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

Junlong Li[1]*, **Yiheng Xu**[2]*, **Tengchao Lv**[2], **Lei Cui**[2], **Cha Zhang**[3], **Furu Wei**[2]
[1]Shanghai Jiao Tong University
[2]Microsoft Research
[3]Microsoft Azure AI
[1]lockonn@sjtu.edu.cn
[2,3]{t-yihengxu,tengchaolv,lecu,chazhang,fuwei}@microsoft.com

DIT → Backbone network for many document image tasks

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, as well as table detection. 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) and table detection (94.23 → 96.55). The code and pre-trained models are publicly available at https://aka.ms/msdit.

1 Introduction

Self-supervised pre-training techniques have been the de facto common practice for Document AI (Cui et al., 2021) in the past several years, where the image, text and layout information is often jointly trained using a unified Transformer architecture (Xu et al., 2020; 2021a;b; Pramanik et al., 2020; Łukasz Garncarek et al., 2021; Hong et al., 2021; Powalski et al., 2021; Wu et al., 2021; Li et al., 2021a;b; Appalaraju et al., 2021). Among all these approaches, a typical pipeline for pre-training Document AI models usually starts 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 label training samples. Although good results have been achieved on benchmark datasets, these vision models are often confronted with the performance gap in the real-world applications due to domain shift and template/format mismatch from the training data. Such accuracy regression (Li et al., 2020a; Zhong et al., 2019) 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 (Dosovitskiy et al., 2021; Touvron et al., 2021; Liu et al., 2021; Chen et al., 2021; Bao et al., 2021; El-Nouby et al., 2021b; He et al., 2021; Zhou et al., 2021) 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 the 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 (Zhong et al., 2019; 2020; Li et al., 2020a;b),

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

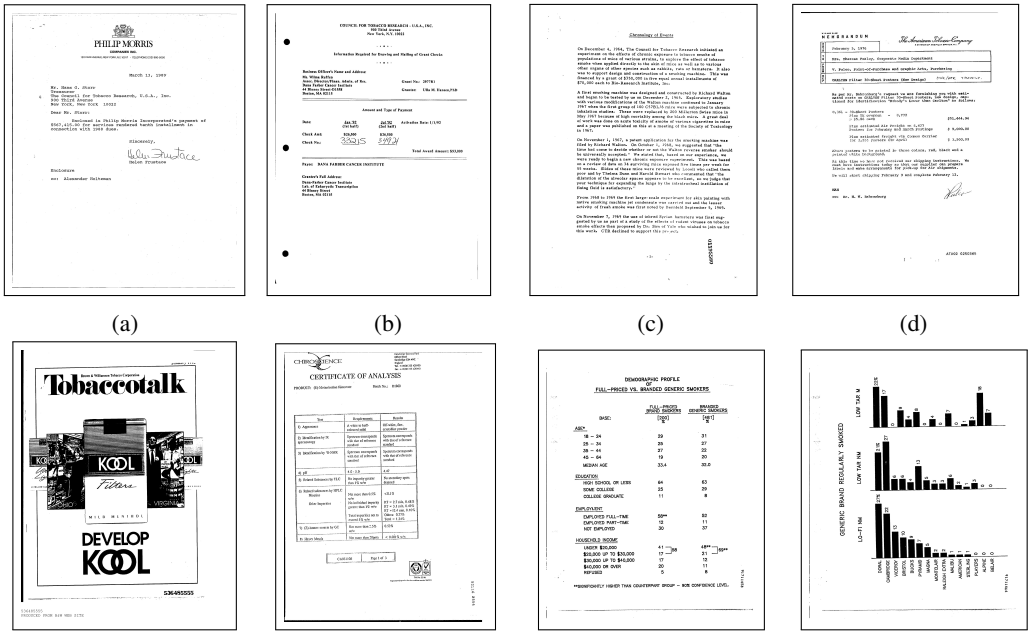


Figure 1: Visually-rich business documents with different layouts and formats for pre-training DiT.

the domain of these datasets is often from the academic papers that share similar templates and formats, which are different from the 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.

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 (Bao et al., 2021), we adopt a similar pre-training strategy using document images. An input text image is first 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 (Ramesh et al., 2021), we re-train the discrete VAE (dVAE) model with large-scale document images, so that the generated visual tokens is 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 leverage large-scale unlabeled data to learn the global patch relationship within each document image. We evaluate the pre-trained DiT models on three publicly available Document AI benchmarks, including the RVL-CDIP dataset (Harley et al., 2015) for document image classification, the PubLayNet dataset (Zhong et al., 2019) for document layout analysis, as well as the ICDAR 2019 cTDaR dataset (Gao et al., 2019) for table 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.

Input:
224x224 → 16x16 patches

Self supervised paradigm used

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 classification, document layout analysis as well as table detection, and achieve new state-of-the-art results.
3. The code and pre-trained models are publicly available at `https://aka.ms/msdit`.

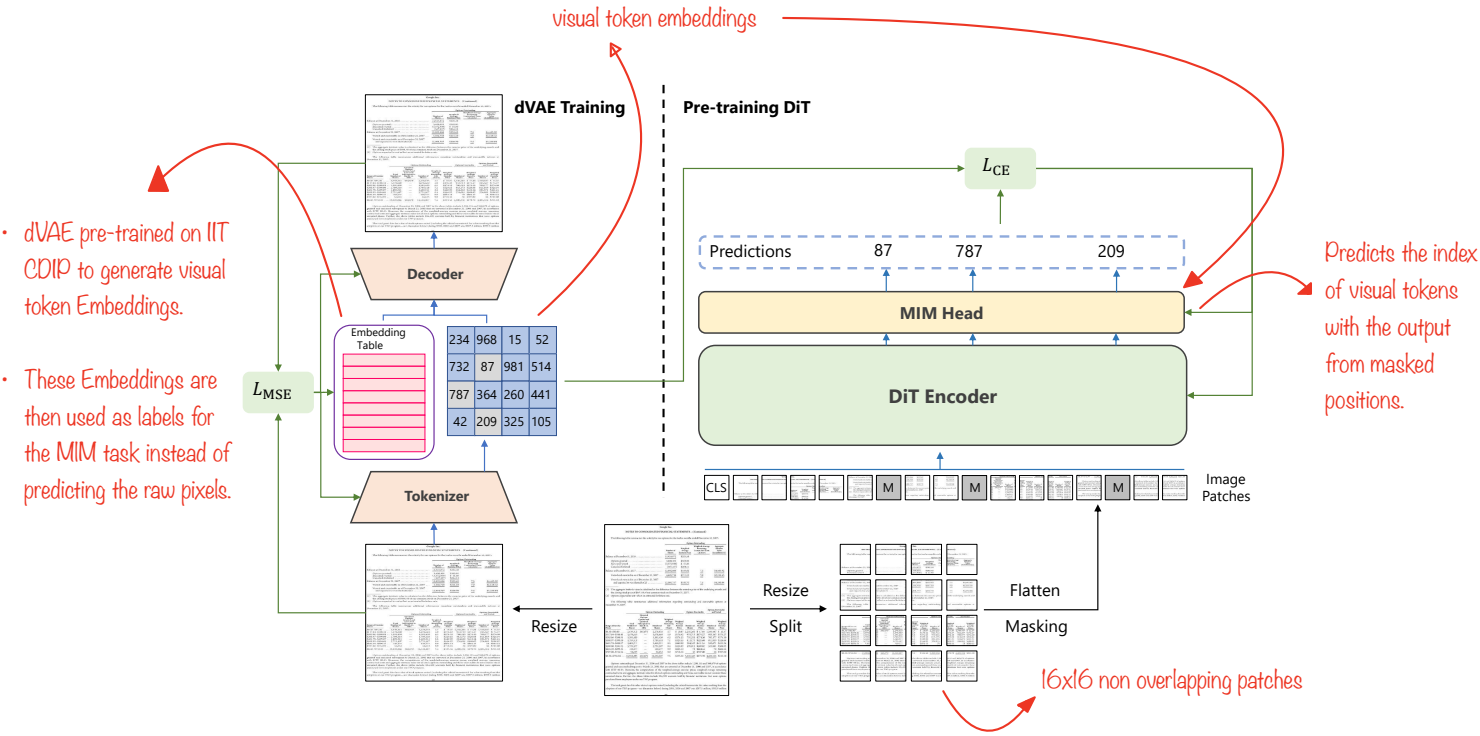


Figure 2: The model architecture of DiT with MIM pre-training.

2 Document Image Transformer

In this section, we first present the architecture of DiT and the pre-training procedure. Then, we describe the application of DiT models in different downstream tasks.

2.1 Model Architecture

Following ViT (Dosovitskiy et al., 2021), we use the vanilla Transformer architecture (Vaswani et al., 2017) 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 the multi-head attention. Finally, we take the output of the Transformer encoder as the representation of image patches, which is shown in Figure 2.

Image → 224x224 → non overlapping 16x16 patches → add 1d position Embeddings → transformer encoder representations of each patch

2.2 Pre-training

Inspired by BEiT (Bao et al., 2021), we use the 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 use the discrete variational auto-encoder (dVAE) from DALL-E (Ramesh et al., 2021) 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 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 (Lewis et al., 2006) 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 image modeling: DiT across image patches as input and predicts visual token output representations. To obtain these representations a dVAE is trained on IIT-CDIP.

Instead of raw pixels

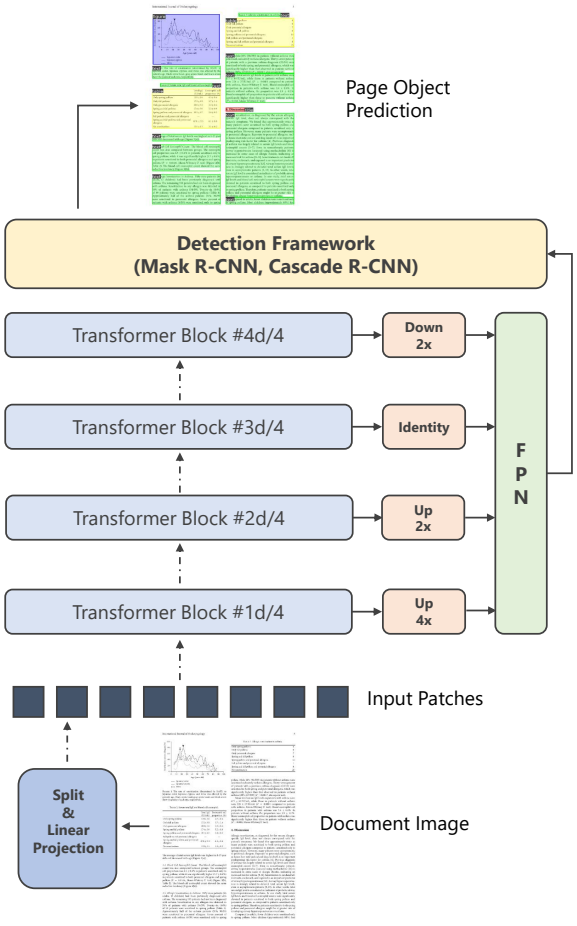


Figure 3: Illustration of applying DiT as the backbone network in different detection frameworks.

masked positions. Instead of predicting the raw pixels, the masked image modeling task requires the model to predict the discrete visual token obtained by the image tokenizer.

2.3 Fine-tuning

To verify the effectiveness of DiT, we fine-tune our model on three Document AI benchmarks, including the RVL-CDIP dataset for document image classification, the PubLayNet dataset for document layout analysis, and the ICDAR 2019 cTDaR dataset for table 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 shown in Figure 3, we leverage Mask R-CNN (He et al., 2017) and Cascade R-CNN (Cai & Vasconcelos, 2018) as detection frameworks and use ViT-based models as the backbone. Our code is implemented based on Detectron2 (Wu et al., 2019). Following (El-Nouby et al., 2021a; Li et al., 2021c), we use four resolution-modifying modules at equally spaced intervals of $d/4$ transformer blocks to adapt the single-scale ViT to the multi-scale FPN, where d is the total number of blocks. The first $1d/4$th block is upsmapled by $4times$ using a module with 2 stride-two 2×2 transposed convolution. For the output of the next $2d/4$th block, we use a single stride-two 2×2 transposed convolution to upsample $2\times$. The next $d/4$th block's output is taken as it is and the final Transformer block's output is downsampled by a factor of two using

stride-two 2×2 max pooling. The output of $3d/4$th block is utilized without additional operation. Finally, the output of $4d/4$th block is downsampled by $2\times$ with stride-two 2×2 max pooling.

3 EXPERIMENTS

3.1 TASKS

We evaluate the performance of our model on three Document AI benchmarks, including the RVL-CDIP dataset for document image classification, the PubLayNet dataset for document layout analysis, and the ICDAR 2019 cTDaR dataset for table detection.

RVL-CDIP The RVL-CDIP (Harley et al., 2015) dataset consists of 400,000 grayscale images in 16 classes, with 25,000 images 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 (Zhong et al., 2019) 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 in a given document image. 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 (Gao et al., 2019) consists of two tracks, including table detection and table structure recognition. In this paper, we focus on the 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 training images and 240 testing images, which contains different kinds of PDF files, such as scientific journals, forms, financial statements, etc. The dataset contains Chinese and English documents with various formats, including scanned document images and born-digital formats. Metrics for evaluating this task are the precision, recall and F1 scores computed from model's ranked output w.r.t. different Intersection over Union (IoU) threshold. We calculate the values with IoU threshold of 0.6, 0.7, 0.8 and 0.9 respectively, and merge them into a final 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 historical set as a whole and calculate all the metrics again to get a final evaluation result.

3.2 SETTINGS

Pre-training Setup We pre-train DiT on the IIT-CDIP Test Collection 1.0 (Lewis et al., 2006). 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 of 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 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

(a) A sample from the PubLayNet dataset

(b) A sample from the ICDAR 2019 cTDaR dataset

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

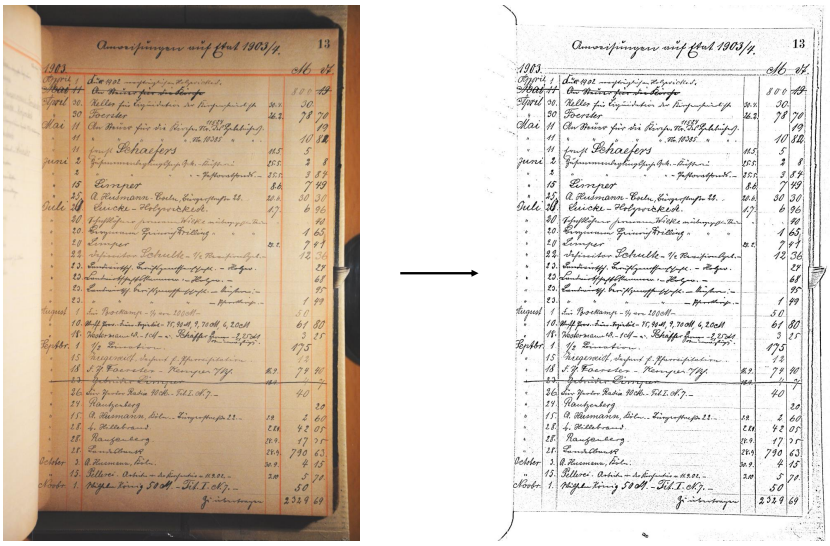
and the image encoder with three layers. Each layer consists of a 2D convolution with the stride of 2 and a ResNet block. Therefore, the tokenizer eventually has a downsampling factor of 8. In this case, give an 112×112 image, it ends up with a 14×14 discrete token map aligning with the 14×14 input patches.

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. 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 tokenizer and our dVAE 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 batch size of 2,048, learning rate of 1e-3, warmup steps of 10K, and weight decay of 0.05. The β_1 and β_2 of Adam (Kingma & Ba, 2015) optimizer are 0.9 and 0.999 respectively. We employ stochastic depth (Huang et al., 2016) with a 0.1 rate and disable dropout as in BEiT pre-training.

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.

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



Raw Document Image Adaptively Binarized Image

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

Fine-tuning on ICDAR 2019 cTDaR We evaluate the pre-trained DiT model and other image backbones on the ICDAR 2019 dataset for table detection. Since the image resolution for object detection tasks are much larger than classification, we limit the batch size to 16. The learning rate is 1e-4 and 5e-5 for historical and modern subsets respectively. In the preliminary experiments, we found that directly using the raw images in the archival subset leads to a suboptimal performance when fine-tuning DiT, so we apply an adaptive image binarization algorithm implemented by OpenCV (Bradski, 2000) 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 (Carion et al., 2020) 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 PubLayNet dataset for document layout analysis. Similar to the ICDAR 2019 cTDaR dataset, the batch size is 16, the learning rate is 4e-4 for base version. and 1e-4 for the large version. The data augmentation method for DETR (Carion et al., 2020) is also used.

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 (Xie et al., 2016). For image Transformers, we choose the base version of DeiT (Touvron et al., 2021), BEiT (Bao et al., 2021) and MAE (He et al., 2021) which are pre-trained on ImageNet-1K dataset with a 224×224 input size. We rerun the fine-tuning of all the baseline models.

3.3 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. We observe that 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.

Model	Type	Accuracy	#Param
(Afzal et al., 2017)	Single	90.97	-
(Das et al., 2018)	Single	91.11	-
(Das et al., 2018)	Ensemble	92.21	-
(Sarkhel & Nandi, 2019)	Ensemble	92.77	-
ResNext-101-32×8d	Single	90.65	88M
DeiT-B (Touvron et al., 2021)	Single	90.32	87M
BEiT-B (Bao et al., 2021)	Single	91.09	87M
MAE-B (He et al., 2021)	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
(Zhong et al., 2019)	0.916	0.840	0.886	0.960	0.949	0.910
ResNext-101-32×8d	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-101-32×8d (Cascade)	0.930	0.862	0.940	0.976	0.968	0.935
DiT-B (Cascade)	0.944	0.889	0.948	0.976	0.969	0.945
DiT-L (Cascade)	**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.

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 on 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 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 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 few-shot learning capability of models under a low-resource scenario. We first analyze the model performance on the archival and modern subset 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 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

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.

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 model.

4 Related Work

Image Transformer has recently achieved significant progress in computer vision problems, including classification, object detection and segmentation. (Dosovitskiy et al., 2021) first applied a standard Transformer directly to images with the fewest modifications. 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 classification in a supervised fashion and outperforms the ResNet baselines. (Touvron et al., 2021) proposed data-efficient 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. (Liu et al., 2021) proposed a hierarchical Transformer whose representation is computed with shifted windows. The shifted windowing scheme brings efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. In addition to supervised pre-trained models, (Chen et al., 2020) trained a sequence Transformer called iGPT to auto-regressively predict pixels without incorporating knowledge of the 2D input structure, which is the first attempt for self-supervised image transformer pre-training. After that, self-supervised pre-training for image Transformer becomes a hot topic in computer vision. (Caron et al., 2021) proposed DINO, which pre-trains the image Transformer using self-distillation with no labels. (Chen et al., 2021) proposed MoCov3 that is based on Siamese networks for self-supervised learning. More recently, (Bao et al., 2021) adopted a BERT-style pre-training strategy, which first tokenizes the original image into visual tokens, then randomly mask some image patches and fed 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. (Zhou et al., 2021) 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 training pipeline where the tokenizer needs to be 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 classification. Due to the lack of large-scale human labeled dataset 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 specific labeled samples. To the best of our knowledge, the pre-trained DiT model is the first 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.

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 and document image classification. 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 LayoutLM models, where a unified Transformer-based architecture may be sufficient for both CV and NLP applications in Document AI.

LayoutLM V3

References

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:883–888, 2017.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers, 2021.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications, 2021.

Arindam Das, Saikat Roy, and Ujjwal Bhattacharya. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3180–3185, 2018.

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. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Xcit: Cross-covariance image transformers. *ArXiv*, abs/2106.09681, 2021a.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. Xcit: Cross-covariance image transformers, 2021b.

Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1510–1515, 2019. doi: 10.1109/ICDAR.2019.00243.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. {BROS}: A pre-trained language model for understanding texts in document, 2021. URL https://openreview.net/forum?id=punMXQEsPr0.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pp. 665–666, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148307. URL http://doi.acm.org/10.1145/1148170.1148307.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding, 2021a.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1918–1925, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.236.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 949–960, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.82. URL https://aclanthology.org/2020.coling-main.82.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning, 2021b.

Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollár, Kaiming He, and Ross B. Girshick. Benchmarking detection transfer learning with vision transformers. *ArXiv*, abs/2111.11429, 2021c.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer, 2021.

Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Ritesh Sarkhel and Arnab Nandi. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3360–3366. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/466. URL https://doi.org/10.24963/ijcai.2019/466.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. Lampret: Layout-aware multimodal pretraining for document understanding, 2021.

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

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2016.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2021a.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403172. URL https://doi.org/10.1145/3394486.3403172.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding, 2021b.

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

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation, 2020.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2021.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: Layout-aware (language) modeling for information extraction, 2021.