7-2023

Do-GOOD: Towards distribution shift evaluation for pre-trained visual document understanding models

Jiabang HE

Yi HU

Lei WANG
Singapore Management University, lei.wang.2019@phdcs.smu.edu.sg

Xing XU

Ning LIU

See next page for additional authors

Citation

HE, Jiabang; HU, Yi; WANG, Lei; XU, Xing; LIU, Ning; and LIU, Hui. Do-GOOD: Towards distribution shift evaluation for pre-trained visual document understanding models. (2023). *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information, Taipei, July 23-27*. 569-579.

Available at: https://ink.library.smu.edu.sg/sis_research/8145

Do-GOOD: Towards Distribution Shift Evaluation for Pre-Trained Visual Document Understanding Models

Jiabang He
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Yi Hu
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Lei Wang*
Singapore Management University
Singapore

Xing Xu†
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Ning Liu
Beijing Forestry University
Beijing, China

Hui Liu
Beijing Rongda Technology Co., Ltd.
Beijing, China

Heng Tao Shen
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
Peng Cheng Laboratory
Shenzhen, China

ABSTRACT

Numerous pre-training techniques for visual document understanding (VDU) have recently shown substantial improvements in performance across a wide range of document tasks. However, these pre-trained VDU models cannot guarantee continued success when the distribution of test data differs from the distribution of training data. In this paper, to investigate how robust existing pre-trained VDU models are to various distribution shifts, we first develop an **o**ut-**o**f-**d**istribution (OOD) benchmark termed Do-GOOD for the fine-**G**rained analysis on **Do**cument image-related tasks specifically. The Do-GOOD benchmark defines the underlying mechanisms that result in different distribution shifts and contains 9 OOD datasets covering 3 VDU related tasks, *e.g.*, document information extraction, classification and question answering. We then evaluate the robustness and perform a fine-grained analysis of 5 latest VDU pre-trained models and 2 typical OOD generalization algorithms on these OOD datasets. Results from the experiments demonstrate that there is a significant performance gap between the in-distribution (ID) and OOD settings for document images, and that fine-grained analysis of distribution shifts can reveal the brittle nature of existing pre-trained VDU models and OOD generalization algorithms. The code and datasets for our Do-GOOD benchmark can be found at https://github.com/MAEHCM/Do-GOOD.

CCS CONCEPTS

• **Information systems** → **Document representation**; • **Applied computing** → **Document analysis**.

KEYWORDS

Visual Document Understanding, Out-of-distribution, Pre-trained Models, Document Information Extraction

ACM Reference Format:
Jiabang He, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. Do-GOOD: Towards Distribution Shift Evaluation for Pre-Trained Visual Document Understanding Models . In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), October 29– November 2, 2023, Ottawa, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591670

*Corresponding author.
†Corresponding author.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591670

1 INTRODUCTION

Background. Document images (e.g., invoices and lease agreements), typically containing rich contextual text and structural information, are commonly seen in modern working and living environments. Automatic processing and understanding of document

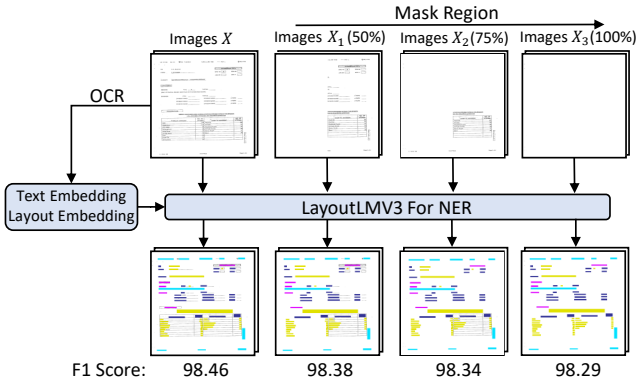


Figure 1: Illustration of the different importance of the input image, text, and layout embedding for the LayoutLMV3 [17] model on the document Named Entity Recognition (NER) task. Notably, though the image x is masked with different proportions, *i.e.* 50%, 75% and 100%, the model prediction (F1 score) just slightly changes.

images have wide-ranging use cases in real-world scenarios, such as document image classification [12, 13, 48], information extraction from document images [18, 31, 36], and document visual question answering [32]. Recently, numerous pre-training techniques concerning document image understanding have been proposed and shown to be effective for various document tasks [9, 17, 20, 26, 28, 44, 48]. Despite the encouraging results achieved by these models, it cannot be guaranteed that models designed under the same training and test data distribution would continue to perform well when the distribution of test data differs from the training data distribution [3, 24]. *However, most document datasets [12, 18, 32, 36] are designed following the i.i.d. assumption, with the training and test data from the same distribution.*

Motivation. To enable the models for document classification to have the ability to handle *out-of-distribution* (OOD) document images, Larson et al. [24] present a new OOD testbed in terms of a widely-used document classification benchmark dataset namely RVL-CDIP. This RVL-CDIP OOD benchmark is only used to develop and evaluate the robustness of methods for document image classification, which just need the models to have the capacity to model coarse-grained information over document images.

Although the RVL-CDIP OOD benchmark reveals that image information is quite important for document classification, image information *plays a relatively minor role on other document imaging tasks*, such as information extraction [39, 53]. As illustrated in Figure 1, taking the NER task on the latest pre-trained visual document understanding (VDU) model LayoutLMv3 [17] for example, when different proportions of an input image x are masked as blank, the F1 score of the LayoutLMv3 model prediction is basically the same. It indicates the prediction of the LayoutLMv3 model relies more on the text and layout information rather than the visual cues. Besides, document images naturally possess three distinct features, including image, text, and layout information. The tasks, such as information extraction from document images and document visual question answering, necessitate a fine-grained understanding of complicated interactions over image, text, and layout information [18, 32, 36]. On the other hand, models designed based on these three types of features require image, text, and layout modules to carry different perspectives of input information for a document image [17, 48, 50]. The uniqueness of document image data calls for the construction of document image specific OOD benchmarks with various distribution shifts. This naturally begs the following question: *How robust are existing pre-trained VDU models to fine-grained distribution shifts occurring on document image tasks?*

Contribution. To answer the above question regarding the robust estimation of the VDU models' capability in the document image OOD scenario, in this paper, we aim to develop a systematic document image OOD benchmark, namely Do-GOOD. To design Do-GOOD, we adhere to the following criteria. In particular, we expect that (1) A large distribution gap between training and test data can result in a substantial drop in model performance; (2) Fine-grained analysis of distribution shifts can expose the brittle nature of existing models; (3) Designed benchmark datasets should be possibly solvable, easily scalable, and human-readable.

To meet criteria (2), as shown in Figure 2, we divide distribution shifts into three categories of different characteristics, i.e., image, text, and layout distribution shifts. The distribution shifts are used to examine the partiality of VDU models on text, image, and layout information, which could compromise the robustness of VDU models. For image shift, we first disentangle the content (e.g., text on form images) from the background (e.g., table borders on form images) and then replace the background with a natural image from MSCOCO. For text shift, we employ common text attacks, such as BERT-Attack [27] and Word Swap [34, 35], to simulate a more realistic scenario where input document images may contain problematic text caused by OCR errors. We have two strategies to induce layout shifts. The first involves merging smaller bounding boxes to form a larger box. Another option is to move a particular box to a different location on the document image. These carefully designed strategies from image, text, and layout perspectives can automatically produce OOD testbeds having substantially different distributions from the training distributions, thus meeting criteria (1) and (2).

Here is a summary of our main contributions: (1) We provide a fine-grained analysis of various distribution shifts in document images from image, text, and layout perspectives; (2) To generate the OOD benchmark that meets the aforementioned three criteria, we introduce a suite of automatic strategies to generate OOD data; (3) We evaluate and compare 5 state-of-the-art pre-trained VDU models and 2 representative OOD algorithms in generated OOD testbeds (i.e., Do-GOOD) across different document image tasks. We hope that the proposed Do-GOOD benchmark, the empirical study, and our in-depth analysis will benefit future research to improve the robustness of pre-trained VDU models.

2 RELATED WORK

Visual Document Understanding. Visual document classification [12], visual document information extraction [18, 36], and visual question answering on documents [32] are among the core tasks of automated document processing. For visual document classification, early works model visual information by various

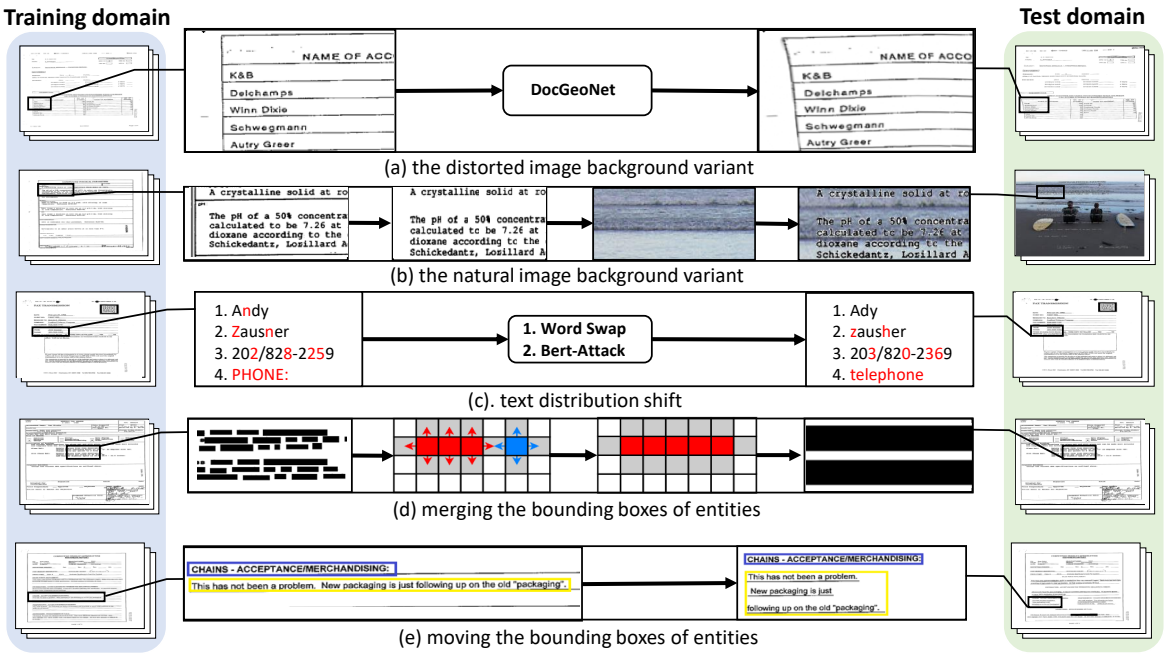


Figure 2: In the Do-GOOD benchmark, each document image is extracted from the training domain. We studied five distribution shifts acting on the three modalities respectively to generate the test domain. The five distribution shift acting includes: two image distribution shifts with (a) the distorted image background or (b) the natural image background; (c) text distribution shift with Bert-Attack and Word Swap; two layout distribution shifts by (d) merging and (e) moving the layouts.

CNN-based methods [13, 19]. Based on the output of OCR, RNN-based [41] and Transformer-based [31] models predict the label for the text. In DocVQA, an LSTM encoder is used to model textual information and CNN encoders are used to model visual information to answer questions about document images [32]. As a result of the recent success of large-scale pre-training in NLP, such as BERT [4] and RoBERTa [30], most methods take pre-train-and-fine-tune schemes for addressing downstream tasks together [7, 9, 17, 20, 25, 26, 28, 29, 38, 44, 47, 48, 50].

The majority of current state-of-the-art models separate scanned document images into text, vision, and layout attributes and design modules to process them individually or together. For example, most methods begin by obtaining text tokens and layouts from OCR tools and then feed the OCR-ed text into pre-trained language models to model the text information. For extracting region features, some works use object detectors [8, 28, 38, 48], while others use the vision transformer [17, 20, 21, 26]. LayoutLM [48] and its followings [49, 50] employ two-dimensional positional vectors for the layout information and fuse their transformed vectors with text embeddings for the multimodal pre-trained model. After collecting text, image, and layout features, most methods leverage a multimodal fusion module to encourage modeling interactions between them. For some exceptions, Donut [20] conducts inference in an end-to-end fashion without OCR processing. LayoutLMv3 [17] makes use of patch-level embeddings for text and image patches for alignment on document images.

Out-of-Distribution Benchmarks. We briefly review benchmarks for distribution shifts in this section. Distribution shifts have been a long-standing problem in the machine learning community [33, 42]. Recently, increasing research has shifted their attention from achieving the highest performance under in-distribution (ID) settings towards assessing models' robustness and generalization capacities [1, 2, 6, 23, 37, 40, 43, 52]. To this end, various OOD benchmarks have been created to encourage the building of more robust models [10, 11, 14, 22, 51]. WILDS [22] creates a curated benchmark of 10 datasets ranging from the categorization of animal species to code completion. This benchmark requires curated datasets that express large distribution shifts, are relevant in the real world, and can potentially be solved. The GOOD benchmark [10] is designed to graph OOD method evaluations based on two shifts. Beyond, Wiles et al. [45] provide a holistic analysis of current SOTA methods by evaluating multiple distinct methods across both synthetic and real-world datasets.

There are also OOD benchmarks for document image tasks. Larson et al. [24] establish an OOD testbed comprised of RVL-CDIP-N and RVL-CDIP-O. RVL-CDIP-N consists of in-domain documents sampled from a different distribution than RVL-CDIP. VL-CDIP-O comprises out-of-domain document images that do not fall into RVL-CDIP categories. The LastDoc4000 [3] is designed for situations in which input document images may contain unknown layouts and keys caused by OCR errors. However, the existing two benchmarks either ignore layout or text distribution shifts and only focus on document IE tasks. In contrast to them, Do-GOOD considers distribution shifts of text, vision, and layout across multiple common document image tasks from image-centric to text-centric perspectives.

3 DO-GOOD BENCHMARK DESIGN

Existing datasets, such as FUNSD [18], prepare training and test samples under the i.i.d. assumption. Given a data distribution p_{train} of training inputs x, the goal of a document image model f is to minimize the risk R as follows:

$$R(f) = \mathbb{E}_{(\boldsymbol{x}, y^l) \sim p_{\text{train}}} \left[\mathcal{L} \left(y^l, f(\boldsymbol{x}) \right) \right], \tag{1}$$

where $\mathcal{L}$ is the loss function for a particular task. Due to confounding factors, such as selection bias in the data collection process and random data splits, it is difficult for train and test data to follow the same data distribution in practice (i.e., $p_{\text{train}} \neq p_{\text{test}}$). As training and test data are distributed differently, models trained on training data are expected to generalize well to test data. This calls for carefully designed OOD benchmarks to accurately assess models' generalization abilities.

Taking inspiration from the recent fine-grained analysis of distribution shifts literature [45], we provide a fine-grained analysis of distribution shifts on document images by dividing them into attributes related to image, text, and layout to investigate why a model f trained on p_{train} should generalize to p_{test}. Specifically, a document image example is considered to be composed of the input x, label y^l, and its three attributes $\{y^{\text{image}}, y^{\text{text}}, y^{\text{layout}}\}$. As a convenience, we use $y^{1:K}$ to denote labels and attributes $\{y^l, y^{\text{image}}, y^{\text{text}}, y^{\text{layout}}\}$. Then, we are able to formalize different distribution shifts associated with image, text, and layout for the generation of true data as follows:

$$p\left(y^{1:K}, \boldsymbol{x}\right) = p\left(y^{1:K}\right) p\left(\boldsymbol{x} \mid y^{1:K}\right) \tag{2}$$

In this way, the data distribution can be expressed as the product of the marginal distributions of the decomposed attributes, which enables us to perform fine-grained analyses of various distribution shifts on document images. With the help of a latent variable model, the formalization can be written as follows:

$$p\left(y^{1:K}, \boldsymbol{x}\right) = p\left(y^{1:K}\right) \int p(\boldsymbol{x} \mid z) p\left(z \mid y^{1:K}\right) dz, \tag{3}$$

where z is the latent vector. Through the above equation, different attributes $y^{1:K}$ can be used to affect latent variables z, thereby affecting the generation of data $\boldsymbol{x}$.

3.1 Image-Specific Distribution Shift

The natural and the distorted image background are two background variants for image distribution shifts. Formally, y^{image} defines the image with the finite set $\mathcal{A} = \{a_{\text{original}}, a_{\text{natural}}, a_{\text{distorted}}\}$. For training, the attribute y^{image} is a_{original}. During testing on out-of-distribution data with natural image backgrounds, we set the attribute $y^{\text{image}} = a_{\text{natural}}$ and then obtain marginal distribution over this attribute $p_{\text{natural}}\left(y^{1:K}\right)$, which is used to induce the joint distribution over latent factors and the attribute *natural*: $p_{\text{natural}}\left(z, y^{1:K}\right) = p(z \mid y^{1:K}) p_{\text{natural}}\left(y^{1:K}\right)$. Subsequently, we can get input data for testing with the joint distribution: $p_{\text{natural}}\left(\boldsymbol{x}, y^{1:K}\right)$ equals to $\int p(\boldsymbol{x} \mid z) p_{\text{natural}}\left(z, y^{1:K}\right)$. On the other hand, the out-of-distribution test set with the background of distorted images $p_{\text{distorted}}\left(z, y^{1:K}\right)$ can be derived in a similar way to the generation of test data with natural images.

Practically, the new OOD benchmark with the joint distribution $p_{\text{natural}}\left(\boldsymbol{x}, y^{1:K}\right)$ can be obtained through a two-stage pipeline: (1) **Disentangling text content from background**. We locate the text content based on the position information provided by an OCR tool and extract pixels of the text content from a document image. The rest of the pixels are viewed as background pixels; (2) **Replacing the original background with natural images**. We randomly select an image from MSCOCO and resize it to match the size of the document image. To compose a new OOD sample, the extracted text content is placed on the sampled natural image (Figure 2 (b)).

Document images may be distorted in real-world scenarios due to uncontrollable physical deformations, uneven illuminations, and various camera angles. To simulate this realistic environment, the OOD benchmark with the joint distribution $p_{\text{distort}}\left(\boldsymbol{x}, y^{1:K}\right)$ is introduced. Inspired by [5], we directly employ well-pretrained DocGeoNet to generate the OOD distorted images (Figure 2 (a)).

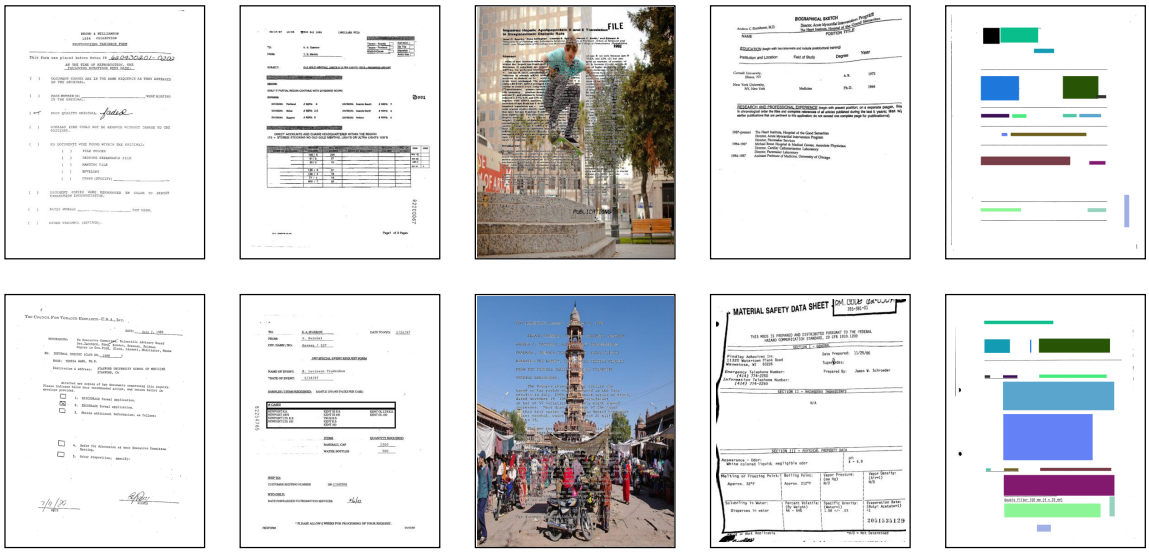
3.2 Text-Specific Distribution Shift

To simulate a realistic scenario where input document images may contain problematic text caused by OCR errors, we employ two text attack strategies for text distribution shifts (Figure 2 (c)): (1) Bert-Attack; (2) Word Swap. Formally, y^{text} defines the text with the finite set $\mathcal{A} = \{a_{\text{original}}, a_{\text{bert}}, a_{\text{swap}}\}$. For training, the attribute y^{text} is a_{original}. Refer to the analysis of image-specific OOD benchmarks, the out-of-distribution test data with BERT-Attack $p_{\text{generation}}\left(\boldsymbol{x}, y^{1:K}\right)$ and Word Swap $p_{\text{swap}}(\boldsymbol{x}, y^{1:K})$ can be obtained in a similar way.

In practice, BERT-Attack, based on pre-trained masked language models exemplified by BERT, is used to produce OOD samples. The advantage of BERT-Attack is that it can generate similar but unseen words while guaranteeing fluency and semantic preservation in the generated samples. For Word Swap, we apply 5 ways to generate OOD samples: (1) *Word Swap by embedding*: Using embedding vectors to find similar words to do swap; (2) *Word Swap by homoglyph*: Replacing words with nearly identical in appearance yet different meaning; (3) *Word Swap for numbers*: Replacing number with another number since numbers play an influential role in document images; (4) *Random character deletion*: Deleting certain characters in words, such as "houses" ⟶ "hoses".

3.3 Layout-Specific Distribution Shift

There are two layout manipulations for layout distribution shifts: Merge and Move. The Merge manipulation is designed to investigate the impact of changing layout information from a fine-grained level to a coarse-grained level while maintaining image and text information. The move operation is used to investigate the effect of neighboring information on the content of a particular bounding box by moving the content to a distinct location. The bounding box is an enclosed area surrounded by lines. Formally, y^{layout} defines the layout with the finite set $\mathcal{A} = \{a_{\text{original}}, a_{\text{merge}}, a_{\text{move}}\}$. For training, the attribute y^{layout} is a_{original}. Refer to the analysis of image-specific and text-specific OOD benchmarks, the OOD test data with the Merge manipulation $p_{\text{merge}}\left(\boldsymbol{x}, y^{1:K}\right)$ and the OOD test data with the Move manipulation $p_{\text{move}}(\boldsymbol{x}, y^{1:K})$ can be obtained in a similar way.



(a) FUNSD-R (b) FUNSD-H (c) RVL-CDIP-I_1 (d) RVL-CDIP-I_2 (e) RVL-CDIP-L

Figure 3: Samples of the distribution shift examples: (a) FUNSD-R containing a real-world OOD dataset variant of FUNSD, (b) FUNSD-H denoting the human-intervened OOD dataset variant of FUNSD, (c) RVL-CDIP-I_1 including samples with the natural image background, (d) RVL-CDIP-I_2 including samples with the distorted image background, and (e) RVL-CDIP-L containing samples with merged bounding boxes. More samples can be found at https://github.com/MAEHCM.

The pseudocode of merging bounding boxes is shown in Algorithm 1. To construct the OOD benchmark $p_{\text{merger}}(\boldsymbol{x}, y^{1:K})$, we perform the process described in Algorithm 1 for each image. The Merge manipulation begins with initializing an empty set S which saves all the bounding boxes that have been traversed. We then traverse each bounding box and get the current bounding box $\mathbf{B_i}$. Then it gets $\mathbf{B_i^{'}}$ by dilating $\mathbf{B_i}$ a litte using predefined horizontal and vertical dilation distances λ_1, λ_2. After that, we see if this dilated bounding box intersects with another bounding box in set S. If there is an intersection, we get $\mathbf{M_i}$ by merging the two bounding boxes, otherwise, we skip this operation. We then add $\mathbf{M_i}$ or $\mathbf{B_i^{'}}$ to set S. After all bounding boxes have been traversed, the merging process is complete (Figure 2 (e)) and we get merged bounding boxes $\mathbf{M_1}$,$\mathbf{M_2}$,...,$\mathbf{M_k}$. For the construction of OOD benchmark $p_{\text{move}}(\boldsymbol{x}, y^{1:K})$, we select a bounding box with strong textual semantics and then move the text content to another location without textual semantics (Figure 2 (e)). A semantic entity is considered to have strong text semantics if its prediction results remain unchanged after its corresponding layout information has been changed ten times.

Algorithm 1 The procedure of merging bounding boxes.

Input: Bounding boxes $\mathbf{B_1}$,$\mathbf{B_2}$,...,$\mathbf{B_n}$ of a image; the collection S of bounding boxes that have been traversed; horizontal and vertical dilation distances λ_1, λ_2.

```
Initialize S to empty set.
repeat
    Get the i-th bounding box B_i[x_1^i, y_1^i, x_2^i, y_2^i].
    Dilate B_i with horizontal and vertical dilation distances as
    B_i'[x_1^i − λ_1, y_1^i − λ_2, x_2^i + λ_1, y_2^i + λ_2].
    if B_i' intersects with a bounding box in the set S_i of S then
        Merge the two bounding boxes in S_i.
    end if
    Mark the area where B_i' is located in S_i.
until All the bounding boxes have been traversed.
```

Output: Merged bounding boxes $\mathbf{M_1}$,$\mathbf{M_2}$,...,$\mathbf{M_k}$.

4 DO-GOOD DATASETS

The purpose of this section is to introduce the datasets used in our proposed Do-GOOD benchmark. we first perform a preliminary study for analysis of the distribution shift options for a specific VDU task. Then, we elaborate on OOD datasets across different VDU tasks. Finally, 9 datasets are constructed across 3 VDU tasks.

4.1 Preliminary Study

We conducted a preliminary study in order to investigate the effect of image, text, and layout information across different datasets and VDU tasks. Thus, for a certain dataset and task, we can determine which distribution shift should be chosen to develop the OOD test dataset. During implementation, we use LayoutLMv3$_{\text{BASE}}$ [17] as the base model and isolate the effects of image, layout, and text information by removing the corresponding input embeddings for inference. Text is necessary for all tasks. Thus, to assess the effect of text information, we retain input text and remove image and layout embeddings.

The overall results are shown in Table 1. While the performance of LayoutLMv3 without (denoted as "w/o") image embeddings on RVL-CDIP drops substantially, the performance of information extraction slightly decreases and the performance of QA tasks may even increase. It indicates that document image classification is largely affected by image information. We observe that without layout information, the performance of LayoutLMv3 drops by a significant margin for information extraction and classification. However, model performance on DocVQA is not affected. We assume that most questions in DocVQA are dependent on textual content to predict the answers. The performance of LayoutLMv3 only with text embeddings is slightly better than that with all input embeddings, which strongly supports our assumption of DocVQA. Besides, using only text embeddings, LayoutLMv3 performs very poorly on information extraction and classification tasks. All of these analyses motivate us to develop image-specific OOD datasets for document image classification, text-specific OOD datasets for all tasks, and layout-specific OOD datasets for document image classification and information extraction.

Table 1: Overall results of the preliminary study on FUNSD, RVL-CDIP, and DocVQA datasets.

Model	FUNSD F1↑	RVL-CDIP Accuracy↑	DocVQA ANLS↑
$\text{LayoutLMv3}_{\text{BASE}}$	90.29	95.44	78.76
w/o Image	90.18	57.07	78.82
w/o Layout	29.87	77.05	78.76
w/ Text	28.65	18.07	78.82

4.2 Document Information Extraction Task

For the visual document information extraction task, we mainly generate OOD datasets based on FUNSD [18]. FUNSD is a dataset sampled from the RVL-CDIP dataset [12] about noisy scanned form understanding, consisting of 199 documents (149 for training and 50 for testing) and 9,743 semantic entities. The task of FUNSD is sequential labeling, which aims to assign labels to words.

FUNSD-L is a variant of FUNSD that includes the OOD samples produced through strategies based on two layout-specific distribution shifts, Merger and Move. As described in Section 3, the Move operation is based on semantic strength determined by the model itself. Specifically, we randomly shuffle the bounding boxes within a document image and employ the fine-tuned model to infer 30 times. For textual content, if the model prediction has fewer errors, its semantic strength is greater. **FUNSD-T** is a variant of FUNSD that contains the OOD samples generated by the two text attack methods described in Section 3.

In fact, we also have OOD samples through observing and selecting from real-world datasets for the visual document information extraction task. Figure 3 (c) and (d) show examples. Specifically, **FUNSD-R** is one real-world OOD dataset variant of FUNSD. The FUNSD-R dataset is used to show the performance gap of VDU models on real-world OOD datasets and generated OOD datasets. First, we sample data examples from the large-scale document classification dataset RVL-CDIP, and then observe and select data examples that differ from the distribution of FUNSD. After that, these selected examples are manually annotated. FUNSD-R contains 50 document images in total. Further, we modify the data examples in FUNSD in order to generate a human-intervented OOD dataset variant of FUNSD, named **FUNSD-H**. Practically, we move some weak textual entities to construct layout and image shift, or add a few semantically linked texts around strong semantic content to construct 3 kinds of shifts. In the end, we obtain 50 OOD samples. Despite the fact that it is expensive and time-consuming to construct OOD datasets such as FUNSD-R and RUNSD-H, these two OOD datasets inspired us to develop a suite of OOD benchmark datasets that can be generated automatically for a wide range of VDU tasks.

4.3 Visual Document Classification Task

RVL-CDIP [12] is a document classification dataset aiming to predict the category of a given document. It includes 400,000 data examples in 16 categories, which are divided into 320,000 training samples, 40,000 validation samples, and 40,000 test samples.

RVL-CDIP-T is one of the OOD dataset variants of RVL-CDIP that contains the OOD samples generated by the two text attack methods. As a variant of RVL-CDIP, **RVL-CDIP-L** includes OOD samples produced through two layout-specific distribution shifts. The image-specific OOD variant, **RVL-CDIP-I**, is generated through natural $\textbf{RVL-CDIP-I}_1$ and distorted $\textbf{RVL-CDIP-I}_2$ image distribution shifts. Examples are illustrated in Figure 3.

4.4 Document Visual Question Answering Task

DocVQA [32] is a dataset for predicting the answer given a document image and a question. To accomplish this, models need to understand the content of documents and learn to reason over them. The original DocVQA dataset consists of 10,194/1,286/1,287 images with 39,463/5,349/5,188 questions for training/validation/test, respectively.

DocVQA-T is the OOD dataset variant of DocVQA. In order to construct DocVQA-T, we first collect text, questions, and answers from OCR results and the Microsoft READ API. Then, we obtain the OOD samples which are generated by the two text attack methods.

5 EXPERIMENT

5.1 Evaluation on state-of-the-art VDU Models

VDU Models. Larger models are generally more robust to OOD data [15]. We thus evaluate the robustness of fine-tuning the popular pre-trained VDU models (large models) for downstream tasks on our Do-GOOD benchmark. The state-of-the-art large models include (1) Pre-trained models with text and layout modalities: BROS [16] and LiLT [44]; (2) Pre-trained models with text, layout and image modalities: LayoutLMv1 [48], LayoutLMv2 [50], and LayoutLMv3 [17].

Implementation Details. We fine-tune the VDU models on the ID datasets while selecting the best checkpoints based on the performance of ID and OOD validation sets. The evaluation metrics we use are the same as those used in the original dataset paper, such as "`F1`" for FUNSD and its OOD variants, "`Accuracy`" for RVL-CDIP and its OOD variants, and "`ANLS`" for DocVQA and its variants. All pre-trained models are based on Hugging Face [46]. For the visual document information extraction task, the learning rate is

Table 2: The ID and OOD performance of existing VDU models on the FUNSD, RVL-CDIP and DocVQA datasets. To compare the models fairly, all VDU models use cell-level layout embedding. Here OOD_R is an OOD dataset of FUNSD which samples 50 images from RVL-CDIP. OOD_H is our handcrafted dataset. OOD_T, OOD_L, OOD_{I_1} and OOD_{I_2} refer to text distribution shift, layout distribution shift, natural image distribution shift and distorted image distribution shift.

Model	FUNSD					RVL-CDIP					DocVQA	
	ID	OOD_R	OOD_H	OOD_T	OOD_L	ID	OOD_T	OOD_L	OOD_{I_1}	OOD_{I_2}	ID	OOD_T
$BROS_{BASE}$ [16]	88.98	60.20	**74.31**	80.58	84.37	90.12	88.43	80.56	90.12	85.32	73.72	60.38
$LiLT_{BASE}$ [44]	88.25	57.45	72.32	78.41	68.83	95.68*	85.31*	51.20*	**95.68***	92.42*	70.43	52.31
$LayoutLM_{BASE}$ [48]	82.82	47.94	68.44	72.23	54.64	94.42	81.35	54.77	94.42	87.59	69.34	59.26
$LayoutLMv2_{BASE}$ [50]	89.91	**62.39**	72.70	79.16	81.33	95.25	86.53	64.78	82.08	**92.16**	78.08	64.67
$LayoutLMv3_{BASE}$ [17]	**90.29**	57.88	73.25	**86.82**	**84.95**	**95.44**	**89.32**	**81.06**	86.27	85.02	**78.76**	**65.69**

* LiLT uses image features with ResNeXt101-FPN backbone in fine-tuning RVL-CDIP.

set to 3e-5, and the training epochs are set to 70. Since the original RVL-CDIP corpus did not provide text information, we used the Tesseract 3 OCR engine to extract words and their positions. The learning rate was set to 1e-6, and the training epoch was 30 rounds. For Doc VQA tasks, the learning rate is set to 2e-5, and the epoch is 40 rounds. All input images have a resolution of 224×224 pixels, and the batch in training is set to 4, while the batch in testing is set to 1.

Main Results. Based on the criteria outlined in Section 1, Do-GOOD is designed to achieve a large distribution gap between training and test data and a substantial performance drop from ID to OOD settings. To verify whether the proposed OOD benchmark meets the criteria, we conduct experiments fine-tuning pre-trained VDU models on the original ID downstream datasets and testing on both the ID and OOD datasets.

Table 2 reports the overall results *w.r.t* comparison of ID and OOD performance of the existing models on the FUNSD, RVL-CDIP and DocVQA datasets. According to the differences between ID and OOD for each distribution shift across all VDU tasks, there is a substantial and consistent performance gap between the ID and OOD settings. In most cases, LayoutLMv3 can achieve the best performance, including ID setting across all datasets, OOD_T and OOD_I settings of FUNSD and RVL-CDIP, and the OOD_T setting of DocVQA, indicating LayoutLMv3 is one of the most robust models on VDU tasks. These motivate us to use LayoutLMv3 as our base model for comparing common OOD algorithms on Do-GOOD benchmark. BROS performs well in 4 OOD settings on FUNSD. The possible reason is that pre-training in BROS uses the relative position of the encoded text and a region masking strategy as the objective. Based on the success of LayoutLMv3 and BROS, we assume that fine-grained modeling such as patch-level or region-level modeling may be very useful for improving the robustness of models in OOD environments.

Results on FUNSD-L Dataset. Furthermore, to explore the robustness of each model under the layout distribution shift condition, we evaluate the performance of each model in each label category on FUNSD-L. As shown in Table 3, we observe that LayoutLMv3 achieves the best performance on the FUNSD-L dataset. BROS obtains the worst Other Error score. LayoutLM [48] and LayoutLMv2[50] have higher Other Error, indicating that the prediction of weak semantic areas can be easily affected by the layout of strong semantic areas in these models. LayoutLM also has higher QA Error, indicating that the prediction of strong semantic entities may still be affected by the surrounding entities. The low header accuracy of all models indicates that the prediction of the current model for the headers largely depends on the location of the entities.

Results on FUNSD-T Dataset. Moreover, to investigate the robustness of each model under the text distribution shift condition, we evaluate the performance of the model against various text attacks. Table 4 shows the performance of each model under various attacks on FUNSD-T. We observe that LayoutLMv3[17] achieves the best performance on 5 out of 6 text distribution shifts, which indicates that LayoutLMv3 is more robust than other models on text attacks. The performance gap between LayoutLMv3 and other VDU models on homoglyph is about 20 to 30 F1 score, which indicates that LayoutLMv3 model is more robust than other models when dealing with the semantic OOD caused by OCR error.

5.2 Evaluation on Typical OOD Algorithms

We further compare the representative OOD algorithms on all OOD datasets across three downstream tasks. Based on the experiment results, we briefly analyze the effect of different OOD methods. All experimental results are based on $LayoutLMv3_{BASE}$ [17].

Baseline Methods. We use empirical risk minimization (ERM) and two OOD algorithms as our baselines. ERM is a systematic process of identifying, assessing, and managing risks that face an organization. The goal of ERM is to maximize the potential of positive events and minimize the impact of negative ones. The 2 OOD methods are Deep Coral [43] and Mixup [52]. Deep Coral achieves domain adaptive effects by aligning second-order statistics between the source and target domains. In our experiment, we only add this method at the last layer of $LayoutLMv3_{BASE}$ model and set λ equal to 1. Mixup [52] achieves data augmentation without excessive overhead by interpolating input features and labels. In our implementation, we simultaneously interpolate input features, including text embedding, layout embedding, and image embedding, and then set α and β equal to 0.4 for the Beta distribution.

Main Results. Table 5 shows the ID and OOD results of ERM, Deep Coral, and Mixup on 3 downstream tasks. According to the observation in Table 5, none of the OOD generalization algorithms consistently outperform ERM, and even ERM is superior to Deep Coral in most cases. Mixup can outperform ERM in OOD_R and

Table 3: Overall comparison results of existing VDU models on the FUNSD and their own FUNSD-L datasets. FUNSD-L is generated by the move operation. Other Error, QA Error and Header Error refer to error rate of entities whose labels are other, question or answer, and header respectively.

Model	FUNSD	FUNSD-L					
	F1↑	Precision↑	Recall↑	F1↑	Other Error↓	QA Error↓	Header Error↓
$BROS_{BASE}$ [16]	89.26	77.13	93.10	84.37	**43.97**	1.79	100.00
$LiLT_{BASE}$ [44]	88.25	60.03	80.66	68.83	54.92	13.18	**59.46**
$LayoutLM_{BASE}$ [48]	82.82	53.68	55.64	54.64	82.88	42.32	72.73
$LayoutLMv2_{BASE}$ [50]	89.91	71.74	**93.89**	81.33	81.73	**1.24**	95.74
$LayoutLMv3_{BASE}$ [17]	**90.29**	**80.40**	90.05	**84.95**	45.46	4.13	100.00

Table 4: Comparison results of existing VDU models under different text attack methods on the FUNSD dataset. All numerical results are averages of 5 runs.

Model	Baseline	BERT-Attack	Embedding	Homoglyph	Change number	Character deletion
	F1↑	F1↑	F1↑	F1↑	F1↑	F1↑
$BROS_{BASE}$ [16]	88.98	**89.04**	82.55	66.56	89.23	75.51
$LiLT_{BASE}$ [44]	88.25	84.54	81.01	70.23	87.28	68.97
$LayoutLM_{BASE}$ [48]	82.82	79.75	75.80	56.99	82.31	66.31
$LayoutLMv2_{BASE}$ [50]	89.91	86.61	83.60	60.83	89.27	75.53
$LayoutLMv3_{BASE}$ [17]	**90.29**	88.14	**86.44**	**84.50**	**90.10**	**84.91**

Table 5: The ID and OOD performances of 3 OOD algorithms on 12 datasets. All numerial results are averages of 5 runs.

Algorithm	FUNSD					RVL-CDIP					DocVQA	
	ID	OOD_R	OOD_H	OOD_T	OOD_L	ID	OOD_T	OOD_L	OOD_{I_1}	OOD_{I_2}	ID	OOD_T
ERM	**90.29**	57.88	73.25	**86.82**	**84.95**	**95.44**	89.32	**81.06**	36.27	85.02	**78.76**	**65.69**
Deep Coral [43]	90.20	58.88	73.92	84.61	83.47	95.12	89.21	76.57	37.82	86.23	78.63	64.21
Mixup [52]	89.28	**61.19**	**74.33**	86.53	84.23	94.69	**89.87**	78.70	**40.44**	**87.09**	77.66	65.34

OOD_H settings on FUNSD, but it underperforms ERM in OOD_T and OOD_L settings, indicating that developing a fine-grained comprehensive evaluation is of importance for OOD generalization.

Next, we take a closer look at the effectiveness of the OOD algorithm on different tasks. For the information extraction task, Deep Coral and Mixup outperform ERM in OOD_R and OOD_H. These results demonstrate the rationality of our manual labeling data set and prove that common OOD algorithms are also applicable to VDU models. In terms of layout distribution shifts on FUNSD, ERM performs slightly better than Deep Coral and Mixup. This may be due to the fact that common OOD algorithms are not capable of coping with excessive layout information changes. For document image classification, Deep Coral and Mixup outperform ERM in OOD_{I_1} and OOD_{I_2} settings while they still perform worse than ERM in OOD_L settings. The results of this study indicate that common OOD algorithms perform well in the task of document classification, which requires a high level of image information modeling. Deep Coral and Mixup score slightly below ERM on the document visual question answering task. The study demonstrates that distribution shifts in complex tasks such as document visual question answering cannot be easily handled using common OOD algorithms.

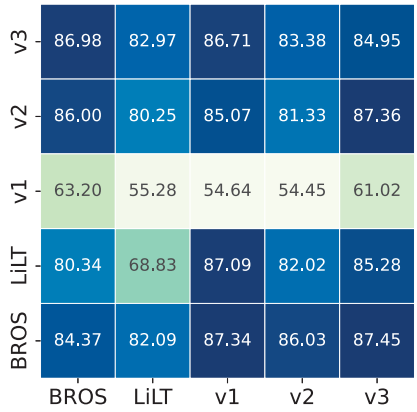
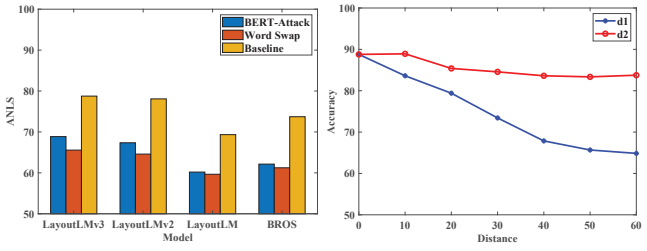


Figure 4: The confusion matrix in terms of F1 score for each VDU model on FUNSD-L data generated by the other models. The columns are the VDU models for generating the data, and the rows are the models for testing the data. v3 means LayoutLMv3, v2 means LayoutLMv2, v1 means LayoutLM.

5.3 Further Analysis

Effect of OOD Samples Generated by Different VDU Models. As the generation of samples in FUNSD-L relies on the model to assess the semantic strength, we conduct experiments to investigate whether the performance of the model also drops substantially when OOD samples *w.r.t* layout distribution shift are generated by other models. Figure 4 shows the confusion matrix for each VDU model on FUNSD-L data generated by the other models. We can observe that LayoutLM consistently performs worse on OOD datasets generated by all models. The results indicate that LayoutLM trained with fixed layout information is strongly dependent on layout information, which makes it difficult to cope with layout distribution shifts. Both LayoutLMv3 and BROS perform well on OOD datasets generated by all models, including themselves. It demonstrates that fine-grained information modeling, such as patch-level and region-level information modeling, can improve the robustness of models.

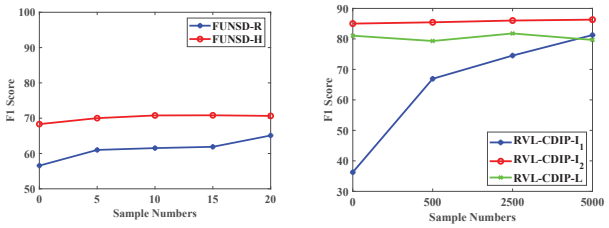


(a) (b)

Figure 5: Further analysis on (a) text distribution shift of VDU models on the DocVQA dataset. v3 means LayoutLMv3, v2 means LayoutLMv2, v1 means LayoutLM and (b) layout shift of LayoutLMv3 on document classification dataset RVL-CDIP.

Effect of Text Shift on Document VQA Task. In Section 4.1, we have demonstrated that LayoutLMv3 rarely uses the visual or layout information in document VQA tasks, thus for DocVQA test sets in this experiment, we only concentrate on the text information. We utilize the same text shift method as FUNSD for text which is not answer. Figure5a shows the results. It can be seen that under the influence of Bert-Attack or Word Swap, the ANLS of all models dropped by about 10 points in terms of Accuracy. It indicates that the existing VDU models are vulnerable to image corruption or OCR errors for document VQA task.

Effect of Merge Distance. We further conduct experiments on the impact of distance parameter d, and the experimental results are shown in Figure 5b. Note that d_1 equals to λ_1 means vertical spacing and d_2 equal to λ_2 means horizontal spacing. We can observe that some overlapping bounding boxes during OCR detection, thus, we explore whether the model needed fine-grained layout coordinates to predict document categories. Here d_1 controls the horizontal stretch length while d_2 controls the vertical stretch length. When d_1 and d_2 are both 0, part of the OCR overlap area merges and the accuracy decreases. It indicates that longitudinal merging reduces prediction accuracy more than horizontal merging.



(a) (b)

Figure 6: Incremental training results of LayoutLMv3 on (a) FUNSD-R and FUNSD-H, and (b)RVL-CDIP-I_1, RVL-CDIP-I_2 and RVL-CDIP-L datasets.

Effect of Incremental Training with Do-GOOD. We finally investigate the impact of the Do-GOOD benchmark on solving the OOD problem for existing VDU models considering the incremental training scheme. Specifically, we divide the FUNSD-R and FUNSD-H datasets into training and test data split. Specially, 20 samples are added to the FUNSD training set for incremental training, 30 samples are tested as OOD samples. We randomly sample five times and take the average of all results. The experimental results show in Figure 6a and we can see that adding OOD sample during training is effective to improve the performance of OOD test sets.

We further sample 5,000 samples on the RVL-CDIP validation set for incremental training and ensure that all document types are evenly distributed. As shown in Figure 6b, the experimental results show that adding OOD data to the training set can significantly improve the performance of the model on such OOD test sets when natural scene background replacement occurs for the document background. For image distortion and Layout shift joining the training set to participate in incremental training, we find that the performance slightly changes on the OOD test set.

6 CONCLUSION

In this paper, we introduced an out-of-distribution (OOD) benchmark, *i.e.* Do-GOOD, that evaluates the robustness of existing VDU models for document image-related tasks. We presented three criteria as well as a general, comprehensive framework for analyzing and benchmarking OOD document images. In this framework, we first broken down document images into image, text, and layout characteristics. Then, we discussed the distribution shifts from image, text, and layout perspectives. We finally obtained 9 OOD datasets covering 3 document image-related tasks. On the basis of these OOD datasets, we conducted experiments using 5 existing pre-trained VDU models and two commonly used OOD generalization algorithms, which demonstrate the brittle nature of existing VDU models and OOD generalization algorithms. We expected that our framework and comprehensive benchmark will facilitate research in document image-related fields, and it can be utilized by practitioners to determine which methods perform best under which distribution shifts.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grants (No. 62222203 and 61976049).

REFERENCES

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems* 34 (2021), 3438–3450.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[3] Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. 2022. Query-driven Generative Network for Document Information Extraction in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4261–4271.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[5] Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. 2022. Geometric Representation Learning for Document Image Rectification. In *European Conference on Computer Vision*. Springer, 475–492.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[7] Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. LAMBERT: Layout-Aware Language Modeling for Information Extraction. In *ICDAR*.

[8] Jiuxiang Gu, Jason Kuen, Vlad Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. UniDoc: Unified Pretraining Framework for Document Understanding. In *NeurIPS*.

[9] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding. In *CVPR*.

[10] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. 2022. GOOD: A Graph Out-of-Distribution Benchmark. *arXiv preprint arXiv:2206.08452* (2022).

[11] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434* (2020).

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

[13] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 991–995.

[14] Yue He, Zheyan Shen, and Peng Cui. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition* 110 (2021), 107383.

[15] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100* (2020).

[16] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A Pre-Trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In *AAAI*.

[17] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv preprint arXiv:2204.08387* (2022).

[18] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*.

[19] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for document image classification. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 3168–3172.

[20] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *European Conference on Computer Vision*. Springer, 498–517.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

[22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[23] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*. PMLR, 5815–5826.

[24] Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating Out-of-Distribution Performance on Document Image Classifiers. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[25] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural Pre-training for Form Understanding. In *ACL*.

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

[27] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984* (2020).

[28] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. SelfDoc: Self-Supervised Document Representation Learning. In *CVPR*.

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

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[31] Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*. 6495–6504.

[32] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

[33] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45, 1 (2012), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[34] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909* (2020).

[35] Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892* (2016).

[36] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In *Document Intelligence Workshop at Neural Information Processing Systems*.

[37] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78, 5 (2016), 947–1012.

[38] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Pałka. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *ICDAR*.

[39] Dongyu Ru, Zhenghui Wang, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2020. QuAChIE: Question Answering based Chinese Information Extraction System. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2177–2180.

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).

[41] Clément Sage, Alexandre Aussem, Haytham Elghazel, Véronique Eglin, and Jérémy Espinas. 2019. Recurrent neural network approach for table field extraction in business documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1308–1313.

[42] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* (2021).

[43] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

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

[45] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. 2021. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328* (2021).

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[47] Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. 2021. LAMPRET: Layout-Aware Multimodal PreTraining for Document Understanding. *arXiv preprint arXiv:2104.08405* (2021).

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

[49] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. *arXiv preprint arXiv:2104.08836* (2021).

[50] 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. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL*.

[51] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. 2021. OoD-Bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721* (2021).

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[53] Yue Zhang, Hongliang Fei, and Ping Li. 2022. End-to-end Distantly Supervised Information Extraction with Retrieval Augmentation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2449–2455.