DiT: Self-supervised Pre-training for Document Image Transformer

Junlong Li*
lockonn@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yiheng Xu*
t-yihengxu@microsoft.com
Microsoft Research Asia
Beijing, China

Tengchao Lv
tengchaolv@microsoft.com
Microsoft Research Asia
Beijing, China

Lei Cui
lecu@microsoft.com
Microsoft Research Asia
Beijing, China

Cha Zhang
chazhang@microsoft.com
Microsoft Azure AI
Redmond, United States

Furu Wei
fuwei@microsoft.com
Microsoft Research Asia
Beijing, China

ABSTRACT

Image Transformer has recently achieved significant progress for natural image understanding, either using supervised (ViT, DeiT, etc.) or self-supervised (BEiT, MAE, etc.) pre-training techniques. In this paper, we propose **DiT**, a self-supervised pre-trained **D**ocument **I**mage **T**ransformer model using large-scale unlabeled text images for Document AI tasks, which is essential since no supervised counterparts ever exist due to the lack of human-labeled document images. We leverage DiT as the backbone network in a variety of vision-based Document AI tasks, including document image classification, document layout analysis, table detection as well as text detection for OCR. Experiment results have illustrated that the self-supervised pre-trained DiT model achieves new state-of-the-art results on these downstream tasks, e.g. document image classification (91.11 → 92.69), document layout analysis (91.0 → 94.9), table detection (94.23 → 96.55) and text detection for OCR (93.07 → 94.29). The code and pre-trained models are publicly available at https://aka.ms/msdit.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

document image transformer, self-supervised pre-training, document image classification, document layout analysis, table detection, text detection, OCR

ACM Reference Format:
Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547911

*Contributions during internship at Microsoft Research Asia. Corresponding authors: Lei Cui and Furu Wei.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547911

1 INTRODUCTION

Self-supervised pre-training techniques have been the de facto common practice for Document AI [10] in the past several years, where the image, text, and layout information is often jointly trained using a unified Transformer architecture [2, 19, 21, 25, 28, 32, 33, 38, 41–44, 48]. Among all these approaches, a typical pipeline for pre-training Document AI models usually start with the vision-based understanding such as Optical Character Recognition (OCR) or document layout analysis, which still heavily relies on the supervised computer vision backbone models with human-labeled training samples. Although good results have been achieved on benchmark datasets, these vision models are often confronted with the performance gap in real-world applications due to domain shift and template/format mismatch from the training data. Such accuracy regression [26, 46] also has an essential influence on the pre-trained models as well as downstream tasks. Therefore, it is inevitable to investigate how to leverage the self-supervised pre-training for the backbone of document image understanding, which can better facilitate general Document AI models for different domains.

Image Transformer [3, 9, 12, 13, 17, 31, 36, 47] has recently achieved great success for natural image understanding including classification, detection and segmentation tasks, either with supervised pre-training on the ImageNet or self-supervised pre-training. The pre-trained image Transformer models can achieve comparable and even better performance compared with CNN-based pre-trained models under a similar parameter size. However, for document image understanding, there is no commonly-used large-scale human-labeled benchmark like ImageNet, which makes large-scale supervised pre-training impractical. Even though weakly supervised methods have been used to create Document AI benchmarks [26, 27, 45, 46], the domain of these datasets is often from the academic papers that share similar templates and formats, which are different from real-world documents such as forms, invoice/receipts, reports, and many others as shown in Figure 1. This may lead to unsatisfactory results for general Document AI problems. Therefore, it is vital to pre-train the document image backbone models with large-scale unlabeled data from general domains, which can support a variety of Document AI tasks.

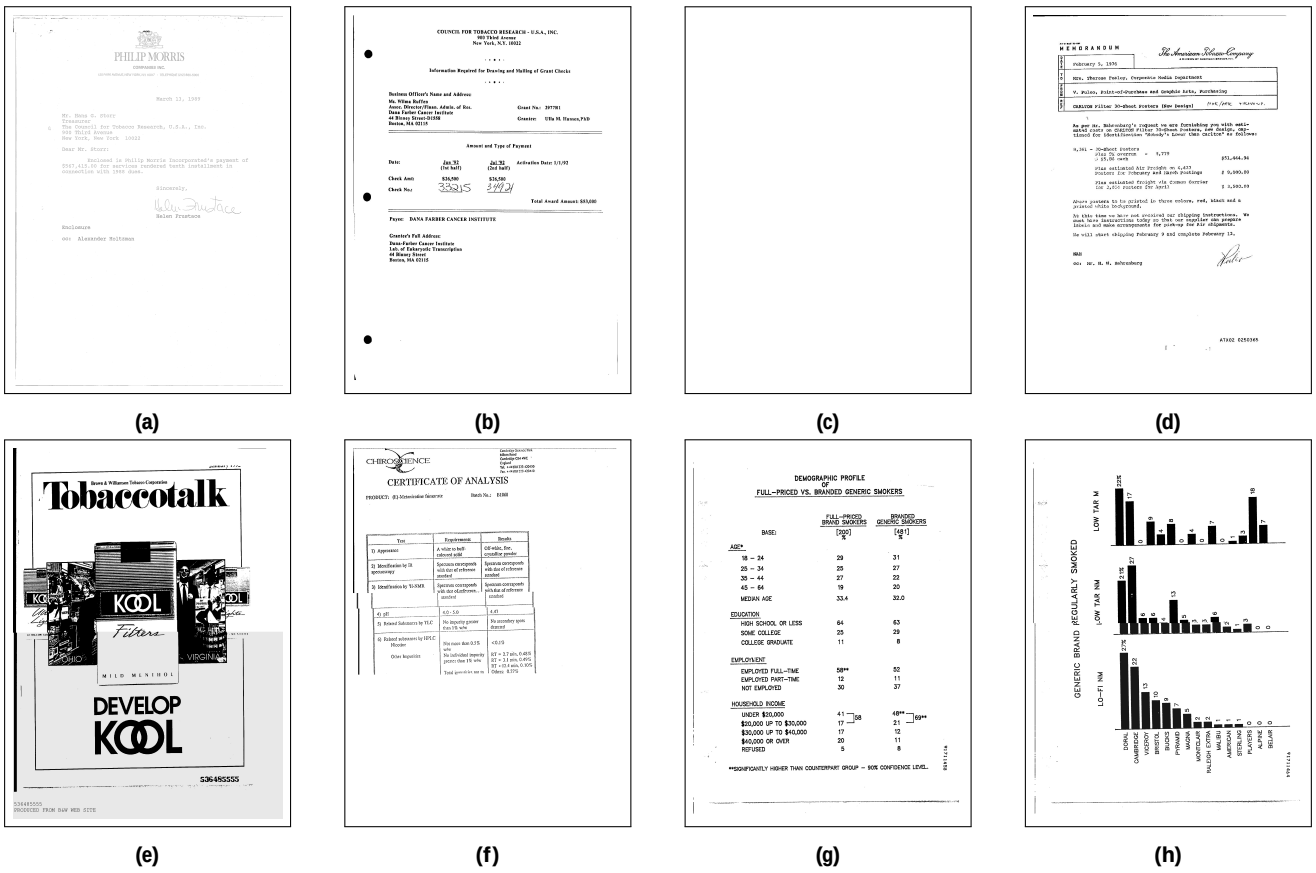


Figure 1: Visually-rich business documents with di□erent layouts and formats for pre-training DiT.

To this end, we propose **DiT**, a self-supervised pre-trained Document Image Transformer model for general Document AI tasks, which does not rely on any human-labeled document images. Inspired by the recently proposed BEiT model [3], we adopt a similar pre-training strategy using document images. An input text image is □rst resized into 224×224 and then the image is split into a sequence of 16×16 patches which are used as the input to the image Transformer. Distinct from the BEiT model where visual tokens are from the discrete VAE in DALL-E [34], we re-train the discrete VAE (dVAE) model with large-scale document images, so that the generated visual tokens are more domain relevant to the Document AI tasks. The pre-training objective is to recover visual tokens from dVAE based on the corrupted input document images using the Masked Image Modeling (MIM) in BEiT. In this way, the DiT model does not rely on any human-labeled document images, but only leverages large-scale unlabeled data to learn the global patch relationship within each document image. We evaluate the pre-trained DiT models on four publicly available Document AI benchmarks, including the RVL-CDIP dataset [16] for document image classi□cation, the PubLayNet dataset [46] for document layout analysis, the ICDAR 2019 cTDaR dataset [15] for table detection, as well as the FUNSD dataset [22] for OCR text detection. Experiment results have illustrated that the pre-trained DiT model has outperformed the existing supervised and self-supervised pre-trained models and achieved new state-of-the-art on these tasks.

The contributions of this paper are summarized as follows:

(1) We propose DiT, a self-supervised pre-trained document image Transformer model, which can leverage large-scale unlabeled document images for pre-training.
(2) We leverage the pre-trained DiT models as the backbone for a variety of Document AI tasks, including document image classi□cation, document layout analysis, table detection, as well as text detection for OCR, and achieve new state-of-the-art results.
(3) The code and pre-trained models are publicly available at https://aka.ms/msdit.

2 RELATED WORK

Image Transformer has recently achieved signi□cant progress in computer vision problems, including classi□cation, object detection, and segmentation. [12] □rst applied a standard Transformer directly to images with the fewest modi□cations. They split an image into 16×16 patches and provide the sequence of linear embeddings of these patches as an input to a Transformer named ViT. The ViT model is trained on image classi□cation in a supervised fashion and outperforms the ResNet baselines. [36] proposed data-e□cient

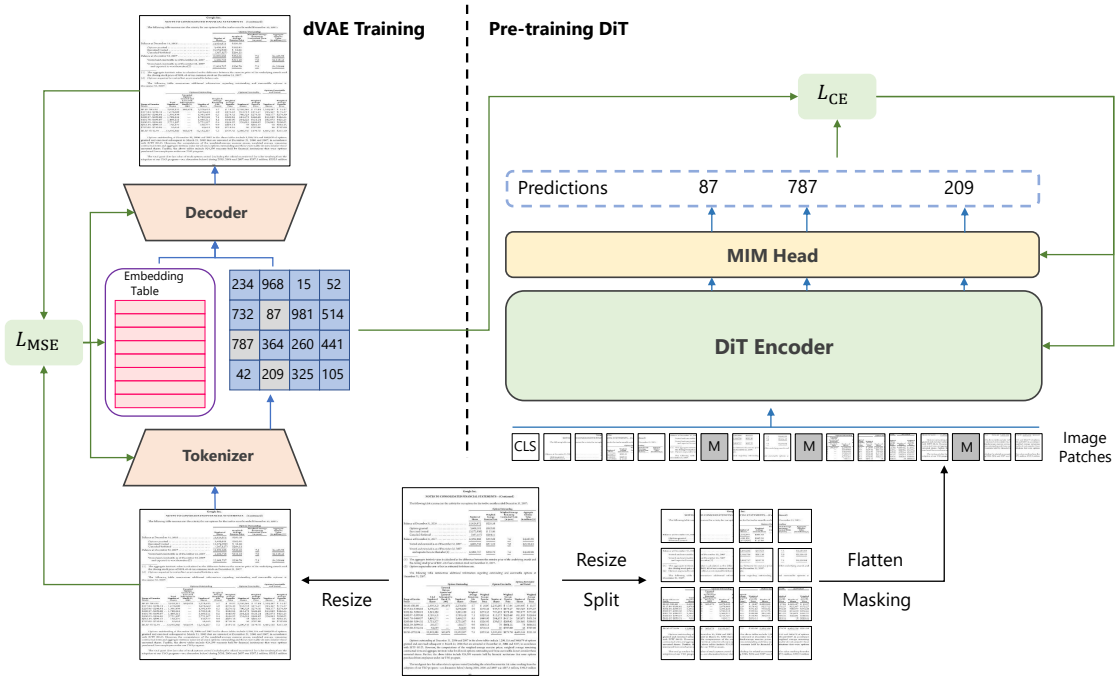


Figure 2: The model architecture of DiT with MIM pre-training.

image transformers & distillation through attention, namely DeiT, which solely relies on the ImageNet dataset for supervised pre-training and achieves SOTA results compared with ViT. [31] proposed a hierarchical Transformer whose representation is computed with shifted windows. The shifted windowing scheme brings e□-ciency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. In addition to supervised pre-trained models, [8] trained a sequence Transformer called iGPT to auto-regressively predict pixels without incorporating knowledge of the 2D input structure, which is the □rst attempt at self-supervised image transformer pre-training. After that, self-supervised pre-training for image Transformer became a hot topic in computer vision. [7] proposed DINO, which pre-trains the image Transformer using self-distillation with no labels. [9] proposed MoCov3 that is based on Siamese networks for self-supervised learning. More recently, [3] adopted a BERT-style pre-training strategy, which □rst tokenizes the original image into visual tokens, then randomly masks some image patches and feeds them into the backbone Transformer. Similar to the masked language modeling, they proposed a masked image modeling task as the pre-training objective that achieves SOTA performance. [47] presented a self-supervised framework iBOT that can perform masked prediction with an online tokenizer. The online tokenizer is jointly learnable with the MIM objective and dispenses with a multi-stage pipeline where the tokenizer is pre-trained beforehand.

The vision-based Document AI usually denote document analysis tasks that leverage the computer vision models, such as OCR, document layout analysis, and document image classi□cation. Due to the lack of large-scale human-labeled datasets in this domain, existing approaches are usually based on the ConvNets models that are pre-trained with ImageNet/COCO datasets. Then, the models are continuously trained with task-speci□c labeled samples. To the best of our knowledge, the pre-trained DiT model is the □rst large-scale self-supervised pre-trained model for vision-based Document AI tasks. Meanwhile, it can be further leveraged for the multimodal pre-training for Document AI.

3 DOCUMENT IMAGE TRANSFORMER

In this section, we □rst present the architecture of DiT and the pre-training procedure. Then, we describe the application of DiT models in di□erent downstream tasks.

3.1 Model Architecture

Following ViT [12], we use the vanilla Transformer architecture [37] as the backbone of DiT. We divide a document image into non-overlapping patches and obtain a sequence of patch embeddings. After adding the 1d position embedding, these image patches are passed into a stack of Transformer blocks with multi-head attention. Finally, we take the output of the Transformer encoder as the representation of image patches, which is shown in Figure 2.

3.2 Pre-training

Inspired by BEiT [3], we use Masked Image Modeling (MIM) as our pre-training objective. In this procedure, the images are represented as image patches and visual tokens in two views respectively. During pre-training, DiT accepts the image patches as input and predicts the visual tokens with the output representation.

Like text tokens in natural language, an image can be represented as a sequence of discrete tokens obtained by an image tokenizer. BEiT uses the discrete variational auto-encoder (dVAE) from DALL-E [34] as the image tokenizer, which is trained on a large data collection including 400 million images. However, there exists a domain mismatch between natural images and document images, which makes the DALL-E tokenizer not appropriate for the document images. Therefore, to get better discrete visual tokens for the document image domain, we train a dVAE on the IIT-CDIP [24] dataset that includes 42 million document images.

To effectively pre-train the DiT model, we randomly mask a subset of inputs with a special token `[MASK]` given a sequence of image patches. The DiT encoder embeds the masked patch sequence by a linear projection with added positional embeddings, and then contextualizes it with a stack of Transformer blocks. The model is required to predict the index of visual tokens with the output from masked positions. Instead of predicting the raw pixels, the masked image modeling task requires the model to predict the discrete visual tokens obtained by the image tokenizer.

3.3 Fine-tuning

We fine-tune our model on four Document AI benchmarks, including the RVL-CDIP dataset for document image classification, the PubLayNet dataset for document layout analysis, the ICDAR 2019 cTDaR dataset for table detection, and the FUNSD dataset for text detection. These benchmark datasets can be formalized as two common tasks: image classification and object detection.

Image Classification. For image classification, we use average pooling to aggregate the representation of image patches. Next, we pass the global representation into a simple linear classifier.

Object Detection. For object detection, as in Figure 3, we leverage Mask R-CNN [18] and Cascade R-CNN [5] as detection frameworks and use ViT-based models as the backbone. Our code is implemented based on Detectron2 [39]. Following [14, 29], we use resolution-modifying modules at four different transformer blocks to adapt the single-scale ViT to the multi-scale FPN. Let d be the total number of blocks, the $1d/3$th block is upsampled by $4\times$ using a module with 2 stride-two 2×2 transposed convolution. For the output of the $1d/2$th block, we use a single stride-two 2×2 transposed convolution to upsample $2\times$. The output of the $2d/3$th block is utilized without additional operations. Finally, the output of $3d/3$th block is downsampled by $2\times$ with stride-two 2×2 max pooling.

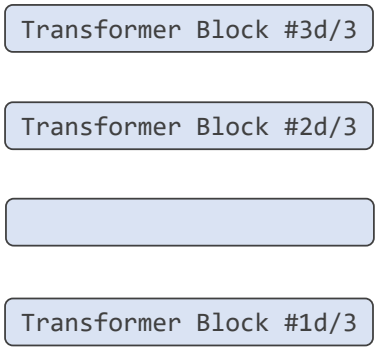


Figure 3: Illustration of applying DiT as the backbone network in different detection frameworks.

4 EXPERIMENTS

4.1 Tasks

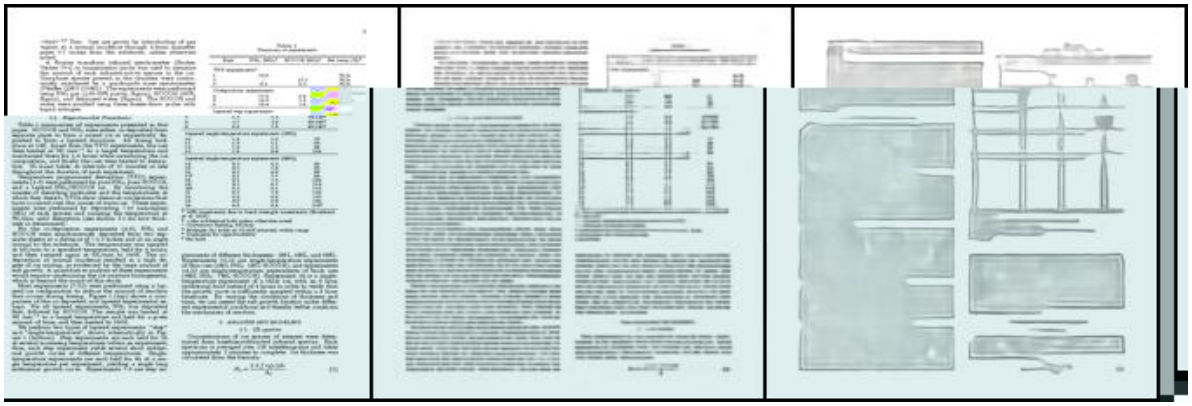
We briefly introduce the datasets mentioned in section 3.3 here.

RVL-CDIP. The RVL-CDIP [16] dataset consists of 400,000 grayscale images in 16 classes, with 25,000 per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images. The 16 classes include {letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo}. The evaluation metric is the overall classification accuracy.

PubLayNet. PubLayNet [46] is a large-scale document layout analysis dataset. More than 360,000 document images are constructed by automatically parsing PubMed XML files. The resulting annotations cover typical document layout elements such as text, title, list, figure, and table. The model needs to detect the regions of the assigned elements. We use the category-wise and overall mean average precision (MAP) @ intersection over union (IOU) [0.50:0.95] of bounding boxes as the evaluation metrics.

ICDAR 2019 cTDaR. The cTDaR datasets [15] consist of two tracks, including table detection and table structure recognition. In this paper, we focus on Track A where document images with one or several table annotations are provided. This dataset has two subsets, one for archival documents and the other for modern documents. The archival subset includes 600 training images and 199 testing images, which shows a wide variety of tables containing hand-drawn accounting books, stock exchange lists, train timetables, production census, etc. The modern subset consists of 600

(a) A sample from the PubLayNet dataset



(b) A sample from the ICDAR 2019 cTDaR dataset

Figure 4: Document image reconstruction with di□erent tokenizers. From left to right: the original document image, image reconstruction using the self-trained dVAE tokenizer, image reconstruction using the DALL-E tokenizer.

training images and 240 testing images, which contain di□erent kinds of PDF □les, such as scienti□c journals, forms, □nancial statements, etc. The dataset contains Chinese and English documents in various formats, including scanned document images and born-digital formats. Metrics for evaluating this task are the precision, recall, and F1 scores computed from the model's ranked output w.r.t. di□erent Intersection over Union (IoU) threshold. We calculate the values with IoU thresholds of 0.6, 0.7, 0.8, and 0.9 respectively, and merge them into a □nal weighted F1 score:

$$wF1 = \frac{0.6F1_{0.6} + 0.7F1_{0.7} + 0.8F1_{0.8} + 0.9F1_{0.9}}{0.6 + 0.7 + 0.8 + 0.9}$$

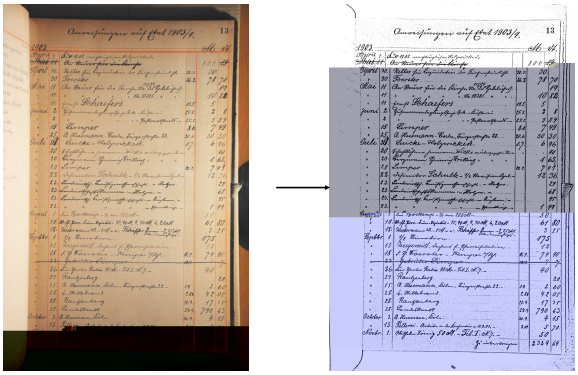
This task further requires models to combine the modern and archival set as a whole to get a □nal evaluation result.

FUNSD. FUNSD [22] is a noisy scanned document dataset labeled for three tasks: Text detection, Text recognition with Optical Character Recognition (OCR), and Form understanding. In this paper, we focus on Task #1 in FUNSD, which aims to detect the text bounding boxes for scanned form documents. FUNSD includes 199 fully annotated forms with 31,485 words, whereas the training set contains 150 forms and the testing set includes 49 forms. The evaluation metrics are the precision, recall, and F1 score at IoU@0.5.

4.2 Settings

Pre-training Setup. We pre-train DiT on the IIT-CDIP Test Collection 1.0 [24]. We pre-process the dataset by splitting multi-page documents into single pages, and obtain 42 million document images. We also introduce random resized cropping to augment training data during training. We train our DiT-B model with the same architecture as the ViT base: a 12-layer Transformer with 768 hidden sizes, and 12 attention heads. The intermediate size of feed-forward networks is 3,072. A larger version, DiT-L, is also trained with 24 layers, 1,024 hidden sizes, and 16 attention heads. The intermediate size of feed-forward networks is 4,096.

The dVAE Tokenizer. BEiT borrows the image tokenizer trained by DALL-E, which is not aligned with the document image data. In this case, we fully utilize the 42 million document images in the IIT-CDIP dataset and train a document dVAE image tokenizer to obtain the visual tokens. Like the DALL-E image tokenizer, the document image tokenizer has the codebook dimensionality of 8,192 and the image encoder with three layers. Each layer consists of a 2D convolution with a stride of 2 and a ResNet block. Therefore, the tokenizer eventually has a downsampling factor of 8. In this



Raw Document Image Adaptively Binarized Image

Figure 5: An example of pre-processing with adaptive image binarization on the ICDAR 2019 cTDaR archival subset.

Model	Type	Accuracy	#Param
[1]	Single	90.97	-
[11]	Single	91.11	-
[11]	Ensemble	92.21	-
[35]	Ensemble	92.77	-
ResNext-101-32×8d	Single	90.65	88M
DeiT-B [36]	Single	90.32	87M
BEiT-B [3]	Single	91.09	87M
MAE-B [17]	Single	91.42	87M
DiT-B	Single	92.11	87M
DiT-L	Single	**92.69**	304M

Table 1: Document Image Classification accuracy (%) on RVL-CDIP, where all the models use the pure image information (w/o text information) with the 224×224 resolution.

Model	Text	Title	List	Table	Figure	Overall
[46]	0.916	0.840	0.886	0.960	0.949	0.910
ResNext	0.916	0.845	0.918	0.971	0.952	0.920
DeiT-B	0.934	0.874	0.921	0.972	0.957	0.932
BEiT-B	0.934	0.866	0.924	0.973	0.957	0.931
MAE-B	0.933	0.865	0.918	0.973	0.959	0.930
DiT-B	0.934	0.871	0.929	0.973	0.967	0.935
DiT-L	0.937	0.879	0.945	0.974	0.968	0.941
ResNext (C)	0.930	0.862	0.940	0.976	0.968	0.935
DiT-B (C)	0.944	0.889	0.948	0.976	0.969	0.945
DiT-L (C)	**0.944**	**0.893**	**0.960**	**0.978**	**0.972**	**0.949**

Table 2: Document Layout Analysis mAP @ IOU [0.50:0.95] on PubLayNet validation set. ResNext-101-32×8d is shortened as ResNext and Cascade as C.

case, given a 112×112 image, it ends up with a 14×14 discrete token map aligning with the 14×14 input patches.

We implement our dVAE codebase from open-sourced DALL-E implementation[1] and train the dVAE model with the entire IIT-CDIP dataset containing 42 million document images. The new dVAE tokenizer is trained with a combination of a MSE loss to reconstructe the input image, and a perplexity loss to increase the use of the quantized codebook representations. The input image size is 224×224, and we train the tokenizer with a learning rate of 5e-4 and a minimum temperature of 1e-10 for 3 epochs. We compare our dVAE tokenizer with the original DALL-E tokenizer by reconstructing the document image samples from downstream tasks, which is shown in Figure 4. We sample images from the document layout analysis dataset PubLayNet and table detection dataset ICDAR 2019 cTDaR. After being reconstructed by the DALL-E and our tokenizer, the image tokenizer by DALL-E is hard to distinguish the border of lines and tokens, but the image tokenizer by our dVAE is closer to the original image and the border is sharper and clearer. We confirm that a better tokenizer can produce more accurate tokens that better describe the original images.

Equipped with the pre-training data and image tokenizer, we pre-train DiT for 500K steps with a batch size of 2,048, a learning rate of 1e-3, warmup steps of 10K, and weight decay of 0.05. The β_1 and β_2 of Adam [23] optimizer are 0.9 and 0.999 respectively. We employ stochastic depth [20] with a 0.1 rate and disable dropout as in BEiT pre-training. We also apply blockwise masking in the pre-training of DiT with 40% patches masked as BEiT.

Fine-tuning on RVL-CDIP. We evaluate the pre-trained DiT models and other image backbones on RVL-CDIP for document image classification. We fine-tune the image transformers for 90 epochs with a batch size of 128 and a learning rate of 1e-3. For all settings, we resize the original images to 224 × 224 with the RandomResizedCrop operation.

Fine-tuning on ICDAR 2019 cTDaR. We evaluate the pre-trained DiT models and other image backbones on the ICDAR 2019 dataset for table detection. Since the image resolution for object detection tasks is much larger than classification, we limit the batch size to 16. The learning rate is 1e-4 and 5e-5 for archival and modern subsets respectively. In the preliminary experiments, we found that directly using the raw images in the archival subset leads to suboptimal performance when fine-tuning DiT, so we apply an adaptive image binarization algorithm implemented by OpenCV [4] to binarize the images. An example of the pre-procession is shown in Figure 5. During training, we apply the data augmentation method used in DETR [6] as a multi-scale training strategy. Specifically, the input image is cropped with probability 0.5 to a random rectangular patch which is then resized again such that the shortest side is at least 480 and at most 800 pixels while the longest at most 1,333.

Fine-tuning on PubLayNet. We evaluate the pre-trained DiT models and other image backbones on the PubLayNet dataset for document layout analysis. Similar to the ICDAR 2019 cTDaR dataset, the batch size is 16, and the learning rate is 4e-4 for the base version. and 1e-4 for the large version. The data augmentation method for DETR [6] is also used.

[1] https://github.com/lucidrains/DALLE-pytorch

Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	96.97	95.99	95.14	90.22	94.23
ResNeXt-101-32×8d	96.42	95.99	95.15	91.36	94.46
DeiT-B	96.26	95.56	94.57	90.91	94.04
BEiT-B	96.82	96.40	95.41	92.44	95.03
MAE-B	96.86	96.31	95.05	91.57	94.66
DiT-B	96.75	96.19	95.62	93.36	95.30
DiT-L	**97.83**	**97.41**	96.29	92.93	95.85
ResNeXt-101-32×8d (Cascade)	96.54	95.84	95.13	92.87	94.90
DiT-B (Cascade)	97.20	96.92	96.78	94.26	96.14
DiT-L (Cascade)	97.68	97.26	**97.12**	**94.74**	**96.55**

(a) Table detection accuracy on ICDAR 2019 cTDaR (combined: archival+modern)

Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	97.16	96.41	95.27	91.12	94.67
ResNeXt-101-32×8d	96.60	96.60	95.09	91.70	94.73
DeiT-B	97.54	97.16	96.41	92.63	95.68
BEiT-B	**98.10**	**98.10**	95.82	94.30	96.35
MAE-B	97.54	97.54	96.03	94.14	96.12
DiT-B	97.53	97.15	96.02	94.88	96.24
DiT-L	97.53	97.15	96.39	95.26	96.46
ResNeXt-101-32×8d (Cascade)	96.76	96.38	95.24	93.71	95.35
DiT-B (Cascade)	96.97	96.97	96.97	95.83	96.63
DiT-L (Cascade)	97.34	97.34	**97.34**	**96.20**	**97.00**

(b) Table detection accuracy on ICDAR 2019 cTDaR (archival)

Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	96.86	95.74	95.07	89.69	93.97
ResNeXt-101-32×8d	96.30	95.63	95.18	91.15	94.30
DeiT-B	95.51	94.61	93.48	89.89	93.07
BEiT-B	96.06	95.39	95.16	91.34	94.25
MAE-B	96.47	95.58	94.48	90.07	93.81
DiT-B	96.29	95.61	95.39	92.46	94.74
DiT-L	**98.00**	**97.56**	96.23	91.57	95.50
ResNeXt-101-32×8d (Cascade)	96.41	95.52	95.07	92.38	94.63
DiT-B (Cascade)	97.33	96.89	96.67	93.33	95.85
DiT-L (Cascade)	97.89	97.22	**97.00**	**93.88**	**96.29**

(c) Table detection accuracy on ICDAR 2019 cTDaR (modern)

Table 3: Table detection accuracy (F1) on ICDAR 2019 cTDaR.

Fine-tuning on FUNSD. We use the same object detection framework for fine-tuning the pre-trained DiT models and other backbones on the text detection task in FUNSD. In document layout analysis and table detection, we use anchor box sizes [32, 64, 128, 256, 512] in the detection process since the detected areas are usually paragraph-level. Different from document layout analysis, text detection aims to locate smaller objects at the word level in document images. Therefore, we use anchor box sizes [4, 8, 16, 32, 64] in the detection process. The batch size is set to 16 and the learning rate is 1e-4 for the base model and 5e-5 for the large model.

The image backbone models selected as baselines have a comparable number of parameters compared with our DiT-B. They include the following two kinds: CNN and image Transformer. For CNN-based models, we choose ResNext101-32×8d [40]. For image Transformers, we choose the base version of DeiT [36], BEiT [3] and MAE [17] which are pre-trained on ImageNet-1K dataset with a 224×224 input size. We rerun the fine-tuning of all baselines.

4.3 Results

RVL-CDIP. The results of document image classification on RVL-CDIP are shown in Table 1. To make a fair comparison, the approaches in the table use only image information from the dataset. DiT-B performs significantly better than all selected single-model baselines. Since DiT shares the same model structure with other image Transformer baselines, the higher score indicates the effectiveness of our document-specific pre-training strategy. The larger version, DiT-L, gets a comparable score with the previous SOTA ensemble model under the single-model setting, which further highlights its modeling capability on document images.

PubLayNet. The results of document layout analysis on PubLayNet are shown in Table 2. Since this task has a large number of training and testing samples and requires a comprehensive analysis of the common document elements, it clearly demonstrates the learning ability of different image Transformer models. It is observed that the DeiT-B, BEiT-B, and MAE-B are obviously better than ResNeXt-101, and DiT-B is even stronger than these powerful image Transformer baselines. According to the results, the improvement mainly comes from the List and Figure category, and on the basis of DiT-B, DiT-L gives out a much higher mAP score. We also investigate the impact of different object detection algorithms, and the results show that a more advanced detection algorithm (Cascade R-CNN in our case) can push the model performance to a higher level. We also apply Cascade R-CNN on the ResNeXt-101-32×8d baseline, and DiT surpasses it by 1% and 1.4% absolute score for the base and large settings respectively, indicating the superiority of DiT on a different detection framework.

ICDAR 2019 cTDaR. The results of table detection on ICDAR 2019 cTDaR dataset are shown in Table 3. The size of this dataset is relatively small, so it aims at evaluating the few-shot learning capability of models under a low-resource scenario. We first analyze the model performance on the archival and modern subsets separately. In Table 3b, DiT surpasses all the baselines except BEiT for the archival subset. This is because in the pre-training of BEiT, it directly uses the DALL-E dVAE which is trained on an extremely large dataset with 400M images with different colors. While for DiT, the image tokenizer is trained with grayscale images, which may not be sufficient for historical document images with colors. The improvement when switching from Mask R-CNN to Cascade R-CNN is also observed which is similar to PubLayNet settings, and DiT still outperforms other baselines significantly. The conclusion is similar to the results on the modern subset in Table 3c. We further combine the predictions of the two subsets into a single set. The results in 3a show DiT-L achieves the highest wF1 score among all Mask R-CNN methods, demonstrating the versatility of DiT under different categories of documents. It is worth noting that the metrics of IoU@0.9 are significantly better, which means DiT has a better fine-grained object detection capability. Under all the three settings, we have pushed the SOTA results to a new level by more than 2% (94.23→96.55) absolute wF1 score with our best model and the Cascade R-CNN detection algorithm.

FUNSD (Text Detection). The results of text detection on the FUNSD dataset are shown in Table 4. Since text detection for OCR has been a long-standing real-world problem, we obtain the word-level text detection results from a popular commercial OCR engine to set a high-level baseline. In addition, DBNet [30] is a widely used text detection model for online OCR engines, we also fine-tune a pre-trained DBNet model with FUNSD training data and evaluate its accuracy. Both of them achieve around 0.85 F1 scores for IoU@0.5. Next, we use the Mask R-CNN framework to compare different backbone networks (CNN and ViT) including ResNeXt-101, DeiT, BEiT, MAE, and DiT. It is shown that CNN-based and ViT-based text detection models outperform the baselines significantly due to advanced model design and more parameters. We also observe that the DiT models achieve new SOTA results compared with other models. Finally, we further train the DiT models with a synthetic dataset that contains 1 million document images, leading to an F1 of 0.9429 being achieved by the DiT-L model.

Model	Precision	Recall	F1
Faster R-CNN [22]	0.704	0.848	0.76
DBNet [30]	0.8764	0.8400	0.8578
A Commercial OCR Engine	0.8762	0.8260	0.8504
ResNeXt-101-32×8d	0.9387	0.9229	0.9307
DeiT-B	0.9429	0.9237	0.9332
BEiT-B	0.9412	0.9263	0.9337
MAE-B	0.9441	0.9321	0.9381
DiT-B	0.9470	0.9307	0.9388
DiT-L	0.9452	**0.9336**	0.9393
DiT-B (+syn)	0.9539	0.9315	0.9425
DiT-L (+syn)	**0.9543**	0.9317	**0.9429**

Table 4: Text detection accuracy (IoU@0.5) on FUNSD Task #1, where Mask R-CNN is used with different backbones (ResNeXt, DeiT, BEiT, MAE and DiT). "+syn" denotes that DiT is trained with a synthetic dataset including 1M document images, then fine-tuned with the FUNSD training data.

5 CONCLUSION AND FUTURE WORK

In this paper, we present DiT, a self-supervised foundation model for general Document AI tasks. The DiT model is pre-trained with large-scale unlabeled document images that cover a variety of templates and formats, which is ideal for downstream Document AI tasks in different domains. We evaluate the pre-trained DiT on several vision-based Document AI benchmarks, including table detection, document layout analysis, document image classification, and text detection. Experimental results have shown that DiT outperforms several strong baselines across the board and achieves new SOTA performance. We will make the pre-trained DiT models publicly available to facilitate the Document AI research.

For future research, we will pre-train DiT with a much larger dataset to further push the SOTA results in Document AI. Meanwhile, we will also integrate DiT as the foundation model in multi-modal pre-training for visually-rich document understanding such as the next-gen layout-based models like LayoutLM, where a unified Transformer-based architecture may be sufficient for both CV and NLP applications in Document AI.

REFERENCES

[1] Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. 2017. Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classi□cation. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), 883–888.

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-End Transformer for Document Understanding. arXiv:2106.11539 [cs.CV]

[3] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 [cs.CV]

[4] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[5] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 6154–6162. https://doi.org/10.1109/CVPR.2018.00644

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294 [cs.CV]

[8] Mark Chen, Alec Radford, Rewon Child, Je□rey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1691–1703. http://proceedings.mlr.press/v119/chen20s.html

[9] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. arXiv:2104.02057 [cs.CV]

[10] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, Models and Applications. arXiv:2111.08609 [cs.CL]

[11] Arindam Das, Saikat Roy, and Ujjwal Bhattacharya. 2018. Document Image Classi□cation with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks. *2018 24th International Conference on Pattern Recognition (ICPR)* (2018), 3180–3185.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[13] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. 2021. XCiT: Cross-Covariance Image Transformers. arXiv:2106.09681 [cs.CV]

[14] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. 2021. XCiT: Cross-Covariance Image Transformers. *ArXiv* abs/2106.09681 (2021).

[15] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 1510–1515. https://doi.org/10.1109/ICDAR.2019.00243

[16] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of Deep Convolutional Nets for Document Image Classi□cation and Retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV]

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

[19] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. BROS: A Pre-trained Language Model for Understanding Texts in Document. https://openreview.net/forum?id=punMXQEsPr0

[20] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. 2016. Deep Networks with Stochastic Depth. In *ECCV*.

[21] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Uni□ed Text and Image Masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisbon,Portugal, October 10-14, 2022*.

[22] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* 2 (2019), 1–6.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[24] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a Test Collection for Complex Document Information Processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. ACM, New York, NY, USA, 665–666. https://doi.org/10.1145/1148170.1148307

[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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6309–6318. https://doi.org/10.18653/v1/2021.acl-long.493

[26] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. TableBank: Table Benchmark for Image-based Table Detection and Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1918–1925. https://aclanthology.org/2020.lrec-1.236

[27] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A Benchmark Dataset for Document Layout Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 949–960. https://doi.org/10.18653/v1/2020.coling-main.82

[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. arXiv:2106.03331 [cs.CV]

[29] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollár, Kaiming He, and Ross B. Girshick. 2021. Benchmarking Detection Transfer Learning with Vision Transformers. *ArXiv* abs/2111.11429 (2021).

[30] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-Time Scene Text Detection with Di□erentiable Binarization and Adaptive Scale Fusion. arXiv:2202.10304 [cs.CV]

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]

[32] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. arXiv:2102.09550 [cs.CL]

[33] Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. 2020. Towards a Multi-modal, Multi-task Learning based Pre-training Framework for Document Representation Learning. arXiv:2009.14457 [cs.CL]

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]

[35] Ritesh Sarkhel and Arnab Nandi. 2019. Deterministic Routing between Layout Abstractions for Multi-Scale Classi□cation of Visually Rich Documents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 3360–3366. https://doi.org/10.24963/ijcai.2019/466

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-e□cient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[38] 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:2104.08405 [cs.CL]

[39] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[40] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 5987–5995. https://doi.org/10.1109/CVPR.2017.634

[41] 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 '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 1192–1200. https://dl.acm.org/doi/10.1145/3394486.3403172

[42] 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:2104.08836 [cs.CL]

[43] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 3214–3224. https://doi.org/10.18653/v1/2022.□ndings-acl.253

[44] 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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

[45] Xu Zhong, Elaheh Sha□eiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. arXiv:1911.10683 [cs.CV]

[46] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166

[47] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. iBOT: Image BERT Pre-Training with Online Tokenizer. arXiv:2111.07832 [cs.CV]

[48] Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. LAMBERT: Layout-Aware (Language) Modeling for information extraction. arXiv:2002.08087 [cs.CL]