WUKONG-READER: Multi-modal Pre-training for Fine-grained Visual Document Understanding

Haoli Bai*, Zhiguang Liu*, Xiaojun Meng*, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang†, Lu Hou†, Jiansheng Wei, Xin Jiang, Qun Liu
Huawei Noah's Ark Lab
{wangliangwei, houlu3}@huawei.com

Abstract

Unsupervised pre-training on millions of digital-born or scanned documents has shown promising advances in visual document understanding (VDU). While various vision-language pre-training objectives are studied in existing solutions, the document textline, as an intrinsic granularity in VDU, has seldom been explored so far. A document textline usually contains words that are spatially and semantically correlated, which can be easily obtained from OCR engines. In this paper, we propose WUKONG-READER, trained with new pre-training objectives to leverage the structural knowledge nested in document textlines. We introduce textline-region contrastive learning to achieve fine-grained alignment between the visual regions and texts of document textlines. Furthermore, masked region modeling and textline-grid matching are also designed to enhance the visual and layout representations of textlines. Experiments show that our WUKONG-READER has superior performance on various VDU tasks such as information extraction. The fine-grained alignment over textlines also empowers WUKONG-READER with promising localization ability.

1 Introduction

Visual document understanding (VDU) handles various types of digital-born or scanned documents like forms, tables, reports, or research papers, and is becoming increasingly important for real-world industrial practices [6]. Multi-modal pre-training on millions of documents is a popular solution for visual document understanding [11, 28, 30, 29, 13, 25]. Unlike the conventional vision-language pre-training over natural images and their paired short and abstractive descriptions [27, 20, 19], the document texts are usually long and highly correlated with the images, since they can be easily obtained from accurate Optical Character Recognition (OCR) engines from the scanned images. Therefore, it is crucial to strengthen the connection between vision and language for VDU with more fine-grained alignment across the two modalities.

(a) Letter. (b) Receipt.

Figure 1: Document textlines from the letter in FUNSD [15] and receipt in SROIE [14], respectively.

Towards that end, existing efforts seek to align the visual and textual knowledge of documents at different levels. A commonly used pre-training objective for documents is masked language modeling [7] over document text tokens [30, 29, 13, 28, 11, 25], often accompanied by the layout information encoded via the positional embedding. Besides, various visual and vision-language multimodal pre-training objectives are also proposed, leveraging the patch-level features [29, 13], object-level features from object detectors [21, 10], or the whole image feature through a global text-image matching loss [29].

However, as an intrinsic granularity for VDU, document textlines have been mostly neglected in past efforts. Intuitively, a textline contains a set of words that are spatially and semantically related. For instance of information extraction, the desired text span (e.g., the names on letters and addresses on receipts in Figure 1) often appears in a single textline. Therefore, the document textline serves

* Equal Contribution.
† Corresponding authors.

as an appealing fine-grained granularity for VDU tasks. While StructualLM [2] similarly considers textlines as cell layout information, they only use the textual features of these textlines in language modeling. Instead, in this work, we seek to enhance the multi-modal representation of a document by aligning the visual region and text span corresponding to the same textline.

In this work, we propose WUKONG-READER, a pre-trained document model with a hybrid dual- and single-stream multimodal architecture. To learn fine-grained document representation, we propose the *Textline-Region Contrastive Learning* to align the visual and textual features of document textlines from the dual-stream encoders. The objective thus connects the spatial and semantic information among document textlines for various VDU tasks. Additionally, we also introduce two other objectives to further improve the textline representation. We design the *Masked Region Modeling* to recover the masked textline regions, so as to enhance the visual features of textline. We also propose the *Textline Grid Matching* to strengthen the layout information of textlines, which localizes each word of textlines to the pre-defined image grids. Similar to previous works [30, 29, 13], the classic masked language modeling objective is also applied over document texts.

Experimental results show that our WUKONG-READER brings a noticeable improvement in the performance of various document understanding tasks. In particular, WUKONG-READER$_{\text{large}}$ with 470M parameters achieves the weighted F1 score of 93.62 on FUNSD [15] and 98.15 on SROIE [14], leading the new state-of-the-art records on information extraction tasks. We also demonstrate that the textline-based pre-training objectives empower the model with meaningful textline features with promising localization ability.

2 Related Work

Visual document understanding (VDU) has been widely studied in recent years [11, 2, 25, 29, 30]. VDU tasks are abundant in textual and visual information, as intensive texts and their layout information can be extracted from documents via Optical Character Recognition (OCR) or other document parsers. Therefore, multi-modal pre-training has been a popular solution for VDU. To deal with the textual input, a pre-trained text encoder (e.g., BERT [7]; RoBERTa [22]) is usually applied to learn contextualized word representation. Meanwhile, a pre-trained visual encoder such as CNN-based [29] and transformer-based [13, 26] models are applied for visual features. Various self-supervised pre-training objectives over millions of documents have shown promising effects for VDU. Reconstructive objectives such as masked language modelling (MLM) [7], and masked image modelling, (MIM) [8], are often used to perform the self-supervised document pre-training [18, 29].

Given the fact that textual knowledge is parsed from the document image, existing efforts explore various document granularities to align the vision and language modalities. They can be generally divided into four categories: 1) **Word-level**: LayoutLM [30] jointly models the inner-relationship between texts and layout 2D positions from documents, via pre-trained language models [7, 22]. However, the visual features are not used in the pre-training architecture. TILT [26] additionally adds a contextualized image embedding to the word embedding. 2) **Grid/Patch-level**: LayoutLMv2 [29], DocFormer [3] and ERNIE-Layout [25] extract image grid features with CNN backbones, and LayoutLMv3 further uses ViT [8] to encode image patches. To achieve the cross-modal alignment, they adopt the text-image alignment (i.e., TIA) and matching (i.e., TIM) objectives during pre-training. 3) **Object-level**: SelfDoc [21] and UniDoc [10] extract object features via document object detectors, and concatenate them with word features. SelfDoc [21] uses two cross-modality attention functions to identify the inner-relationships from one modality to another. UniDoc [10] designs the similarity-preserving knowledge distillation to encourage alignment between words and visual features. 4) **Cell-level**: StructualLM [2] uses the textual features of cell layout information, which is similar to document textlines. However, it only considers the textual feature without the visual information.

Different from existing works, we target at the textline-level features of both textual and visual modalities. We propose a hybrid dual- and single-stream multimodal architecture to achieve fine-grained alignment over textlines. By leveraging the structural knowledge nested in document textlines, we believe such an important granularity of documents can benefit both language and visual representation learning in VDU tasks.

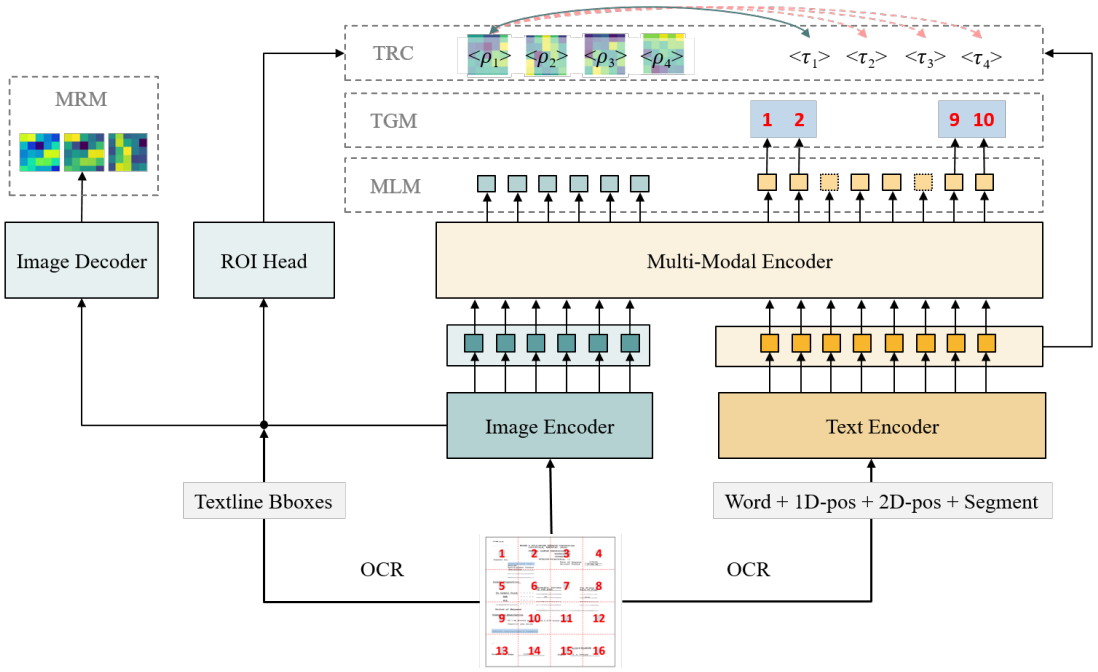


Figure 2: Architecture of the proposed WUKONG-READER. The scanned document is sent to the image encoder to extract visual features. Meanwhile, OCR tools are applied to extract words, bounding boxes as 2D positional embeddings to the text encoder. WUKONG-READER is pre-trained with 1) masked language modeling (MLM); 2) textline-region contrastive learning (TRC) to learn fine-grained textline alignment; 3) masked region modeling (MRM) to enhance the visual representation of textlines; and 4) textline grid matching (TGM) which classifies the words of selected textlines (blue) into different image grids (red). More details in Section 3.2.

3 Methodology

In this work, we propose WUKONG-READER, a new pre-trained multi-modal model for visual document understanding. Our model jointly encodes the visual image and textual tokens via two monomodal encoders, followed by a multi-modal encoder to fuse the two modalities. To leverage the structural information nested in document textlines, WUKONG-READER is pre-trained with several novel pre-training objectives for fine-grained representation learning of documents.

3.1 Model Architecture

The overall architecture of the proposed WUKONG-READER is shown in Figure 2. WUKONG-READER encodes the document image and text through separate encoders and then fuse the two modalities via the multi-modal encoder. Besides, we also deploy an RoIhead and an image decoder for fine-grained learning over document textlines.

Image Encoder. We use the Mask-RCNN model trained on PubLayNet[1] to learn the visual representations for WUKONG-READER. Specifically, we use the visual backbone of Mask-RCNN as the image encoder. The visual features from the image encoder are adaptively pooled into 49 visual tokens. The RoIHead of Mask-RCNN then extracts the regional features of document textlines for contrastive learning with texts. Meanwhile, an image decoder is also deployed to recover the visual features over textline regions.

Text Encoder. Given a document image, we use an off-the-shelf OCR tool to extract the textual information from the image, which includes both the words and their corresponding bounding boxes. Following [30, 29], we normalize the bounding boxes within $[0, 1000]$ and use 2D positional embedding layers to encode the layout information. We initialize the text encoder with the first six layers of the RoBERTa model, and employ the spatial-aware self-attention mechanism following [29] in the Transformer layers. We calculate the input embedding as the summation of the token embedding from RoBERTa tokenizer, 1D positional embedding, 2D positional embedding and the segment embedding following [29]. The input embedding is then fed to the text encoder to get textual features.

[1]We adopt the configuration of "MaskRCNN ResNeXt101 32x8d FPN 3X" as provided in https://github.com/hpanwar08/detectron2.

Multimodal Encoder. We concatenate the token-level features from both vision and text, and feed them to the multi-modal encoder to jointly fuse the two modalities. We initialize the multi-modal encoder with the rest layers of the RoBERTa model. Before concatenation, we also add 1D and 2D positional embeddings to visual features following [29].

3.2 Pre-training Objectives

As the fundamental pre-training objective in modeling languages, we use the Masked Language Modeling (MLM) to recover the masked word tokens in the document text. We follow the standard masking strategy in BERT [7] and mask out 15% word tokens. Besides, to prevent information leakage, we also cover the corresponding image regions and set their bounding boxes to zeros, following [29].

Despite the powerful effect of MLM, it fails to explicitly leverage the visual information. Previous attempts [31, 29, 13] consider multi-modal pre-training objectives, but they usually lack fine-grained multi-modal alignment, which hinders a deeper understanding of the document. For instance, in [29], the TIA loss only predicts whether a token is covered or not, without requiring the model to understand the content. The TIM loss measures only the alignment between the global document image and text, without considering more detailed content. Below we propose to mine the fine-grained image-text alignment through multiple new pre-training objectives.

3.2.1 Textline-Region Contrastive Learning

As is shown Figure 1, a textline of a document returned by OCR usually contains a set of words that are semantically related. We are thus motivated to exploit structural knowledge within it by textline-region contrastive learning (TRC). Specifically, to obtain the textual representation of a textline, we average the features of tokens within that textline. Besides the textual feature, we also employ a multi-layer perception based RoIHead on top of the image encoder to extract the visual feature corresponding to the textline region in the document image.

Contrastive representation learning has been widely used for vision-language cross-modal pre-training [27, 32]. To enhance the alignment of a document image and its textual content, we also utilize contrastive learning to align the textline-region and texts. For ease of presentation, we suppose there is a batch of N document image-text pairs, and each document has L textlines. For the n-th document, denote $\boldsymbol{\rho}_n$ and $\boldsymbol{\tau}_n$ as the visual and textual feature of document textlines, respectively. Note that we pad $\boldsymbol{\rho}_n$ and $\boldsymbol{\tau}_n$ with 0 to length L for documents with fewer than L textlines. For each document image, its paired text is used as its positive, and the texts from other documents are used as its negatives. The contrastive learning from image to text can be formulated as

$$\mathcal{L}(\boldsymbol{\rho}_m, \boldsymbol{\tau}_{1:N}) = -\frac{1}{N} \log \frac{\exp(s(\boldsymbol{\rho}_m, \boldsymbol{\tau}_m))}{\sum_{n=1}^{N} \exp(s(\boldsymbol{\rho}_m, \boldsymbol{\tau}_n))},$$

where $s(\boldsymbol{\rho}_m, \boldsymbol{\tau}_n)$ represents the similarity of the m-th image to the n-th text computed in the granularity of textlines. By symmetry, the contrastive objective from text to image can be similarly established as

$$\mathcal{L}(\boldsymbol{\tau}_m, \boldsymbol{\rho}_{1:N}) = -\frac{1}{N} \log \frac{\exp(s(\boldsymbol{\tau}_m, \boldsymbol{\rho}_m))}{\sum_{n=1}^{N} \exp(s(\boldsymbol{\tau}_m, \boldsymbol{\rho}_n))}.$$

The TRC objective is the summation of the two loss terms:

$$\mathcal{L}_{\text{TRC}} = \frac{1}{2} \sum_{m=1}^{N} \big(\mathcal{L}(\boldsymbol{\rho}_m, \boldsymbol{\tau}_{1:N}) + \mathcal{L}(\boldsymbol{\tau}_m, \boldsymbol{\rho}_{1:N})\big). \quad (1)$$

The cross-modal interaction is reflected in how the similarity between the image and text is computed. Existing contrastive learning methods simply calculate the similarity based on the global feature of the image or text [29, 13, 25]. To establish fine-grained alignment over textlines, the key lies in the following similarity metric. Inspired by [32, 9] we adopt the average textline maximum similarity which is computed as

$$s(\boldsymbol{\rho}_m, \boldsymbol{\tau}_n) = \frac{1}{L} \sum_{l=1}^{L} \max_{1 \le k \le L} \big(\boldsymbol{\rho}_{m,l}^{\top} \boldsymbol{\tau}_{n,k}\big),$$

$$s(\boldsymbol{\tau}_m, \boldsymbol{\rho}_n) = \frac{1}{L} \sum_{l=1}^{L} \max_{1 \le k \le L} \big(\boldsymbol{\rho}_{m,k}^{\top} \boldsymbol{\tau}_{n,l}\big),$$

where $\boldsymbol{\rho}_{m,l}$ represent the l-th textline of the m-th visual feature, and $\boldsymbol{\tau}_{n,k}$ similarly denotes the k-th textline of the n-th textual feature, respectively. The defined similarity shows that for each image region of textlines, we find their most similar text segments. Similarly, for each textline text, we also find its closest image region of textlines. With the objective in Equation (1), such design intrinsically encourages the fine-grained alignment between the visual and textual features of textlines.

3.2.2 Masked Region Modeling

To enhance the visual representation of document textlines, we further propose the Masked Region Modeling (MRM) to recover the masked pixels of textline regions during pre-training.

Specifically, for the n-th document image, we randomly mask 15% textlines of the document for recovery. A document textline is usually dominated by white background pixels instead of foreground characters. To avoid trivial solutions and balance the foreground and background pixels in a textline, we mask all black strokes as well as 15% of background pixels within each textline. Our pre-training objective is to predict these masked pixels based on their surroundings. On top of the image encoder, we use three deconvolution layers as the image decoder to recover the textline visual features $\tilde{\boldsymbol{\rho}}_n^{\text{mask}}$. As the pre-training objective of MRM, we adopt the ℓ_1 loss [21] between the reconstructed $\tilde{\boldsymbol{\rho}}_n^{\text{mask}}$ and the original $\boldsymbol{\rho}_n$:

$$\mathcal{L}_{\text{MRM}} = \sum_{n=1}^{N} \ell_1(\boldsymbol{\rho}_n, \tilde{\boldsymbol{\rho}}_n^{\text{mask}}). \tag{2}$$

Note that if a masked textline contains masked tokens introduced in the MLM task, we do not calculate the reconstruction loss for this token.

3.2.3 Textline Grid Matching

Aside from enhancing the visual representations of textlines, layout information of textlines also plays an important role for visual document understanding. We thus introduce the Textline Grid Matching (TGM) to explicitly model the layout of each word in textlines.

Specifically, we first split each document image into G pre-defined grids. Then we randomly sample 15% textlines that are not used in MLM and MRM, and predict which grid each output token in the selected textline belongs to. For the n-th document, suppose we sampled L' textlines. We first transform the output from the multi-modal encoder to obtain a set of grid logits $\boldsymbol{y}_{l,1:T_l}$, where T_l is the number of words in the l-th textline. To avoid leakage of position information, we set the 2D bounding boxes of tokens in the selected textlines as [0, 0, 0, 0]. We then classify the grid logits into the G classes over the image, by minimizing the cross-entropy loss ℓ_{ce} as

$$\mathcal{L}_{\text{TGM}_n} = \sum_{l=1}^{L'} \sum_{t=1}^{T_l} \ell_{ce}(\boldsymbol{y}_{l,t}, \boldsymbol{g}_{l,t}),$$

where $\boldsymbol{g}_{l,t}$ is the corresponding ground-truth label of $\boldsymbol{y}_{l,t}$. The Texline Grid Matching loss for a mini-batch is the summation over all the documents in this batch:

$$\mathcal{L}_{\text{tgm}} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\text{TGM}_n}. \tag{3}$$

Compared with the previous TIA loss in LayoutLMv2 [29] which simply classifies whether a token is masked, TGM enhances the layout information via explicit grid localization from both nearby unmasked textual tokens and visual regions.

The total pre-training loss is the combination of the four pre-training objectives introduced above:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLM}} + \lambda_1 \mathcal{L}_{\text{TRC}} + \lambda_2 \mathcal{L}_{\text{MRM}} + \lambda_3 \mathcal{L}_{\text{TGM}},$$

where λ_1, λ_2 and λ_3 are the scaling parameters that control the weights of different loss terms. For simplicity, we choose $\lambda_1 = 0.2, \lambda_2 = \lambda_3 = 1$ for all our experiments. It is possible that better performance can be achieved with a more careful tuning of these scaling parameters.

4 Experiments

4.1 Experimental Setup

Model Configuration. We provide two kinds of model sizes: WUKONG-READER$_{\text{base}}$ and WUKONG-READER$_{\text{large}}$. For both sizes, we use the pre-trained MaskRCNN model to initialize the image encoder, including the ResNet-101 visual backbone and the multi-layer-perception based RoI-Head. We adopt the RoBERTa-base and RoBERTa-large[2] as backbones to initialize the rest parts of the base and large models, respectively. Specifically, WUKONG-READER$_{\text{base}}$ adopts six transformer layers for the textual encoder, and another six layers for vision-language encoder. For WUKONG-READER$_{\text{large}}$, we keep the six-layer Transformer architecture for the text encoder, and extend the multi-modal encoder to the rest 18 Transformer layers. Following [29, 13], the image resolution is set as 224×224, which is then adaptively pooled into 49 visual tokens after the image encoder. The textual sequence length is to 512. For textline-region contrastive learning, we choose the first 64 textlines for each document. We evaluate WUKONG-READER

[2]RoBERTa-base and RoBERTa-large are from https://huggingface.co/roberta-base/tree/main and https://huggingface.co/roberta-large/tree/main, respectively.

Model	# Param.	Modality	Granularity	FUNSD (F1↑)	CORD (F1↑)	SROIE (F1↑)	RVL-CDIP (Acc↑)
$\text{BERT}_{\text{base}}$ [7]	110M	T	Word	60.26	89.68	90.99	89.91
$\text{RoBERTa}_{\text{base}}$ [22]	125M	T	Word	66.48	93.54	-	-
$\text{UniLMv2}_{\text{base}}$ [4]	125M	T	Word	66.48	90.92	90.06	
SelfDoc [21]	137M	T+I	Object	83.36	-	-	93.81
UniDoc [10]	272M	T+I	Object	87.93	**98.94**	-	95.05
$\text{TILT}_{\text{base}}$ [26]	230M	T+I	Word	-	95.11	-	95.25
$\text{DocFormer}_{\text{base}}$ [3]	183M	T+I	Grid/Patch	83.34	96.33	-	
$\text{LayoutLM}_{\text{base}}$ [30]	160M	T+I	Grid/Patch	79.27	-	94.38	94.42
$\text{LayoutLMv2}_{\text{base}}$ [29]	200M	T+I	Grid/Patch	82.76	94.95	96.25	95.25
$\text{LayoutLMv3}_{\text{base}}$ [13]	133M	T+I	Grid/Patch	90.29	96.56	-	**95.44**
$\text{WUKONG-READER}_{\text{base}}$	237M	T+I	Textline	**91.52**	96.54	**96.88**	94.91
$\text{BERT}_{\text{large}}$ [7]	340M	T	Word	65.63	90.25	92.00	89.81
$\text{RoBERTa}_{\text{large}}$ [22]	355M	T	Word	70.72	-	92.80	-
$\text{UniLMv2}_{\text{large}}$ [4]	355M	T	Word	72.57	82.05	94.88	90.20
$\text{TILT}_{\text{large}}$ [26]	780M	T+I	Word	-	96.33	98.10	95.52
$\text{StructuralLM}_{\text{large}}$ [2]	355M	T	Textline	85.14	-	-	96.08
$\text{LayoutLM}_{\text{large}}$ [30]	343M	T+I	Grid/Patch	78.95	94.93	95.24	94.43
$\text{LayoutLMv2}_{\text{large}}$ [29]	426M	T+I	Grid/Patch	84.20	96.01	97.81	95.64
$\text{LayoutLMv3}_{\text{large}}$ [13]	368M	T+I	Grid/Patch	92.08	**97.46**	-	95.93
$\text{ERNIE-Layout}_{\text{large}}$ [25]	-	T+I	Grid/Patch	93.12	97.21	97.55	**96.27**
$\text{WUKONG-READER}_{\text{large}}$	470M	T+I	Textline	**93.62**	**97.27**	**98.15**	95.26

Table 1: The entity-level F1 scores for information extraction on form (FUNSD) and receipt understanding (CORD and SROIE), and accuracies on the document classification task (RVL-CDIP). "T" and "I" refer to the text and image modality, respectively.

on various document understanding tasks: information extraction and document classification in Section 4.2, and document visual-question answering in Appendix B.1. We implement WUKONG-READER based on MindSpore [1].

Compared Methods. We compare WUKONG-READER against the following methods with different granularities: (i) Word-level features: BERT [7] and RoBERTa [22] adopt the conventional masked-language modeling objective over words. LayoutLM [30] and TILT [26] obtains words' bounding boxes from OCR and add them to the paired text embeddings. (ii) Grid/patch-level features: LayoutLMv2 [29] and DocFormer [3] extract image grid features with a CNN backbone, and LayoutLMv3 uses ViT [8] to encode image patches; (iii) Object-level features: SelfDoc [21] and UniDoc [10] concatenate text embeddings with region features from object detectors; and (iv) Textline-level features: StructuralLM [2] first leverages the cell-level layout information, the most similar to our textline-level features. However, they do not explicitly encode visual features, but only use this cell-level information of texts.

Pre-training. Following previous studies [30, 29], we adopt the IIT-CDIP Test Collection dataset [17] for pre-training, which contains 11M document images from various industrial domains. We extract the texts and bounding boxes using our internal OCR tool. We use 64 AI processors for pre-training, and the batch size of 24 per device. We use the Adam optimizer [16]. The learning rate is linearly warmed up to 1e-4 within the first 10% iterations, and then linearly decayed to 0. The weight decay is set as 1e-2. To save running memory we also enable gradient checkpointing [5] and FP16 training. We conduct pre-training for 10 epochs, which takes around 3 days and 5 days on 64 processors respectively.

4.2 Main Results

4.2.1 Information Extraction.

Datasets and Evaluation Metric. For information extraction, we evaluate over three datasets: FUNSD [15], CORD [24], and SROIE [14]. Following [30, 29, 13], we build a token classification layer on top of the multi-modal encoder, and predict the BIO tags for each entity field for FUNSD, CORD and SROIE. The weighted F1 score is

Pre-training Objectives	FUNSD	CORD	SORIE	RVL-CDIP
MLM	89.70	93.48	97.23	92.67
MLM+MRM	91.97(+2.27)	96.84(+3.36)	97.64(+0.41)	94.36(+1.69)
MLM+MRM+TRC	92.81(+0.84)	97.16(+0.32)	97.64(+0.00)	94.47(+0.11)
MLM+MRM+TRC+TGM	**93.62**(+0.81)	**97.27**(+0.11)	**98.15**(+0.51)	**95.26**(+0.79)

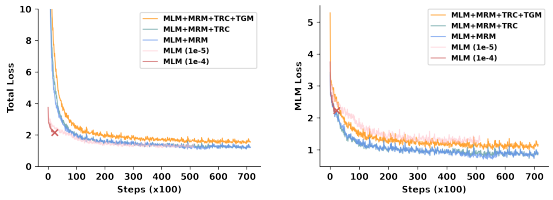
Table 2: Ablation study on the pre-training objectives with WUKONG-READER$_{\text{large}}$. All models are pre-trained for 10 epochs, and the fine-tuning settings are consistent with Table 1. The subscript numbers in the brackets represent the relative improvement with the ablated objectives.

used as the evaluation metric. Following StructuralLM [2] and LayoutLMv3 [13], we use the cell bounding box of each token in substitution of word bounding boxes. Similar to LayoutLMv2 [29], we use entity-level F1 score on SROIE, and correct OCR mismatch as the official OCR annotations are inconsistent with the test set provided by the official evaluation site. More details of these datasets can be found in Appendix A.

Results. According to Table 1, our model generally outperforms existing baselines on both model scales. Specifically, we achieve 91.52 and 93.62 weighted F1 score on FUNSD for WUKONG-READER$_{\text{base}}$ and WUKONG-READER$_{\text{large}}$, respectively. Both results are 1.23 to 1.56 points higher than LayoutLMv3, the previous SOTA models on document understanding. On CORD, our models also achieve comparable performances to state-of-the-art methods like LayoutLMv3. For SROIE, we again lead the performance with 96.88 and 98.15 weighted F1 scores on the base and large model, superior to LayoutLMv2 by 0.63 and 0.34 points, respectively.

4.2.2 Document Classification.

Datasets and Evaluation Metric. For document classification, we use the RVL-CDIP dataset [12]. RVL-CDIP contains around 400K industrial document images in 16 classes, such as forms, advertisement, letters, e.t.c. Following [29], we use the pre-encoder and post-encoder visual features, together with the [CLS] token of the multi-modal encoder for document classification. By default, we perform fine-tuning for 10 epochs over 8 computing processors, with the batch size of 24 per processor. The classification accuracy is used for evaluation. We set the learning rate to 5e-5 with the same scheduler to pre-training, and the weight decay is 1e-2.



(a) Total loss. (b) MLM loss.

Figure 3: The training curves in terms of total loss and MLM loss for pre-training with different training objectives.

Results. From the last column in Table 1, our WUKONG-READER$_{\text{base}}$ and WUKONG-READER$_{\text{large}}$ achieve 94.91% and 95.26% accuracies on RVL-CDIP, respectively. The results are competitive among the baseline models and have space for further improvement.

4.3 Discussions

Ablation Study of Training Objectives. We provide a comprehensive study on the effect of the different pre-training objectives on WUKONG-READER$_{\text{large}}$ over each downstream dataset. To better understand how these proposed objectives affect visual document understanding, we compare with the following settings: (i) the MLM objective; and (ii) the MLM and MRM objectives; and (iii) the MLM, MRM and TRC objectives; and (iv) the MLM, MRM, TRC and TGM objectives.

From Table 2, it can be found that training with only MLM objective leads to a significant performance drop. When MRM is used, the performance of each task is consistently improved, e.g., 2.27 and 3.36 F1 scores on FUNSD and CORD, respectively. Moreover, the TRC objective enhances the fine-grained visual and textual representation learning, and further improves the F1 score of FUNSD by 0.84. Finally, the TGM objective can further boost the performance of sequence labeling tasks, improving the F1 score by 0.81 on FUNSD.

Further Analysis of MRM. We visualize the training curves of both total loss and MLM loss in Figure 3(a) and Figure 3(b). It can be found that with only the MLM objective, the training fails as a result of NaN errors at early training steps, as indicated by the red $\times$. Thus we have to lower the learning rate to 1e-5 to finish the pre-training. However, when armed with MRM loss, the training stabilize and the overall process can be easily finished with a larger learning rate of 1e-4. We hypothesize that the enhanced visual features can help stabilize the pre-training. In addition, the MRM objective significantly improves the task performance. We notice that even only using self-reconstruction losses such as MLM and MRM, the pre-trained model can still achieve a relatively good performance. It shows the self-reconstruction objective on each separate modality serves to facilitate the implicit cross-modal interaction.

Visualization of TRC. We also study the WUKONG-READER's capability of capturing fine-grained cross-modal localization. We use the WUKONG-READER$_{\text{large}}$ model, and visualize the textline-region alignment in Figure 4, where the green and red boxes denote the correctly and incorrectly aligned pairs. Specifically, we perform the visualization similarly as [32], and compute the textline-region alignment based on the textline-wise similarity between the image regions and textlines. Note that only the dual-stream encoders are used to compute this similarity. It can be found that WUKONG-READER automatically learns to align the textline with its corresponding regions, with above 80+% accuracies across various kinds of document images. The learned alignment between two modalities implicitly explains the powerful effect of WUKONG-READER in various downstream tasks. This ability of WUKONG-READER provides a promising multimodal solution towards document localization tasks, instead of using naive text matching based on OCR results.

5 Conclusion

In this paper, we propose WUKONG-READER, a multi-modal pre-trained model for fine-grained visual document understanding. Unlike existing solutions that ignore the intrinsic textual segment information, our WUKONG-READER aims to leverage the semantics in textline regions of documents, by aligning with the visual and textual contents over document textlines via textline-region contrastive learning. Meanwhile, we also propose masked region modeling and textline grid matching to further enhance the visual and layout information of document textlines. We evaluate WUKONG-READER on various visual document understanding tasks such as information extraction and document classification, and the proposed model demonstrates superior performance against previous counterparts.

(a) Align Acc = 83.3%. (b) Align Acc = 83.9%.

Figure 4: Visualization of learned textline-region alignment. The green and red textline bounding boxes denote the correct and incorrect alignment, respectively.

References

[1] Mindspore. https://www.mindspore.cn.

[2] 2021. StructuralLM: Structural pre-training for form understanding, author = "li, chenliang and bi, bin and yan, ming and wang, wei and huang, songfang and huang, fei and si, luo. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistic*, pages 6309–6318.

[3] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.

[4] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

[5] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

[6] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[9] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*.

[10] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

[11] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4583–4592.

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

[13] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4083–4091.

[14] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition*, pages 1516–1520. IEEE.

[15] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops*, volume 2, pages 1–6. IEEE.

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

[17] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson 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*, pages 665–666.

[18] 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*, page 3530–3539.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

[20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705.

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

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5652–5660.

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

[23] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

[24] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

[25] Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

[26] 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. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[28] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744.

[29] 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*, pages 2579–2591.

[30] 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

[31] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

[32] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*.

A Downstream Datesets

FUNSD [15] consists of noisy scanned documents and aims at understanding the structure of textual content of forms. It contains 199 fully labelled real scanned images, including 149 training samples and 50 test documents. We follow [29] to use the entity-level F1 to evaluate the model performance.

CORD [24] is a consolidated dataset for receipt parsing. CORD collected over 11,000 Indonesian receipt images from shops and restaurants. The dataset comprises 800, 100, and 100 receipt samples for training, validation, and testing. We adopt entity-level F1 and transcript of CORD for training and evaluation.

SROIE [14] contains 1000 scanned receipt images for text recognition and key information extraction. SROIE annotated 626 and 347 receipts for training and test, respectively. The dataset labelled four entities: company, date, address, and total. We correlate the entity annotation files with OCR results to generate ground-truth BIO labels for training and testing. During inference, we extract entities according to BIO labeling results and employ the entity-level F1 for evaluation. We use the official OCR annotations, however which contain OCR mismatch and are inconsistent with test set provided by the official evaluation site. Therefore, LayoutLMv2 [29] and other top methods on SROIE leaderboard[3] claim to exclude OCR mismatch and fix total entities. We thus follow the same evaluation protocol as these methods to correct OCR mismatch via post-processing on entities.

RVL-CDIP [12] contains around 400K industrial document images in 16 classes, such as forms, advertisements, and letters, among which 360K and 40K are selected for training and testing. We extract text and layout information using Huawei-developed text recognition algorithms. We use the overall classification accuracy as the evaluation metric. We use the official OCR annotations, however which are inconsistent with test set provided by the official evaluation site. We thus follow LayoutLMv2 [29] to post-process extracted entities and correct OCR mismatch.

DocVQA [23] contains 50,000 manually designed questions over 12,767 industrial document images. These scanned documents include various categories: figure/diagram, form, table/list, layout, free text, image/photo, handwritten characters, yes or no and others. We use the Microsoft OCR tool to extract the text and their bounding boxes. We also re-organize the OCR recognized text based on reading order of human, i.e., we heuristically cluster the word bounding box based on their intervals. This can be beneficial for documents with irregular layouts. For instance, reading from left to right in double column documents may fail to produce natural text.

[3] https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=3

Model	ANLS
LayoutLMv2$_{\text{base}}$	78.0
LayoutLMv2$_{\text{base}}$*	74.0
WUKONG-READER$_{\text{base}}$	73.7
WUKONG-READER$_{\text{large}}$	78.9

Table 3: Results on the DocVQA dataset.

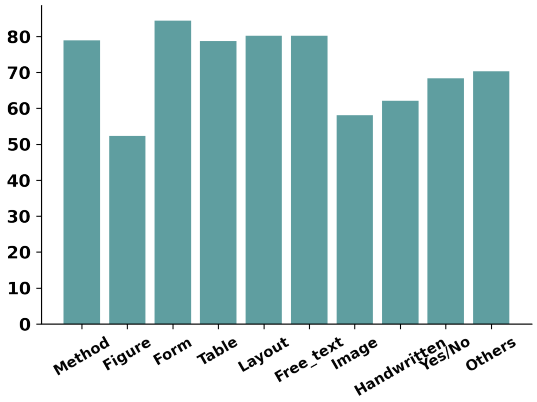


Figure 5: The ANLS scores of each category in DocVQA achieved by WUKONG-READER$_{\text{large}}$.

B More Experiments

B.1 Document Question Answering

Datasets and Evaluation Metric. For document question answering, we use the DocVQA dataset [23], which contains 50,000 questions over 12,000 pages of various industrial documents. We use the official website for evaluation[4], which compares the extracted answer span with the ground-truth and reports the averaged normalized Levesitein distance (ANLS).

Results. The results on DocVQA are listed in Table 3. For LayoutLMv2-base [29], we report the

[4] https://rrc.cvc.uab.es/?ch=17&com=introduction

best reproduced result marked as *. As suggested by existing methods [29], leveraging the additional techniques of post-processing, data augmentation and model ensemble contributes a lot to this performance, while we leave this exploration to the future work. Overall, our WUKONG-READER$_{base}$ and WUKONG-READER$_{large}$ achieve 73.7 and 78.9 ANLS score, respectively. This is comparable to the competitive LayoutLMv2 without using additional techniques. For instance, LayoutLMv2 is initialized from UniLMv2 [4] that naturally owns a more powerful question answering ability than RoBERTa. Unfortunately, we are unable to access UniLMv2 model since it is not publicly released yet and thus our model was initialized from RoBERTa. We also visualize the ANLS score of each class in DocVQA returned by our WUKONG-READER$_{large}$ in Figure 5. It can be found that our model can perform reasonably well on "Form" and "Layout" with around 80.0 ANLS scores, yet there is still room for improvement for categories such as "Figure" and "Image".