MASKED VISUAL-TEXTUAL PREDICTION FOR DOCUMENT IMAGE REPRESENTATION PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we present Masked Visual-Textual Prediction for document image representation pretraining, called MaskDoc. It comprises of two self-supervised pretraining tasks: Masked Image Modeling and Masked Language Modeling, *based on text region-level image masking*. Our approach randomly masks some words or texts and accordingly the corresponding image regions, and the pretraining task is reconstructing the masked image regions as well as the corresponding words. In comparison to masked image modeling which usually predict the image patches or tokens, the encoder pretrained by our approach captures more textual semantics. Compared to the masked multi-modal modeling methods for document image understanding, e.g., LayoutLM and StrucTexT, that need both the image and text inputs, our approach is able to model image-only input, and potentially can deal with more application scenarios free from OCR pre-processing. We demonstrate the effectiveness of MaskDoc on several document image understanding tasks such as image classification, layout analysis, table structure recognition, document OCR, and end-to-end information extraction. Experimental results show that MaskDoc achieves state-of-the-art performance. Our code and models will be released soon.

1 INTRODUCTION

In the Document Artificial Intelligence, how to make the machine understand visually rich document images and extract structured information from them has gradually become a popular research topic. Its main associated tasks include document image classification Harley et al. (2015b), layout analysis Zhong et al. (2019), document OCR (also called text spotting) Li et al. (2017); Lyu et al. (2018), end-to-end information extraction (usually consists of OCR and entity labeling) Jaume et al. (2019), etc. To solve these tasks well, it is necessary to fully exploit both visual and textual cues. Meanwhile, large-scale self-supervised pretraining is a recently rising technical means to enhance multi-modal knowledge learning of document images.

There are two mainstream self-supervised pretraining frameworks for document image understanding. As illustrated in Fig. 1, (a) the first is the masked multi-modal modeling like layoutLM series Xu et al. (2020b;a), UDoc Gu et al. (2021), DocFormer Appalaraju et al. (2021) and StrucTexT Li et al. (2021b), whose inputs mainly consists of OCR results and image embeddings. These methods depend heavily on front-end OCR engines with certain computing costs. What is more, the two models of OCR and document understanding are separately optimized, which cannot guarantee the best end-to-end performance. (b) The second framework is masked image modeling which inherits the concept of vision-based self-supervised learning such as BEiT Bao et al. (2022), MAE He et al. (2022b), CAE Chen et al. (2022), and DiT Li et al. (2022), etc. The input modal is image-only, which is simple yet powerful to learn a visual contextualized representation of document images. Because of the insufficient consideration of textual contents, it is often only used for some vision-dominated tasks like image classification and layout analysis. (c) We propose the third pretraining framework, which learns visual-textual representations using only the images, to optimize the performance of various downstream tasks in an end-to-end manner.

Due to the massive differences between vision and language, the existing methods of document understanding either consider single modality, or introduce a OCR engine to capture textual contents in advance. Researchers used tokens of text as the input in the language pretraining step, such as

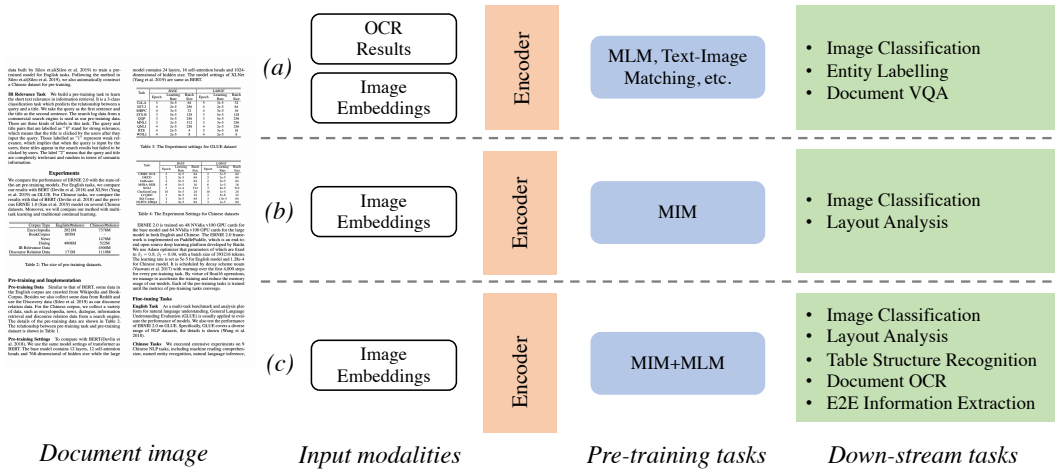

Figure 1: Comparisons with the main-stream pretraining models of document image understanding. (a) shows the masked multi-modal modeling methods which input both OCR results and image embeddings. (b) is the framework that input image embeddings only, suitable for some vision dominated tasks like document image classification and layout analysis. (c) is our method which attempts to learn visual-textual representations using only the information from images in the pretraining phase and then optimizes various downstream tasks of document image understanding end-to-end.

masked language modeling (MLM), while selected predefined image patches as the granularity of vision pretraining tasks (such as BEiT Bao et al. (2022) and MAE He et al. (2022b)). However, the textual contents of document image also originates from its visual information. For example, via a text decoder, the visual features of one text region can convert into the corresponding text. Thus, we perform a text region-level image masking scheme at document images, which can bridge vision modeling to language modeling with the shared representations.

This paper proposes a novel multi-modal knowledge learning framework called MaskDoc, performing text region-based image masking with dual parallel self-supervised tasks of image reconstruction and language modeling, as shown in Fig. 2. First of all, we adopt an off-the-shelf OCR toolkit to perform word-level text detection and text recognition on large public document images such as IIT-CDIP Lewis et al. (2006). Next, we randomly select a set of word regions for image masking, and extract 2D feature maps with the masked image. Then we apply the RoIAlign He et al. (2017) operation on the feature maps to obtain the representations of masked word regions. Finally, the representations are used for reconstructing the image and text of masked word regions.

To take full advantage of the proposed pretraining tasks, we introduce a backbone network for MaskDoc. Inspired by ViBERTGrid Lin et al. (2021), the backbone uses FPN Lin et al. (2017) to integrate visual features of Convolutional Neural Networks (CNN) and semantic features from Transformer. We evaluate and verify our pretrained model in five tasks including document image classification, layout analysis, table structure recognition, document OCR, and end-to-end information extraction, all of which have achieved significant gains. The experimental results confirm that the framework of MaskDoc can produce fundamental pretrained models for document image understanding.

The major contributions of our work can be summarized as following:

- First, we propose a novel self-supervised pretraining framework with performing text region-based document image masking, named MaskDoc. The framework can make the encoder learn visual-textual representations in an end-to-end manner.

- Second, by fine-tuning our pretrained model and achieving SOTA performance on the five downstream tasks, we demonstrate the effectiveness of the proposed model in the field of document image understanding.

2 RELATED WORK

Self-supervised Learning Thanks to the development of self-supervised tasks and the Transformer architectures, the computer vision (CV) fields and natural language processing (NLP) have achieved breakthroughs in knowledge learning from large-scale unlabeled data in the past few years. In the domain of NLP, Masked Language Modeling (MLM) task has been widely used in pretrained models like BERT Devlin et al. (2018) and GPT Radford et al. (2018) due to its efficiency and effectiveness. In detail, the MLM task randomly masks a set of input tokens and reconstructs them according to the context around the masked tokens. In the CV domain, taking a similar idea, Masked Image Modeling (MIM) has also been successfully verified. There are several variants of MIM, for example, BEiT Bao et al. (2022) randomly masks some image patches and predicts the discrete token representations that are learned by dVAE Ramesh et al. (2021). MAE He et al. (2022b) and CAE Chen et al. (2022) take a simpler behavior, predicting the RGB values of masked regions.

Document pretrained models As introduced in the previous section, the cutting-edge document pretraining models can be roughly separated into two categories: masked multi-modal modeling and masked image modeling approaches. The representative works of the former are LayoutLM v1/v2 Xu et al. (2020b;a), Docformer Appalaraju et al. (2021), UDoc Gu et al. (2021) and StrucTexT Li et al. (2021b), they usually need a mature OCR engine to extract textual contents from large-scale unlabeled document images, and fed them to the Transformer-based backbone with image embeddings. These methods can achieve good results, but they often need to be built on a high-performance OCR engine, and they are cumbersome to solve some vision dominated tasks. The latter methods, which directly use a general CV pretraining framework, such as DiT Li et al. (2022), almost follow the idea of BEiT Bao et al. (2022), and learn on large-scale document image data. Due to the lack of guidance for text understanding tasks, the model is likely to be deficient in semantic understanding.

The framework of MaskDoc is a new solution that integrates the advantages of CV and NLP pretraining methods in an end-to-end manner. Since the input modality of the backbone is image-only, our framework can avoid the interference of false OCR results, compared to the first kind of document pretrained models. In addition, although the supervision labels partially come from OCR results, only some high-confidence word-level OCR results need to be randomly selected, so the dependence on the OCR engine is alleviated to a certain extent.

3 APPROACH

3.1 MODEL ARCHITECTURE

Fig. 2 illustrates the overall framework of our MaskDoc, which consists of two main components: a backbone network consist of Convolutional Neural Networks and Vision Transformers and two prediction branches, Mask Language Modeling and Mask Image Modeling.

The proposed backbone networks consist of a visual branch and a context branch. Given an input document image, MaskDoc extract visual-textual representations through this backbone networks. Specifically, the features of the last four down-sampled stages with scales on CNN are extracted from the visual branch. In the context branch, the standard Transformer receives the last stage feature maps of CNN and flatten them into a 1D sequence of token embeddings. Following ViT Dosovitskiy et al. (2020), to handle 2D feature maps, the features of the last stage $x \in R^{h \times w \times C}$ in CNN are firstly flattened 2D patches $x^{'} \in R^{N \times C}$, where (h, w) is the resolution of the input feature maps, C is the number of channels, and $N = h \times w$ is the resulting number of flattened patches, which also serves as the effective input sequence length for the Transformers. Then, the flattened patches are fed to a Transformer Network and enhanced by multiple Multi-head self-attention (MSA) and output enhanced features $x^{''} \in R^{N \times C}$. We reshape the shape of $x^{''}$ back to $h \times w \times C$ and up-sample the context feature map with the factors of $8\times$. We adopt FPN strategy Lin et al. (2017) to merge visual features of different resolutions and then concatenate context features with them, deriving a feature map with 1/4 size of the input image. Finally, two successive 1×1 convolutional layers are performed to take a further fusion.

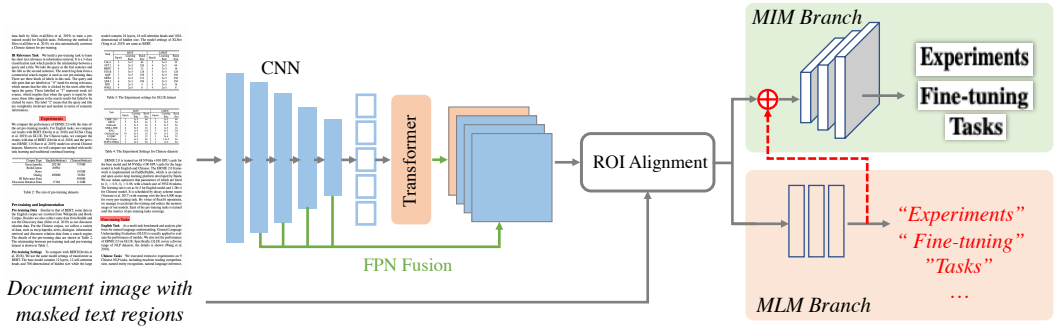

Figure 2: The architecture and dual-modal pretraining tasks of MaskDoc. It consists of a visual branch and a context branch. The features of document image are extracted by CNN and modulated by Transformer. The multi-model features of each text region are fed into MIM branch and MLM branch to reconstructs the text region in pixels and content, respectively.

3.2 PRETRAINING

To enhance the document representation and understanding of MaskDoc proposed above, we perform pretraining on a large-scale document dataset IIT-CDIP Test Collection 1.0 dataset Lewis et al. (2006). Unlike MAE He et al. (2022b), BEiT Bao et al. (2022), DIT Li et al. (2022), we do not use patch level masking strategy for pretraining. Instead, we use a novel text-region level masking strategy and employ two self-supervised pretraining tasks to learn visual-textual representations. The subsequent experiment results suggest that fine-grained text-region level masking strategy is more suitable than coarse-grained patch level masking strategy in the document understanding domain.

3.2.1 TASK #1: MASKED LANGUAGE MODELING

To encourage the model to learn contextual representation, similar to BERT Devlin et al. (2018), we mask a portion of the text region of the input document with RGB values [255, 255, 255] at random. The masking ratio in BERT is 15%, however, we find that 30% is the best choice in our experiments. The encoder-decoder architecture is trained to predict the token indexes (in a predefined vocabulary) of the masked text regions with cross-entropy loss, according to the information provided by the contexts (the unmasked regions around the masked text region) in a document. To avoid the interference of sub-words produced by WordPiece Song et al. (2020), we only choose the first sub-word for keeping the number of words in the document images. Notice that the Masked Language Modeling task in MaskDoc does not need convert the image to text tokens as input, which is necessary for the NLP domain.

Decoder for Task #1. For Mask Language Modeling, we employ an extremely lightweight 2-layer MLP as decoder. The multi-model feature of each text region is extracted from the fused feature maps F_{fuse} by ROI-Align Ren et al. (2015) and the prediction head is implemented by a 2-layer MLP. Then, the prediction is computed as follows,

$$P_i^{mlm} = MLP(ROI - Align(F_{fuse}, b_i)), \tag{1}$$

where b_i is the bounding box of the $i-th$ text region, and the P_i^{mlm} is optimized by cross-entropy loss with 30,522 token categories.

3.2.2 TASK #2: MASKED IMAGE MODELING

In MAE He et al. (2022a), BEiT Bao et al. (2022) and SimMIM Xie et al. (2022), patch-level Masked Image Modeling has shown strong potential in representation learning. However, in the document understanding domain, patch-level feature learning is too coarse to represent the details of each text or word region. Therefore, we introduce a text-region level visual representation learning task called Masked Image Modeling to enhance the document representation and understanding. Instead of classifying the classes defined by tokenization like LayoutLMv3 Huang et al. (2022) and BEiT, we regress the raw pixel values of the masked region with mean square error loss following SimMIM

and MAE. Specifically, we mask the rectangle text regions and predict the RGB values of missing pixels, leading to significant performance improvements for representation learning.

Decoder for Task #2. We develop a Fully Convolutional Network with Transpose Convolution to perform the masked document image reconstruction. Specifically, we apply global average pooling to aggregate each text region's feature, and generate the embedding Emb_{style} that mainly represents the "style" information for each text region. To strength the accuracy of the textual content of restored image regions. We encode the MLM prediction P_i^{mlm} to $Emb_{content}$ by using a embedding layer, representing "content" information. At last, we concatenate Emb_{style} and $Emb_{content}$ and feed it to a Fully Convolutional Networks, generating the final image prediction P_i^{mim}. The Mask Image Modeling process is formulated as follows,

$$Emb_{style} = GAP(ROI - Align(F_{fuse}, b_i)), \tag{2}$$

$$Emb_{content} = EmbeddingLayer(P_i^{mlm}), \tag{3}$$

$$P_i^{mim} = FCN(Concat(Emb_{style}, Emb_{content})), \tag{4}$$

where GAP is the global average pooling operator. In MIM, we follow MAE and predict the missing pixels of masked text regions. For example, we resize the spatial resolutions of each masked text-region to 64×64, that means each text-region's regression target is $12,288 = 64 \times 64 \times 3$ standing for the RGB values of pixels. The P_i^{mim} is optimized by MSE loss in the training phrase.

3.3 Downstream Tasks

The MaskDoc pretraining scheme contributes to a visual-textual representation with a single image input. The pretrained feature is available to model fine-tuning and profits numerous tasks of document image understanding.

3.3.1 Task #1: Document Image Classification

Document image classification aims to predict the category of each document image, which is one of the fundamental tasks in office automation. For this task, we downsample the output feature maps of backbone networks by four 3×3 convolutional layers with stride 2. Next, the image representation is fed into the final linear layer with softmax to predict the class label.

3.3.2 Task #2: Document Layout Analysis

Document layout analysis aims to identify the layout components of document images by object detection. Following DiT, we leverage Cascade R-CNN Cai & Vasconcelos (2018) as the detection framework to perform layout element detection and replace its backbone with MaskDoc. Thanks to the multi-scale context design of backbone networks, four resolution-modifying features (P2, P3, P4, and P5) of FPN fusion layers on backbone networks are sent into the iterative structure heads of the detector.

3.3.3 Task #3 Table Structure Recognition

Table structure recognition aims to recognize the internal structure of a table which is critical for document understanding. Specifically, We employ Cascade R-CNN for cell detection in our table structure recognition framework while replacing the feature encoder with backbone networks. Since some table images are collected by cameras and many cells are deformed, we modify the final output of Cascade R-CNN to the coordinate regression of four vertexes of cells.

3.3.4 Task #4: Document OCR

We tend to read the text in an end-to-end manner based on MaskDoc. Our OCR method consists of both the word-level text detection and recognition modules. They share the features of backbone networks and are connected by ROI-Align operations. The text detection module adopts the standard DB Liao et al. (2020) algorithm, which predicts the binarization mask for word-level bounding boxes. Similar to NRTR Sheng et al. (2019), the text recognition module is composed of multi-layer Transformer decoders to predict character sequences for each word.

3.3.5 TASK #5: END-TO-END INFORMATION EXTRACTION

The aim of the task is to extract entity-level contents of key fields from given documents without predefined OCR information. We evaluate the MaskDoc model based on the architecture of document OCR and devise a new branch for semantic entity extraction. Concretely, another DB detection is developed to identify the entity bounding boxes. An additional MLP block is performed with the ROI features to classify entity label. These bounding boxes are utilized for word grouping to merge the text contents from Task #4. At length the key information is obtained by grouping words according to the reading order.

4 EXPERIMENTS

In order to demonstrate the effectiveness of MaskDoc, we pretrain the model on the large-scale public dataset and conduct extensive experiments on four document relevant benchmarks for five downstream tasks mentioned above. Additionally, we provide ablation study to explore the impacts of each strategy and component.

4.1 DATASETS

Pretraining Data By following DiT Li et al. (2022), we pretrain MaskDoc on the IIT-CDIP Test Collection 1.0 dataset Lewis et al. (2006), whose multi-page documents is split into single pages, totally 42 million document images.

RVL-CDIP Harley et al. (2015a) contains 400,000 grayscale document images in 16 classes, with 25,000 images per class. We adopt RVL-CDIP as the benchmark to carry out experiments on document classification task. Average classification accuracy is used evaluate model performance.

PubLayNet Zhong et al. (2019) is a large dataset of more than 360,000 paper images built by automatically parsing PubMed XML files. The five typical document layout elements (text, title, list, figure, and table) are annotated with bounding boxes. Mean average precision (mAP) @ intersection over union (IOU) is used as the evaluation metric of document layout analysis.

WTW Long et al. (2021) covers unconstrained table in natural scene, requiring table structure recognizer to have both discriminative and generative capabilities. It has a total of 14581 images in a wide range of real business scenarios and the corresponding full annotation (including cell coordinates and row/column information) of tables.

FUNSD Jaume et al. (2019) consists of 199 real, fully annotated, scanned forms. The purpose of FUNSD is to identify four categories of semantic entity (questions, answers, headers, and other) and entity linking presented in the form. In this work, we focus on the document OCR and end-to-end information extraction. For evaluation of the two tasks, We compute the normalized Levenshtein similarity (1-NED) between the predictions and the ground truth.

4.2 IMPLEMENTATION DETAILS

Pretraining We extract both visual features and content features based on the proposed encoder network, which is composed mainly of CNN and Transformer. The small and large parameter settings are built for MaskDoc. To balance efficiency and effectiveness, $\text{MaskDoc}_{\text{Small}}$ consists of ResNet-50 and a 12-layer Transformers with 128 hidden size and 8 attention heads. It introduces only 28M parameters which still gets superior performance in downstream tasks. A larger version, $\text{MaskDoc}_{\text{Large}}$ is set up as ResNeXt-101 Xie et al. (2016) and a 24-layer Transformers and its total parameters are 238M. The initialization of ResNet and ResNeXt are based on the ImageNet Deng et al. (2009) pretrained weights and the 24-layer Transformers are initialized from existing language models Sun et al. (2019). The rest parameters of MaskDoc are trained from scratch.

Fine-tuning on RVL-CDIP We evaluate MaskDoc for document image classification. We fine-tune the model on RVL-CDIP for 20 epochs with cross-entropy loss. The learning rate is set for 3e-4 and the batch size is 28. The input images are resized to 960×960 and maintain its aspect ratio. We use label smoothing=0.1 in the loss function. Besides, the data augmentation methods such as CutMix and MixUp with 0.3 probability are applied on training phase.

Table 1: Performance comparisons on the RVL-CDIP dataset. We report classification accuracy on the test set. MaskDoc gets the highest performance with image-only input.

Modality	**Methods**	**Accuracy**	**#Param.**
T	BERT$_{Base}$ *Devlinet al.* (2018)	89.8%	110M
T	BERT$_{Large}$ *Devlinet al.* (2018)	89.9%	340M
V+T	SelfDoc Li et al. (2021a)	93.8%	-
V+T	LayoutLM$_{Large}$ *Xuet al.* (2020*b*)	94.4%	390M
V+T	UDoc Gu et al. (2021)	95.1%	272M
V+T	LayoutLMv2$_{Large}$ *Xuet al.* (2020*a*)	95.6%	426M
V+T	LayoutLMv3$_{Large}$ *Huanget al.* (2022)	95.9%	368M
V+T	**DocFormer**$_{Base}$ *Appalarajuet al.* (2021)	**96.2%**	183M
V	ResNeXt-101 Xie et al. (2016)	90.7%	88M
V	BEiT$_{Base}$ *Baoet al.* (2022)	91.1%	87M
V	DiT$_{Base}$ *Liet al.* (2022)	92.1%	87M
V	DiT$_{Large}$ *Liet al.* (2022)	92.7%	304M
V	MaskDoc$_{Small}$	93.4%	**28M**
V	**MaskDoc**$_{Large}$	**94.5%**	238M

Fine-tuning on PubLayNet We evaluate on the validation set of PubLayNet for document layout analysis. We fine-tune the Cascade R-CNN and initialize the backbone with the pretrained MaskDoc. The detector is trained with Momentum optimizer and a batch size of 8 for 8 epochs. We set learning rate as 1e-2 with 500 warmup steps, decay to 1e-3 and 1e-4 on 3 and 6 epoch. We introduce random resized cropping to augment the training images while the short edges does not exceed 800 pixels.

Fine-tuning on WTW We conduct experiments on WTW for table structure recognition. We also employ Cascade R-CNN to detect the cells of table whose backbone is replaced by pretrained MaskDoc. We fine-tune our model end-to-end using ADAM Kingma & Ba (2014) optimizer for 20 epochs with a batch size of 16 and a learning rate of 1e-4. The input images are resized to 640 $\times$ 640 after random scaling and the long size being resized to 640 pixels.

Fine-tuning on FUNSD On account of the full annotations, both the document OCR and end-to-end information extraction tasks are measured on FUNSD. We set the text recognition network as a 6-layers Transformer and fine-tune the whole model for 1200 epochs with a batch size of 32. We follow a cosine learning rate policy and set the initial learning rate to 5e-4. Extra position embeddings are appended to roi-features and we pass it to each layer of decoders. The training losses except DB detector are the cross-entropy function. Additionally, the same loss is estimated for each decoder layer in the text recognition module for better convergence training.

4.3 Comparisons with the State-of-the-art

To investigate the effect of visual-textual representations, we benchmark the MaskDoc with several state-of-the-art techniques on different downstream tasks. Since only small datasets (149 training documents on FUNSD) and to avoid exceeding GPU memory (a tremendous number of table cells on WTW), we only evaluate MaskDoc$_{\text{Small}}$ on the FUNSD and WTW datasets.

RVL-CDIP As the Tab. 1 shows, OCR-based approaches such as DocFormer Appalaraju et al. (2021) with multi-modal modeling takes superior performance on RVL-CDIP. Compared with image-only methods, it relies on stand-alone OCR systems which are low efficiency for practical applications. Our MaskDoc$_{\text{Small}}$ achieves an accuracy of 93.4% and outperforms previous image-only methods with a much smaller model size. Besides, MaskDoc$_{\text{Large}}$ drives an accuracy of 94.5%, which brings a 1.8% significantly improvement over DiT$_{\text{Large}}$ Li et al. (2022). Intuitively, the results show that MaskDoc is effective for semantic understanding on document image classification.

PubLayNet The experiment results on PubLayNet are presented in Tab. 2. It is observed that MaskDoc achieves new SOTA performance of 95.4% and 95.5% mAP scores for both small and large settings. $\text{MaskDoc}_{\text{Small}}$ beats even $\text{LayoutLMv3}_{\text{Base}}$ Huang et al. (2022) (the result of $\text{LayoutLMv3}_{\text{Large}}$ is not released in the paper) which contains multi-modal inputs by 0.3%. We suggest that our dual-modal pretraining tasks can learn rich visual-textual representations of document images and performs excellently in confusing situations. Notably, $\text{MaskDoc}_{\text{Large}}$ gets 0.1% mAP gain over that on $\text{MaskDoc}_{\text{Small}}$. We attribute this subtle improvement to the over-fitting to a certain extent with increasing model size.

WTW Tab.3 shows quantitative results of table structure recognition on the WTW dataset. MaskDoc achieves the physical 78.9% F1-score among all published methods. We reconstruct the table structure based on the detection results of table cells. The superior performance of MaskDoc largely due to the proposed pretraining framework.

Table 2: Performance comparisons on the PubLayNet validation set. The mAP @ IOU [0.50:0.95] is used as the metric.

Methods	**mAP**
BEiT_{Base}	93.1%
UDoc	93.9%
DiT_{Large}	94.9%
LayoutLMv3_{Base}	95.1%
MaskDoc_{Small}	95.4%
MaskDoc_{Large}	**95.5%**

Table 3: Performance comparisons on the WTW dataset. The F1-Score is used to measure the accuracy of cell coordinate when IOU=0.9.

Methods	**F1-Score**
Split+Heuristic Tensmeyer et al. (2019)	3.4%
Faster-RCNN Ren et al. (2015)	64.4%
TridenNet Li et al. (2019)	65.0%
CenterNet Duan et al. (2019)	73.1%
Cycle-CenterNet Long et al. (2021)	78.3%
MaskDoc_{Small}	**78.9%**

FUNSD We evaluate $\text{MaskDoc}_{\text{Small}}$ on both the document OCR and end-to-end information extraction tasks. As shown in Tab. 4, $\text{MaskDoc}_{\text{Small}}$ achieves outstanding performance, 84.1% 1-NED for document OCR and 55.0% 1-NED for information extraction. Significantly, the whole network is end-to-end trainable. Compared to StrucTexT Li et al. (2021b) and LayoutLMv3 with need of separately stage-wise training strategies, our model alleviates the error propagation in a documental system with key information parsing.

Table 4: Performance comparisons on FUNSD. We present the Normalized Edit Distance (1-NED) for the word-level document OCR and the entity-level information extraction. The * denotes a multistage process in which the methods are applied using our OCR results and entity boxes for word grouping in information extraction.

The results of Document OCR:

Methods	**1-NED**
DB+NRTR Sheng et al. (2019)	73.8%
Google Vision Google (2019)	76.4%
MaskDoc_{Small}	**84.1%**

The results of Information Extraction:

Methods	**1-NED**
$\text{StrucTexT}^{*}_{Base}$	46.8%
$\text{LayoutLMv3}^{*}_{Base}$	53.1%
MaskDoc_{Small}	**55.0%**

4.4 Ablation Study

We conducted several experiments to evaluate the effectiveness of our design. All pretrain models in experiments are pretrained on a subset of 1 million sampled from IIT-CDIP Test Collection 1.0 dataset Lewis et al. (2006). We conduct ablation studies on document layout analysis(i.e., PubLayNet) and document image classification(i.e., RVL-CDIP).

Ablation for Different Backbone Architectures In this study, we evaluate the impact of the backbone architectures of the proposed MaskDoc by replacing them with some popular backbone, e.g, ViT Dosovitskiy et al. (2020) and SwinTransformer Liu et al. (2021). As shown in Tab. 5. Under the same pretraining configuration, the proposed backbone network of $\text{MaskDoc}_{\text{Small}}$ even achieves the better results of 92.5% accuracy and 94.9% mAP on RVL-CDIP and PubLayNet, respectively. Replaced with ViT backbone, the performance of two benchmarks dropped by 3.9% accuracy and

Table 5: Ablation for Different Backbone Architectures and Pretraining Tasks of MaskDoc.

Backbone	**Pretrain Tasks**	**RVL-CDIP**	**PubLayNet**	**#Param.**
ViT$_{base}$ + FPN	MLM+MIM	88.6%	93.2%	116M
SwinT$_{base}$ + FPN	MLM+MIM	82.4%	92.2%	176M
MaskDoc$_{Small}$	MIM	91.8%	94.1%	28M
MaskDoc$_{Small}$	MLM	92.0%	94.5%	28M
MaskDoc$_{Small}$	MLM+MIM	92.5%	94.9%	28M
MaskDoc$_{Large}$	MLM+MIM	94.1%	95.5%	238M

2.3% mAP. Replaced with SwinTransformer backbone, the degradation is more obvious. However, switching to the backbone network of MaskDoc$_{\text{Large}}$, it improves peformance by 1.6% on RVL-CDIP and 1.0% on PubLayNet.

Ablation for Different pretraining Tasks In this study, we evaluate the impact of different pretraining tasks. As shown in Tab. 5, Mask Image Modeling achieves an accuracy of 91.8% on RVL-CDIP and an mAP of 94.1% on PubLayNet with the backbone of MaskDoc$_{\text{Small}}$, Mask Language Modeling achieves 92.0% and 94.5% respectively. By combining MLM and MIM, the accuracy can be improved to 92.5% on RVL-CDIP and the mAP achieves 94.9% on PubLayNet. Jointly exploring both MLM and MIM makes the feature representation of MaskDoc more powerful.

Table 6: Ablation for Different Masking Ratio.

Masking Ratio	**Pretrain Tasks**	**RVL-CDIP**	**PubLayNet**
0.15	MLM+MIM	92.1%	94.7%
0.30	MLM+MIM	92.5%	94.9%
0.45	MLM+MIM	91.7%	94.8%
0.60	MLM+MIM	92.4%	94.8%

Ablation for Different Masking Ratio We investigate the impact of training with different masking ratios. As shown in Tab.6, by replacing the masking ratio with 0.15, 0.30, 0.45 and 0.60, the accuracy of RVL-CDIP is 92.1%, 92.5%, 91.7% and 92.4%, respectively. We also report the results on PubLayNet, the mAP of PubLayNet is 94.7%, 94.9%, 94.8%, and 94.8%, respectively. It suggests that the best masking ratio of our pretraining tasks is 0.30. At the same time, It also suggests that the performance of downstream tasks is less sensitive to the selection of masking ratio.

Ablation for Different Masking Strategies In this study, we evaluate the impact of masking strategies by replacing them with patch-level masking strategy. The results in performance drop, e.g., 88.6% to 78.5% accuracy score on RVL-CDIP and 93.4% to 91.5% mAP on PubLayNet. This suggests that the proposed Text-Region masking strategy is critical to document classification. Notice that we both use ViT without any features fusion strategy as the backbone for a fair comparison.

5 Conclusion

This work successfully explores a novel pretraining framework named MaskDoc to learn visual-textual representations for document image understanding with a single image modality input. By performing text region-based image masking, and then predicting both corresponding visual and textual contents, the proposed encoder, can benefit from large-scale document images efficiently. And the SOTA performances on five document understanding tasks including document image classification, layout analysis, table structure recognition, document OCR, and end-to-end information extraction verify the effectiveness of MaskDoc.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *ICCV*, pp. 993–1003, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pp. 6154–6162, 2018.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 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. *CoRR*, abs/2010.11929, 2020.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pp. 6569–6578, 2019.

Google. Google cloud vision. `https://cloud.google.com/vision`, 2019.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. Unified pretraining framework for document understanding. *NeurIPS*, 2021.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, pp. 991–995, 2015a.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, pp. 991–995, 2015b.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022a.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022b.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM Multimedia*, 2022.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*, volume 2, pp. 1–6, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a test collection for complex document information processing. In *ACM SIGIR*, pp. 665–666, 2006.

Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *CVPR*, pp. 5238–5246, 2017.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*, 2022.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *CVPR*, pp. 5652–5660, 2021a.

Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *ICCV*, pp. 6054–6063, 2019.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *ACM Multimedia*, pp. 1912–1920, 2021b.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, volume 34, pp. 11474–11481, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents. In *ICDAR*, pp. 548–563, 2021.

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. In *ICCV*, pp. 10012–10022, 2021.

Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *ICCV*, pp. 944–952, 2021.

Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *ECCV*, pp. 67–83, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pp. 8821–8831, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015.

Fenfen Sheng, Zhineng Chen, and Bo Xu. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *ICDAR*, pp. 781–786, 2019.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

Chris Tensmeyer, Vlad I Morariu, Brian Price, Scott Cohen, and Tony Martinez. Deep splitting and merging for table structure decomposition. In *ICDAR*, pp. 114–121, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pp. 9653–9663, 2022.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020a.

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 *ACM SIGKDD*, pp. 1192–1200, 2020b.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *ICDAR*, pp. 1015–1022, 2019.