	mRD↓
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	73.4	01.4	58.8	46.3	67.4	61.0	70.1	66.2	63.2	27.7	05.3	09.9	27.5	42.1	221.2
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	69.3	02.6	49.1	42.7	65.8	32.5	60.3	39.8	44.7	44.6	05.4	23.4	22.0	36.1	314.9
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	73.5	08.4	58.7	53.4	67.3	60.7	70.5	66.4	62.5	26.7	05.3	08.4	27.5	43.0	220.3
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	76.9	07.2	54.6	55.9	68.0	58.2	74.9	67.8	71.0	31.9	05.0	25.2	35.7	46.3	208.1
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	62.1	01.3	51.4	44.1	55.9	49.1	58.4	58.5	57.7	61.9	45.6	48.3	51.5	48.7	271.0
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	75.1	21.8	67.8	59.5	74.2	59.3	59.6	63.8	07.0	47.3	00.6	70.1	67.8	49.9	234.7
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	80.5	04.3	59.4	59.1	74.5	70.4	80.0	77.1	73.4	77.3	44.7	26.3	48.7	57.9	157.3

Table 8. The detailed per-level **P-Avg**↑ results on DocLayNet-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Clean	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	73.4	37.6	11.0	01.4	65.2	60.4	58.8	62.1	52.7	46.3	70.2	68.6	67.4	65.8	61.6	61.0	70.6	70.5	70.1	72.1	69.8	66.2	72.5	70.0	63.2	72.3	66.5	27.7	69.4	19.6	05.3	48.3	21.5	09.9	35.5	27.5	27.7
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	69.3	40.3	23.7	02.6	60.5	58.4	49.1	62.0	51.7	42.7	69.9	68.2	65.8	65.3	54.1	32.5	64.3	67.6	60.3	46.3	42.6	39.8	56.9	49.3	44.7	61.7	45.2	44.6	26.5	24.2	05.4	49.7	23.4	24.6	35.1	32.9	22.0
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	73.5	44.2	18.1	08.4	65.2	60.4	58.7	62.6	53.4	56.9	70.4	68.7	67.3	65.3	61.3	60.7	71.1	71.0	70.5	72.6	70.3	66.4	72.7	69.5	62.5	72.9	66.3	26.7	69.3	21.0	05.3	44.7	18.1	08.4	35.2	28.2	27.5
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	76.9	48.6	19.0	07.2	67.4	58.6	54.6	68.6	61.9	55.9	68.0	69.1	69.9	67.8	59.7	58.2	74.9	75.1	74.9	75.2	72.1	67.8	76.7	75.6	71.0	75.2	65.9	31.9	70.6	21.5	05.0	47.6	32.7	25.2	39.1	35.7	37.0
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	62.1	35.2	08.3	01.3	58.5	54.1	51.4	58.3	50.3	44.1	61.1	55.9	55.9	52.5	49.3	49.1	60.2	59.1	58.4	61.4	59.9	58.5	62.3	61.9	57.7	61.9	62.8	63.0	63.0	58.0	45.6	48.3	52.2	52.3	56.3	52.8	51.5
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	75.1	56.7	33.2	21.8	75.1	73.0	67.8	65.8	60.9	59.5	75.2	75.1	74.2	67.5	61.4	59.3	75.1	75.2	59.6	69.1	65.1	63.8	75.5	41.5	07.0	62.6	54.4	47.3	39.6	07.5	00.6	72.3	71.8	70.1	73.7	68.9	67.8
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	80.5	49.6	17.8	04.3	72.6	64.2	59.4	73.2	65.8	59.1	79.0	76.4	74.5	74.2	71.1	70.4	80.3	80.2	80.0	80.7	78.4	77.1	81.6	79.0	73.4	81.5	82.0	77.3	81.8	67.5	44.7	58.9	37.8	26.3	59.7	51.3	48.7

Table 9. The detailed per-level **RD**↓ results on DocLayNet-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	128.2	121.7	122.3	146.3	146.9	141.7	126.0	109.7	105.5	283.6	247.6	230.1	163.2	156.0	156.3	221.0	202.3	197.2	290.7	236.4	190.9	282.1	179.1	114.6	299.5	263.6	263.1	276.0	265.3	241.3	241.5	216.6	203.0	196.2	177.9	171.7
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	122.8	104.3	121.0	166.2	154.4	174.8	126.6	111.9	112.6	286.6	250.8	241.2	165.6	186.5	270.6	268.6	222.0	261.9	560.1	450.3	339.6	442.7	302.6	172.3	414.2	431.4	201.8	662.5	250.1	240.9	234.6	211.3	169.8	197.7	164.5	185.1
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	114.7	111.9	113.7	146.3	146.9	141.8	124.5	108.0	084.6	281.9	246.7	230.7	165.5	157.3	157.5	217.2	198.7	195.0	286.4	232.9	189.5	280.0	182.0	116.9	293.6	265.4	266.9	277.0	260.6	241.2	258.0	225.9	206.4	197.3	176.0	172.2
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	105.6	110.7	115.2	137.2	153.8	156.1	104.4	088.4	086.6	304.2	243.5	212.1	153.5	163.9	167.4	188.7	170.6	165.5	258.6	218.9	181.7	239.5	145.8	090.4	268.4	268.2	248.1	264.6	258.9	242.0	244.6	185.7	168.5	185.5	157.8	149.5
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	133.2	125.3	122.5	174.5	170.2	166.9	138.6	115.4	109.7	370.1	347.2	311.3	226.7	206.0	204.0	299.9	280.2	274.5	402.4	314.3	234.2	387.0	227.7	131.7	411.9	293.0	134.7	333.1	138.5	138.6	241.5	132.0	107.5	133.1	115.7	115.1
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	089.0	091.3	097.0	104.7	100.2	110.7	113.8	090.6	079.6	235.8	196.4	182.3	155.2	157.1	163.0	187.6	170.2	266.1	322.3	273.8	204.5	252.0	349.4	289.7	404.7	358.9	191.8	544.3	305.2	253.1	129.4	077.9	067.4	080.0	076.2	076.5
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	103.6	112.4	118.8	115.2	132.9	139.5	089.2	079.3	080.3	199.8	186.0	179.9	123.0	117.5	118.7	148.3	135.7	132.0	201.4	169.3	129.3	188.9	125.4	082.8	200.1	141.7	082.7	164.1	107.2	140.8	191.8	171.6	166.1	122.7	119.4	121.8

4.2. Detailed Results on DocLayNet-P

To comprehensively compare the performance of different methods across various datasets, we also present the best and worst outcomes on the DocLayNet-P in Table 6 and Table 7, respectively. Furthermore, comparisons between different models performance in **P-Avg** and **RD** at different severity levels are provided in Table 8 and Table 9.

4.3. Detailed Results on M^{6}Doc-P

We showcase the best and worst model performances on the M^{6}Doc-P dataset in Table 10 and Table 11. We hope the elaborate result from three datasets illustrates the DLA model's robustness using data. Detailed **P-Avg** and **RD** metrics for M^{6}Doc-P are in Table 12 and Table 13, respectively.

5. Comparison with image rectification

To fully demonstrate the significance of researching the robustness of DLA models, we will compare the direct prediction results using our RoDLA model with the results obtained by first rectifying the document images using a Document Image Rectification model, followed by applying the DLA model. For the Document Image Rectification task, we employ the state-of-the-art DocTr [5] model. However, it is important to note that, currently, the task of document image rectification only considers the recovery of four types of perturbations in our robustness benchmark: **Rotation**, **Warping**, **Keystoning**, and **Illumination**. Therefore, the comparison will be located in these four perturbations.

5.1. Pipeline Comparison

Fig. 2 visualizes the two different pipelines, with Fig. 2a showing the inference flow of our RoDLA model and

Table 10. The robustness benchmark with the **best-case** result on M^{6}Doc dataset.

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-Avgt↑	mRD↓
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	62.0	44.6	60.1	56.6	61.7	54.0	59.3	62.5	61.9	62.1	59.6	52.2	40.7	56.3	177.2
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	60.3	38.4	52.1	48.5	52.7	32.6	72.7	88.3	43.7	89.2	58.1	41.3	48.8	55.5	168.6
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	61.9	61.7	60.6	62.3	60.3	61.8	55.8	58.9	55.9	51.7	61.9	59.6	62.2	59.4	144.2
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	47.1	31.4	45.5	42.9	47.0	40.3	44.8	47.3	46.8	47.2	46.3	40.7	39.1	43.3	196.5
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	70.2	61.0	61.1	66.7	60.5	50.6	70.9	62.4	61.4	64.7	53.7	66.2	69.9	62.4	141.9
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	64.3	60.7	64.6	63.1	64.4	56.1	64.8	52.2	63.4	59.1	45.2	64.0	68.4	60.5	156.3
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	70.0	58.4	68.2	66.3	68.8	62.4	68.2	69.7	69.9	69.2	68.5	62.9	55.6	65.7	125.3

Table 11. The robustness benchmark with the **worst-case** result on M^{6}Doc dataset.

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-Avgt↑	mRD↓
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	62.0	06.2	57.8	48.8	59.3	50.0	56.6	54.3	53.7	34.0	09.0	19.5	29.7	39.9	223.1
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	60.3	07.0	32.2	44.0	33.9	24.5	67.2	45.8	39.2	12.6	07.5	19.3	42.8	31.3	260.0
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	61.9	17.8	53.0	57.2	39.4	29.8	35.0	50.2	08.7	44.2	07.0	55.7	59.4	38.1	298.7
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	47.1	07.9	40.1	35.2	45.8	37.5	43.2	46.1	42.9	40.0	20.7	30.6	35.7	35.5	295.6
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	70.2	51.0	56.9	60.2	52.6	49.6	69.0	60.1	39.8	58.6	12.3	64.8	68.9	53.6	203.7
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	64.3	46.4	54.4	60.6	62.0	48.6	55.7	50.3	29.6	51.6	06.2	62.5	60.1	49.0	217.7
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	70.0	23.7	64.0	60.9	67.5	58.5	67.5	67.5	68.3	63.6	34.4	47.7	51.7	56.3	178.0

Table 12. The detailed per-level **P-Avg**↑ results on M^{6}Doc-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Clean	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	62.0	44.6	24.4	06.2	60.1	58.4	57.8	56.6	50.5	48.8	61.7	60.1	59.3	54.0	50.0	52.1	59.3	58.1	56.6	62.5	60.6	54.3	61.9	60.0	53.7	62.1	56.9	34.0	59.6	20.6	09.0	52.2	32.1	19.5	40.7	31.5	29.7
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	60.3	38.4	23.4	07.0	52.1	37.3	32.2	48.5	47.8	44.0	52.7	34.3	33.9	30.3	32.6	24.5	72.7	67.9	67.2	88.3	86.7	45.8	42.7	43.7	39.2	89.2	50.6	12.6	58.1	14.6	07.5	41.3	30.4	19.3	48.8	46.6	42.8
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	61.9	31.4	17.8	61.7	60.5	53.0	60.6	59.3	57.2	62.3	60.3	53.7	39.4	32.8	29.8	61.8	55.8	35.0	51.8	50.2	51.7	58.9	20.8	08.7	55.9	51.7	49.1	44.2	24.7	07.0	61.9	59.6	58.6	55.7	62.2	59.8	59.4
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	47.1	31.4	16.3	07.9	45.5	42.3	40.1	42.9	37.3	35.2	47.0	46.2	45.8	40.3	38.7	37.5	44.7	44.8	43.2	47.3	46.9	46.1	46.8	45.6	42.9	47.2	45.2	40.0	46.3	33.9	20.7	40.7	34.4	30.6	39.1	35.7	37.0
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	70.2	61.0	54.8	51.0	61.1	60.6	56.9	66.7	62.5	60.2	60.2	60.5	52.6	50.6	50.1	49.6	70.4	70.9	69.0	62.4	61.6	60.1	61.4	54.5	39.8	64.7	62.6	58.6	53.7	34.9	12.3	64.8	66.2	64.9	69.9	69.1	68.9
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	64.3	60.7	53.2	46.4	64.6	62.8	54.4	63.1	61.3	60.6	64.4	63.6	62.0	56.1	50.8	48.6	74.8	62.7	55.7	52.2	50.3	51.8	63.4	48.8	29.6	59.1	54.7	51.6	45.2	22.6	06.2	64.0	62.8	62.5	68.4	64.4	60.1
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	70.0	58.4	40.9	23.7	68.2	66.0	64.0	66.3	63.4	60.9	68.8	68.7	67.5	62.4	60.2	58.5	67.5	68.2	67.8	69.7	67.8	67.5	68.8	69.9	68.3	69.2	68.8	63.6	68.5	54.2	34.4	62.9	53.6	47.7	55.6	51.7	52.3

Table 13. The detailed per-level **RD**↓ results on M^{6}Doc-P dataset. **L1**, **L2**, and **L3** stand for the severity levels from light to heavy.

Backbone	Method	Modality V	L	T	Ext.	Rotation L1	L2	L3	Warping L1	L2	L3	Keystoning L1	L2	L3	Watermark L1	L2	L3	Background L1	L2	L3	Illumination L1	L2	L3	Ink-Bleeding L1	L2	L3	Ink-Holdout L1	L2	L3	Defocus L1	L2	L3	Vibration L1	L2	L3	Speckle L1	L2	L3	Texture L1	L2	L3
ResNet [6]	Faster R-CNN [12]	✓	✗	✗	✗	119.5	110.1	118.8	156.6	150.7	100.6	074.9	077.1	102.2	287.1	257.2	241.2	184.7	175.6	171.5	238.5	223.8	220.8	293.1	248.8	210.9	286.9	199.2	131.3	299.6	271.0	260.1	281.5	264.9	239.4	237.7	207.6	195.5	190.5	173.7	169.6
ResNet [6]	DocSegTr [2]	✓	✗	✗	✗	132.9	111.5	117.8	187.9	226.9	161.6	088.9	081.3	112.0	355.1	423.1	391.9	280.0	236.9	270.6	160.1	171.6	166.8	091.3	084.0	249.7	431.3	280.2	172.4	085.3	310.5	344.5	292.0	284.9	243.4	291.5	212.9	195.9	164.5	135.5	138.0
ResNet [6]	Mask R-CNN [7]	✓	✗	✗	✗	148.1	119.7	048.5	155.1	170.0	093.9	070.2	066.7	075.3	298.0	298.0	359.6	270.2	246.6	137.1	259.5	347.5	245.1	389.0	305.3	189.5	596.1	454.4	125.0	381.5	320.0	220.1	525.4	310.2	100.2	200.5	126.6	107.6	121.5	102.0	098.1
Swin [10]	SwinDocSegmenter [1]	✓	✗	✗	✗	148.1	121.9	116.7	214.0	208.8	142.7	098.5	097.7	129.5	398.0	346.7	321.4	239.8	215.2	224.0	324.3	295.2	288.9	411.4	335.1	248.4	399.9	271.0	161.8	417.5	344.2	236.3	374.5	220.6	208.8	294.4	200.5	168.4	195.8	163.1	152.0
DiT [9]	Cascade R-CNN [3]	✓	✗	✗	✓	084.2	065.8	062.1	152.5	142.6	102.8	057.4	058.5	079.6	298.8	254.3	281.0	198.5	175.3	180.7	173.8	155.7	157.7	293.6	242.7	184.1	290.0	226.7	170.7	278.6	235.3	163.1	323.1	217.0	230.7	175.1	103.3	085.1	096.7	078.4	075.1
LayoutLMv3 [8]	Cascade R-CNN [3]	✓	✓	✓	✓	084.9	068.2	067.9	138.9	134.7	108.8	063.8	060.3	078.7	267.5	234.7	225.1	176.2	172.9	184.1	206.7	199.3	225.4	372.9	313.6	222.4	275.2	255.0	199.6	323.0	284.7	190.8	381.8	258.1	246.7	178.9	113.6	090.9	101.5	090.4	096.3
InternImage [14]	RoDLA (Ours)	✓	✗	✗	✗	089.8	086.1	096.6	124.8	123.0	085.8	058.2	057.0	078.1	234.1	201.7	192.7	151.1	139.8	148.8	190.6	170.0	163.8	236.6	203.3	149.8	234.7	149.9	089.9	243.4	196.1	143.5	219.7	152.8	172.6	184.3	141.8	126.9	142.7	122.5	115.1

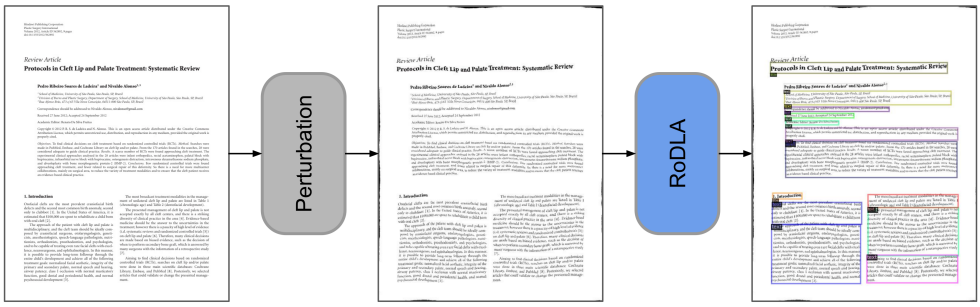
Fig. 2c depicting the combined use of DocTr [5] for image rectification followed by Faster R-CNN [12] inference. Notably, while DocTr [5] quickly rectifies spatial perturbations (**0.45s** per image), it is slow in recovering documents with illumination disturbances, typically requiring **8.2s** to rectify a single image. Furthermore, as evidenced in Fig. 2, it is apparent that for spatial perturbations, the DocTr [5] fails to adequately restore the original document images. Instead, it introduces additional spatial distortions, leading to erroneous predictions by Faster R-CNN [12].

5.2. Result Comparison

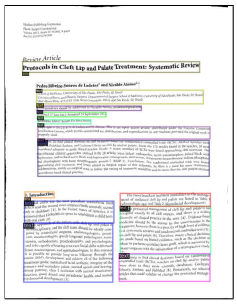
To elucidate the distinctions between these two inference pipelines in greater detail, Table 14 showcases a comparative analysis of the inference outcomes for our RoDLA model and Faster R-CNN [12] on three perturbation datasets. Specifically, it contrasts our model's performance against the combined DocTr [5] and Faster R-CNN [12] approach for four types of perturbations: **Rotation**, **Warping**, **Keystoning**, and **Illumination**. From Table 14, DocTr [5] with Faster R-CNN [12] Shows a significant drop in performance across all perturbations at the PubLayNet-P, DocLayNet-P, and M^{6}Doc datasets. This suggests DocTr [5], while rectifying images, may not be effectively handling these perturbations. Considering the time required for document image rectification, our model RoDLA not only demonstrates exceptional robustness but also exhibits high efficiency.

6. Visualization Results

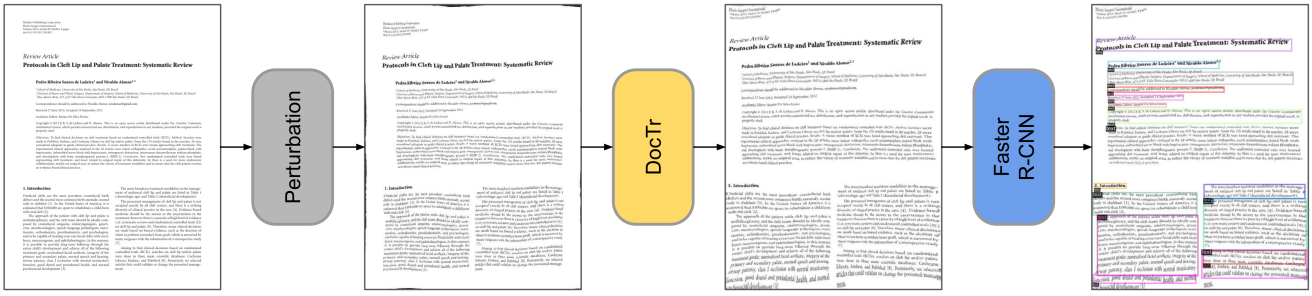
Beyond providing detailed performance data, to facilitate a more intuitive understanding, this section includes visualizations of selected model predictions on our robustness benchmark (*e.g.*, the PubLayNet-P dataset) for one document image perturbation in each of the following categories: Warping from Spatial, Watermark from Content, Illumina-



(a) Our RoDLA inference pipeline under perturbations



(b) The ground truth



(c) Two-stage inference pipeline under perturbations, including a document image rectification model (DocTr [5]) and a DLA model (Faster R-CNN [12]).

Figure 2. **Pipeline comparison** between (a) our one-stage inference (RoDLA) and (c) two-stage inference (first rectification, then DLA). Compared to (b) the ground truth, our RoDLA can obtain better DLA results.

Table 14. Result comparison between two-stage and one-stage pipelines. The **P-Avg**↑ are evaluated on **PubLayNet-P**, **DocLayNet-P**, and **M^{6}Doc-P** datasets. '-' means no document image rectification model has been implemented. Here, **P-Avg** only refers to the result for four types of perturbations: Rotation, Warping, Keystoning, and Illumination.

Rectification Model	**DLA Model**	**Clean**	Rotation			Warping			Keystoning			Illumination			**P-Avg**↑
			L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	
PubLayNet-P															
DocTr [5]	Faster R-CNN [12]	90.2	37.5	50.6	49.6	18.9	19.8	20.0	20.6	27.1	31.4	74.1	72.7	72.1	41.2
-	Faster R-CNN [12]	90.2	67.9	44.1	20.6	79.7	75.2	71.0	80.0	74.1	68.8	81.9	81.3	81.1	68.8
-	RoDLA (Ours)	96.0	71.9	19.9	02.9	89.0	80.4	68.5	88.4	81.2	72.1	92.0	91.4	91.5	**70.8**
DocLayNet-P															
DocTr [5]	Faster R-CNN [12]	73.4	10.9	21.0	20.9	04.5	04.2	04.0	04.8	06.5	07.8	62.1	61.4	60.8	22.4
-	Faster R-CNN [12]	73.4	37.6	11.0	01.4	65.2	60.4	58.8	62.1	52.7	46.3	70.6	70.5	70.1	50.6
-	RoDLA (Ours)	80.5	49.6	17.8	04.3	72.6	64.2	59.4	73.2	65.8	59.1	80.3	80.2	80.0	**58.9**
M^{6}Doc-P															
DocTr [5]	Faster R-CNN [12]	62.0	12.8	25.1	24.2	06.2	07.0	07.0	07.0	08.7	09.1	53.7	52.6	52.0	22.1
-	Faster R-CNN [12]	62.0	44.6	24.4	06.2	60.1	58.4	57.8	56.6	50.5	48.8	59.3	58.1	56.6	48.5
-	RoDLA (Ours)	70.0	58.4	40.9	23.7	68.2	66.0	64.0	66.3	63.4	60.9	67.5	68.2	67.8	**59.6**

tion from Inconsistency, Defocus from Blur, and Speckle from Noise. We further compare the visualization of predictions from Faster R-CNN [12], Mask R-CNN [7], and SwinDocSegmenter [1]. Observing the contrasts in Fig. 3, it becomes evident that our RoDLA model exhibits strong robustness against various perturbations, consistently yielding better predictions that align closely with the ground truth. In contrast, the other three models, *i.e.*, Faster R-CNN [12], Mask R-CNN [7], and SwinDocSegmenter [1], demonstrate varying degrees of erroneous predictions when confronted with different perturbations. Notably, SwinDocSegmenter displays particularly pronounced inaccuracies under the influence of these perturbations. The qualitative analysis proves the effectiveness of our proposed RoDLA in enhancing the robustness of DLA.

7. Discussion

7.1. Limitations and Future Works

Our benchmark, designed for testing the robustness of Document Layout Analysis (DLA) models, currently simulates only a subset of perturbations commonly encountered in document images. It does not account for content tampering or content replacement, which could significantly impact model robustness. Additionally, our robustness benchmark is limited to only three severity levels. This granularity might be too coarse, and further subdivision into more nuanced levels could reveal subtler variations in DLA model robustness across different intensities of perturbations. In addition to our current robustness benchmark, we have only separately identified potential perturbations and assessed their impacts individually. We have not yet combined multiple perturbations on a single image with a certain probability, which would more accurately reflect real-world scenarios. A comprehensive evaluation of these combined perturbations is necessary for a more realistic assessment of the robustness of DLA. Moreover, in the robustness evaluation metric, we have incorporated only two image quality assessment methods and have used the performance of the Faster R-CNN [12] as the reference for calculating the Mean Perturbation Effect (mPE). To enhance the objectivity of mPE and more accurately reflect the impact of perturbations on document images, it would be beneficial to include a broader range of image quality assessment methods and performances from various models.

7.2. Societal Impacts

Our robustness benchmark and RoDLA model for document layout analysis bear significant implications. While demonstrating a strong ability to withstand various perturbations and achieving impressive robustness metrics, the current evaluation of our model is confined to three document layout datasets. This step is critical to ensuring its applicability in diverse practical scenarios, such as digitizing historical documents, streamlining administrative processes, or enhancing accessibility for visually impaired individuals. As we transition from controlled datasets to varied real-life environments, continuous model refinement is necessary to address any unforeseen challenges, ensuring the model's relevance and effectiveness. Ethical considerations, particularly data integrity and impartiality, are paramount in avoiding biases and incorrect predictions that could have significant societal consequences. While our model demonstrates technical robustness, its deployment in real-world settings requires careful consideration of its broader societal implications to ensure it contributes positively and responsibly.

8. Acknowledgments

This work was supported in part by Helmholtz Association of German Research Centers, in part by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, and in part by BMBF through a fellowship within the IFI programme of DAAD. This work was partially performed on the HoreKa supercomputer funded by the MWK and by the Federal Ministry of Education and Research, partially on the HAICORE@KIT partition supported by the Helmholtz Association Initiative and Networking Fund, and partially on bwForCluster Helix supported by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

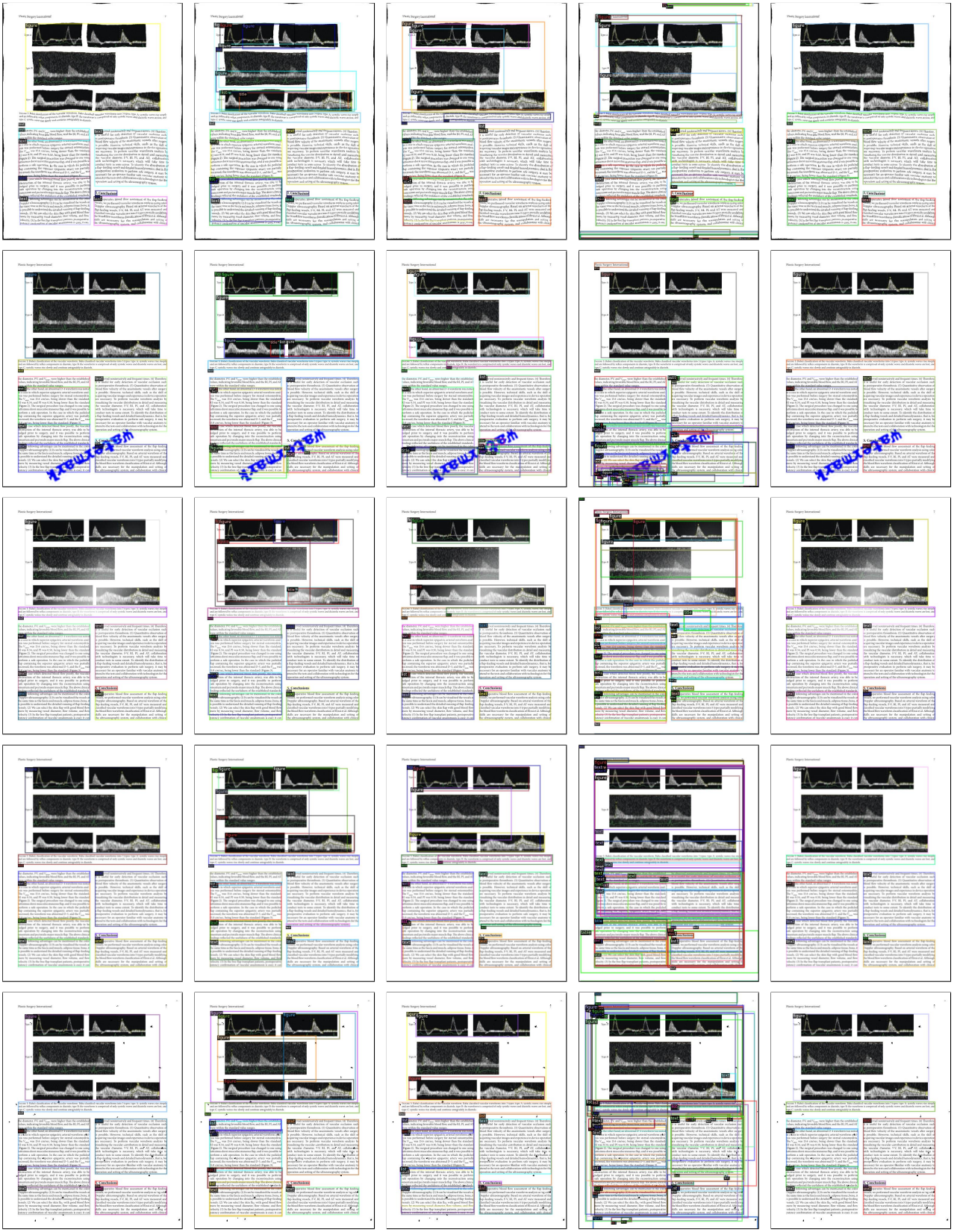


Figure 3. Visualizations on PubLayNet-P. From top to bottom: **Warping**, **Watermark**, **Illumination**, **Defocus**, and **Speckle**. From left to right: **the ground truth**, predictions from, *i.e.*, **Faster R-CNN** [12], **Mask R-CNN** [7], **SwinDocSegmenter** [1], and our **RoDLA**.

References

[1] Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. SwinDocSegmenter: An End-to-End Unified Domain Adaptive Transformer for Document Instance Segmentation. *ICDAR*, 2023. 5, 6, 7, 8, 10

[2] 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. 5, 6, 7

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 5, 6, 7

[4] 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. 1

[5] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. Deep Unrestricted Document Image Rectification. *arXiv preprint arXiv:2304.08796*, 2023. 6, 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5, 6, 7

[7] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5, 6, 7, 8, 10

[8] 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. 5, 6, 7

[9] 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. 5, 6, 7

[10] 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. 5, 6, 7

[11] 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. 1

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI*, 2017. 5, 6, 7, 8, 9, 10

[13] 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. 5

[14] 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

[15] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. 5

[16] 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

[17] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: largest dataset ever for document layout analysis. In *ICDAR*, 2019. 1

[18] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training. In *ICCV*, 2023. 5