Unifying Vision, Text, and Layout for Universal Document Processing

Zineng Tang[1,2] Ziyi Yang[2*] Guoxin Wang[3] Yuwei Fang[2] Yang Liu[2]
Chenguang Zhu[2] Michael Zeng[2] Cha Zhang[3] Mohit Bansal[1*]
[1]University of North Carolina at Chapel Hill
[2]Microsoft Azure Cognitive Services Research
[3]Microsoft Azure Visual Document Intelligence

Abstract

We propose Universal Document Processing (UDOP), a foundation Document AI model which unifies text, image, and layout modalities together with varied task formats, including document understanding and generation. UDOP leverages the spatial correlation between textual content and document image to model image, text, and layout modalities with one uniform representation. With a novel Vision-Text-Layout Transformer, UDOP unifies pretraining and multi-domain downstream tasks into a prompt-based sequence generation scheme. UDOP is pretrained on both large-scale unlabeled document corpora using innovative self-supervised objectives and diverse labeled data. UDOP also learns to generate document images from text and layout modalities via masked image reconstruction. To the best of our knowledge, this is the first time in the field of document AI that one model simultaneously achieves high-quality neural document editing and content customization. Our method sets the state-of-the-art on 8 Document AI tasks, e.g., document understanding and QA, across diverse data domains like finance reports, academic papers, and websites. UDOP ranks first on the leaderboard of the Document Understanding Benchmark.[1]

1. Introduction

Document Artificial Intelligence studies information extraction, understanding, and analysis of digital documents, e.g., business invoices, tax forms, academic papers, etc. It is a multimodal task where text is structurally embedded in documents, together with other vision information like symbols, figures, and style. Different from classic vision-language research, document data have a 2D spatial layout: text content is structurally spread around in different locations based on diverse document types and formats (e.g., invoices vs. tax forms); formatted data such as figures, tables and plots are laid out across the document. Hence, effectively and efficiently modeling and understanding the layout is vital for document information extraction and content understanding, for example, title/signature extraction, fraudulent check detection, table processing, document classification, and automatic data entry from documents.

Document AI has unique challenges that set it apart from other vision-language domains. For instance, the cross-modal interactions between text and visual modalities are much stronger here than in regular vision-language data, because the text modality is visually-situated in an image. Moreover, downstream tasks are diverse in domains and paradigms, e.g., document question answering [45], layout detection [57], classification [13], information extraction [28], etc. This gives rises to two challenges: (1) how to utilize the strong correlation between image, text and layout modalities and unify them to model the document as a whole? (2) how can the model efficiently and effectively learn diverse vision, text, and layout tasks across different domains?

There has been remarkable progress in Document AI in recent years [1, 10–12, 15, 16, 24, 26, 29, 30, 36, 37, 48, 52–55]. Most of these model paradigms are similar to traditional vision-language frameworks: one line of work [1, 11, 29, 30, 36, 37, 52–55] inherits vision-language models that encode images with a vision network (e.g., vision transformer) and feed the encodings to the multimodal encoder along with text [17, 27, 44, 47]; another line of work uses one joint encoder [22, 46] for both text and image [16]. Some models regard documents as text-only inputs [10, 12, 15, 26, 48]. In these works, the layout modality is represented as shallow positional embeddings, e.g., adding a 2D positional embedding to text embeddings. The strong correlation between modalities inherent in document data are not fully exploited. Also to perform different tasks, many models have to use task-specific heads, which is inefficient and requires manual design for each task.

To address these challenges, we propose Universal Docu-

[*] Corresp. authors: ziyiyang@microsoft.com, mbansal@cs.unc.edu

[1] Code and models: https://github.com/microsoft/i-Code/tree/main/i-Code-Doc

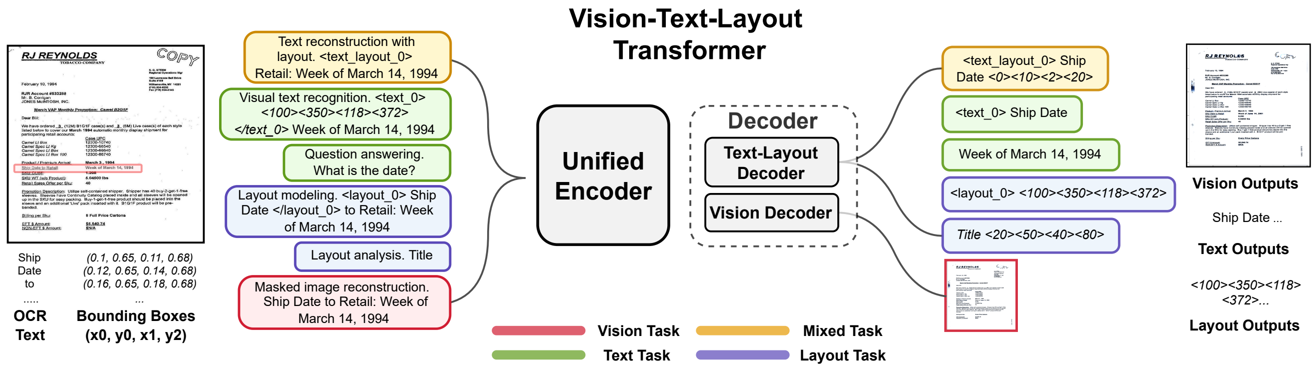

Vision-Text-Layout Transformer
RJ REYNOLDS
COPY
Ship
Date
to
.....
(0.1, 0.65, 0.11, 0.68)
(0.12, 0.65, 0.14, 0.68)
(0.16, 0.65, 0.18, 0.68)
...
OCR Text
Bounding Boxes (x0, y0, x1, y2)
Text reconstruction with layout. <text_layout_0> Retail: Week of March 14, 1994
Visual text recognition. <text_0> <100><350><118><372> </text_0> Week of March 14, 1994
Question answering. What is the date?
Layout modeling. <layout_0> Ship Date </layout_0> to Retail: Week of March 14, 1994
Layout analysis. Title
Masked image reconstruction. Ship Date to Retail: Week of March 14, 1994
Unified Encoder
Decoder
Text-Layout Decoder
Vision Decoder
<text_layout_0> Ship Date <0><10><2><20>
<text_0> Ship Date
Week of March 14, 1994
<layout_0> <100><350><118><372>
Title <20><50><40><80>
Vision Outputs
Ship Date ...
Text Outputs
<100><350><118> <372>...
Layout Outputs
Vision Task
Mixed Task
Text Task
Layout Task

Figure 1. UDOP unifies vision, text, and layout through vision-text-layout Transformer and unified generative pretraining tasks including vision task, text task, layout task, and mixed task. We show the task prompts (left) and task targets (right) for all self-supervised objectives (joint text-layout reconstruction, visual text recognition, layout modeling, and masked autoencoding) and two example supervised objectives (question answering and layout analysis).

ment Processing (UDOP), a foundation Document AI model that unifies vision, text, and layout and different document tasks. Different from regarding image and document text as two separate inputs in previous works, in UDOP we propose to model them with the uniform layout-induced representation (Sec. 3.1): in the input stage, we add embeddings of text tokens with the features of the image patch where the tokens are located. This simple and novel layout-induced representation greatly enhances the interaction between the text and vision modalities.

Besides the layout-induced representation, to form a uniform paradigm for different vision, text, layout tasks, UDOP first builds a homogeneous vocabulary for texts and document layout that converts layout, i.e. bounding boxes, to discretized tokens. Second, we propose Vision-Text-Layout (VTL) Transformer, consisting of a modality-agnostic encoder, text-layout decoder and vision decoder. VTL Transformer allows UDOP to jointly encode and decode vision, text, and layout. UDOP unites all downstream tasks with a sequence-to-sequence generation framework.

Besides the challenges of modality unification and task paradigms, another issue is previous works utilized self-supervised learning objectives that were originally designed for single-modality learning, e.g., masked language modeling, or classical vision-language pretraining, e.g., contrastive learning. We instead propose novel self-supervised learning objectives designed to allow holistic document learning, including layout modeling, text and layout reconstruction, and vision recognition that account for text, vision and layout modeling together (Sec. 4). Besides sequential generation, UDOP can also generate vision documents by leveraging masked autoencoders (MAE) [14] by reconstructing the document image from text and layout modalities. With such generation capacity, UDOP is the first document AI model to achieve high-quality customizable, joint document editing and generation.

Finally, our uniform sequence-to-sequence generation framework enables us to conveniently incorporate all major document supervised learning tasks to pretraining, i.e., document layout analysis, information extraction, document classification, document Q&A, and Table QA/NLI, despite their significant differences in task and data format. In contrast, pretraining in previous document AI works is constrained to unlabeled data only (or using one single auxiliary supervised dataset such as FUNSD [55]), while abundant labeled datasets with high quality supervision signals are ignored due to the lack of modeling flexibility. Overall, UDOP is pretrained on 11M public unlabeled documents, together with 11 supervised datasets of 1.8M examples. Ablation study in Table 4 shows that UDOP only pretrained with the proposed self-supervised objectives exhibits great improvements over previous models, and adding the supervised data to pretraining further improves the performance.

We evaluate UDOP on FUNSD [18], CORD [34], RVL-CDIP [13], DocVQA [33], and DUE-Benchmark [2]. UDOP ranks the 1st place on the DUE-Benchmark leaderboard with 7 tasks, and also achieves SOTA on CORD, hence making UDOP a powerful and unified foundation Document AI model for diverse document understanding tasks,

To summarize, our major contributions include:

1. Unified representations and modeling for vision, text and layout modalities in document AI.

2. Unified all document tasks to the sequence-to-sequence generation framework.

3. Combined novel self-supervised objectives with supervised datasets in pretraining for unified document pretraining.

4. UDOP can process and generate text, vision, and layout modalities together, which to the best of our knowledge is first one in the field of document AI.

5. UDOP is a foundation model for Document AI, achieving SOTA on 8 tasks with significant margins.

2. Related Work

Unifying Model Architectures in Multimodal Learning. Unifying model architectures for different modalities, such as vision, language, and speech, is an emergent direction. Inspired by the immense success in natural language processing, computer vision and speech processing, model architectures in multimodal learning is converging to Transformers. One type of works concatenates text token embeddings and projected image patches as the input [6, 42] to a multimodal Transformer. Other models uses two-tower or three-tower architecture where each modality is encoded respectively. Projection heads or fusion networks on top of the two-tower architecture generate multimodal representations [38, 56].

Unifying Tasks with the Generative Framework. Research on unifying training processes across different tasks and domains recently has made significant progress. [8] finetunes language models with instructions on 1.8k tasks. [7] unifies several vision-language tasks by converting training objectives to sequence generation. [31, 49, 50] further combines more tasks, e.g., image generation, by converting images and bounding boxes to discrete tokens.

Document Artificial Intelligence. LayoutLM [53] pretrains BERT models on document data with masked language modeling and document classification task, with 2D positional information and image embeddings integrated. Subsequent works [15, 16, 55] also adopt VL-BERT alike architecture and includes additional pretraining tasks, e.g., masked image/region modeling proposed, and leverages the reading order in layout information [12]. [11, 29] use a multimodal encoder to model region features extracted by CNN with sentence-level text representations and train with self-supervised objectives. [20] proposes an OCR-free model to directly generate textual output from document images. [36] trains generative language models on both unlabeled and labeled document data using generative training objectives. [10] proposed to model documents as collections of tokens bounding boxes.

3. Universal Document Processing

We introduce UDOP, a novel document AI framework with unified learning objectives and model architecture for text, vision, and layout as shown in Figure 1. In this section, we will concretely discuss the proposed Vision-Text-Layout Transformer in UDOP, and will introduce the unified generative pretraining method in the next section. In document processing, given a document image $\boldsymbol{v}$, typically optical character recognition (OCR) is used on $\boldsymbol{v}$ to extract text tokens $\{s_i\}$ in the document and their bounding boxes $\{(x_i^1, y_i^1, x_i^2, y_i^2)\}$, i.e., the layout information for each token. (x_i^1, y_i^1) and (x_i^2, y_i^2) respectively represent the coordinates of the left-upper and right-bottom corner of the bounding box. Thus, suppose we have M word tokens, the input is the triple, $(\boldsymbol{v}, \{s_i\}_{i=1}^M, \{(x_i^1, y_i^1, x_i^2, y_i^2)\}_{i=1}^M)$. Figure 1 shows an example document (left) and downstream tasks (right).

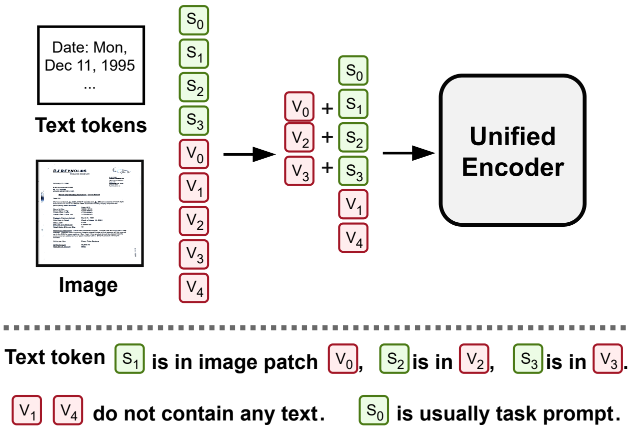

Date: Mon,
Dec 11, 1995
...
Text tokens
Image
S0
S1
S2
S3
V0
V1
V2
V3
V4
V0 + S1
V2 + S2
V3 + S3
S0
V1
V4
Unified Encoder
Text token S1 is in image patch V0, S2 is in V2, S3 is in V3.
V1 V4 do not contain any text.
S0 is usually task prompt.

Figure 2. Layout-induced vision-text embedding.

3.1. A Unified Vision, Text, and Layout Encoder

We fuse the vision, text, and layout modalities in the input stage using one unified transformer encoder. For traditional vision-text data, the text modality is usually the high-level description of the corresponding image or task prompt (e.g., question). While in document images, text is embedded inside the image, i.e., text and image pixels have one-to-one correspondence. To leverage this correspondence, we propose a new Vision-Text-Layout (VTL) Transformer architecture to dynamically fuse and unite the image pixels and text tokens based on the layout information.

Concretely, given the document image $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$, M word tokens $\{s_i\}_{i=1}^M$ inside the image and the extracted layout structure $\{(x_i^1, y_i^1, x_i^2, y_i^2)\}_{i=1}^M$, we first partition $\boldsymbol{v}$ into $\frac{H}{P} \times \frac{W}{P}$ image patches, where each patch is of size $P \times P \times C$. We then encode each patch with a D-dim vector and group all patch embeddings into a sequence of vectors $\{\boldsymbol{v}_i \in \mathbb{R}^D\}_{i=1}^N$ where $N = \frac{H}{P} \times \frac{W}{P}$. Text tokens are also converted to numerical D-dim embeddings $\{\boldsymbol{s}_i\}_{i=1}^M$ by vocabulary look-up.

Layout-Induced Vision-Text Embedding. Next, we build a unified representation for vision, text, and layout as shown in Figure 2. We define the layout indicator function ϕ of image patch and token embeddings as follows:

$$\phi(\boldsymbol{s}_i, \boldsymbol{v}_j) = \begin{cases} 1, & \text{if the center of } \boldsymbol{s}_i\text{'s bounding box} \\ & \text{is within the image patch } \boldsymbol{v}_j. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Then for each text token embedding $\boldsymbol{s}_i$, the joint representation is the sum of its image patch feature[2] and the text

[2] Some text token like manually crafted prompts have no locations. So, we set their layout bounding boxes to be $(0, 0, 0, 0)$, i.e., they fall into a pseudo image patch.

feature:

$$\boldsymbol{s}'_i = \boldsymbol{s}_i + \boldsymbol{v}_j, \text{ where } \phi(\boldsymbol{s}_i, \boldsymbol{v}_j) = 1.$$

For image patches $\boldsymbol{v}_j$ without any text tokens, i.e. $\forall i,\ \phi(\boldsymbol{s}_i, \boldsymbol{v}_j) = 0$, the joint representation, $\boldsymbol{v}'_j$ is itself:

$$\boldsymbol{v}'_j = \boldsymbol{v}_j.$$

Note we do not have a designated joint representation for image patch containing tokens, since features of these image patches are already integrated with the text embeddings. Then $\{\boldsymbol{s}'_i\}$ and $\{\boldsymbol{v}'_j\}$ are fed into the VTL transformer encoder. These joint representations greatly enhance the interaction between vision, text and layout in the model input stage by explicitly leveraging their spatial correlations.

To further unify layout and text representation, inspired by the recent progress in generative object detection [4, 49], we discretize the layout modality, i.e., continuous coordinates text bounding box, to layout tokens. Suppose we have bounding box $(x^1_i, y^1_i, x^2_i, y^2_i)$ normalized in $[0, 1]$. The resulting layout token will be each coordinate multiplied by vocabulary size and then rounded to nearest integer. For example, if we have bounding box $(0.1, 0.2, 0.5, 0.6)$ with layout vocabulary size 500, the layout tokens will then be *<50><100><250><300>*. Layout tokens can be conveniently inserted into text context, and elegantly used for layout generation tasks (e.g., location detection). More details are discussed in Section 4.

Position Bias. We follow TILT [36] to encode 2D text token position as 2D relative attention bias, similar to the relative attention bias used in T5. However, unlike T5, TILT, or transformer models in previous Document AI works [16, 36], we do not use 1D position embeddings in VTL transformer encoder, since the joint embedding and the 2D position bias already incorporate the layout structure of the input document.

3.2. Vision-Text-Layout Decoder

As introduced in the previous section, the VTL encoder is able to compactly and jointly encode vision, text, and their layout. To perform various document generative tasks (will be discussed in Section 4), the VTL decoder is designed to jointly generate all vision, text, and layout modalities.

The VTL decoder consists of a text-layout decoder and a vision decoder, as shown in Figure 1 (middle). The text-layout decoder is a uni-directional Transformer decoder to generate text and layout tokens in a sequence-to-sequence manner. For the vision decoder, we adopt the decoder of MAE [14] and directly generate the image pixels with text and layout information. Details of the image decoding process will be discussed in the segment **"Masked Image Reconstruction with Text and Layout "** of Section 4.1. Both text-layout decoder and vision decoder will cross-attend to the VTL encoder.

Information such as model configurations are presented in Section 5.1.

4. Unified Generative Pretraining

To unify across different training objectives and datasets, we create a universal generative task format with task prompt. We pretrain UDOP on large-scale documents with and without human labels. We summarize the tasks prompts and targets in Table 1 which includes all self-supervised and supervised tasks respectively in upper and lower blocks.

4.1. Self-Supervised Pretraining Tasks

We propose various innovative self-supervised learning objectives for unlabeled documents. The unlabeled document contains OCR text inputs with token-level bounding boxes and the document image. In the rest of this subsection, we use the following input text as example:
"Ship Date to Retail: Week of March 14, 1994"

(1) Joint Text-Layout Reconstruction requires the model to reconstruct the missing texts and locate them in the document image. Concretely, we mask a percentage of text tokens and ask the model to both the tokens and their bounding boxes (i.e. layout tokens). E.g., assume masking "Ship Date" and "of", the input sequence and target sequence is given below:

Input Sequence:
"*Joint Text-Layout Reconstruction*. <text_layout_0> to Retail: Week <text_layout_1> March 14, 1994"

Target Sequence:
"<text_layout_0> Ship Date *<100><350><118><372>* <text_layout_1> of *<100><370><118><382>*"

Here <text_layout_0> and <text_layout_1> denote the text-layout sentinel tokens, *<100><350><118><372>* and *<100><370><118><382>*" represent the layout tokens of "Date to" and "of" respectively. We use masking ratio 15% similar to Masked Language Modeling (MLM) [9] as this task can be interpreted as masked text-layout modeling.

(2) Layout Modeling asks the model to predict positions of (group of) text tokens, given the document image and context text. E.g., to predict positions of "Ship Date" and "of", the input sequence and target sequence is given below:

Input Sequence:
"*Layout Modeling*. <layout_0> Ship Date </layout_0> to Retail: Week <layout_1> of </layout_1> March 14, 1994"

Target Sequence:
"<layout_0> *<100><350><118><372>* <layout_1> *<100><370><118><382>*"

Table 1. A summary of all generative pretraining objectives with task names, task prompts, and task targets.

Self-Supervised Tasks	**Task Prompts**	**Task Targets**
Layout Modeling	*Layout Modeling*. <layout_0> Ship Date to Retail </layout_0> Week of March 14, 1994	<layout_0> *<100><350><118><372>*
Visual Text Recognition	*Visual Text Recognition*. <text_0> *<100><350><118><372>* </text_0> to Retail: Week of March 14, 1994	<text_0> Ship Date
Joint Text-Layout Reconstruction	*Joint Text-Layout Reconstruction*. <text_layout_0> to Retail: Week of March 14, 1994	<text_layout_0> Ship Date *<100><350><118><372>*
Masked Image Reconstruction	*Masked Image Reconstruction*. Ship Date to Retail: Week of March 14, 1994	[Pixels of the original image]
Supervised Tasks		
Classification	*Document Classification*. Ship Date to Retail: Week of March 14, 1994	Memo.
Layout Analysis	*Layout Analysis*. Paragraph.	Paragraph *<82><35><150><439>*
Information Extraction	*Information Extraction*. Ship Date to Retail	Week of March 14, 1994
Question Answering	*Question Answering*. What is the ship year?	1994
Document NLI	*Document Natural Language Inference*. Ship Date to Retail: Week of March 14, 1994	Entailment.

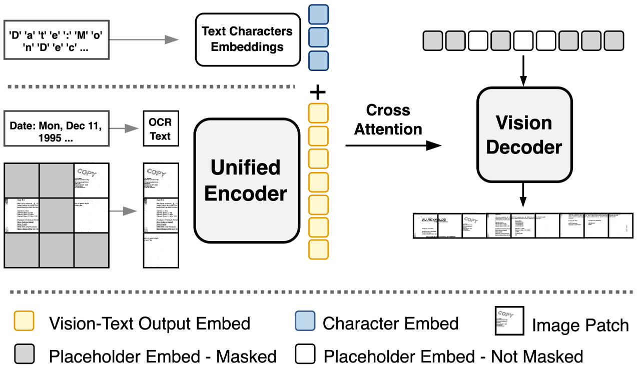

'D' 'a' 't' 'e' ':' 'M' 'o' 'n' 'D' 'e' 'c' ...
Text Characters Embeddings
+
Date: Mon, Dec 11, 1995 ...
OCR Text
COPY
Unified Encoder
Cross Attention
Vision Decoder
Vision-Text Output Embed
Character Embed
Image Patch
Placeholder Embed - Masked
Placeholder Embed - Not Masked

Figure 3. Masked autoencoding with text and layout.

Note this pretraining task has a different sentinel token, <layout_sent_0>, from the previous task "Joint Text-Layout Reconstruction" because the generation content is different (layout vs. text + layout). We use large masking ratio 75% since masking with small ratio results in an easy task.

(3) Visual Text Recognition identifies text at given location in the image. E.g., to recognize the text tokens at *<100><350><118><372>* and *<100><370><118><382>*, the input and target is:

Input Sequence:
"*Visual Text Recognition*. <text_0> *<100><350><118><372>* </text_0> to Retail: Week <text_1> *<100><370><118><382>* </text_1> March 14, 1994"

Target Sequence:
"<text_0> Ship Date <text_1> of"

Note this pretraining task also has a different sentinel token, <text_0> . We use masking ratio 50% to distinguish this task from "Joint Text-Layout Reconstruction" and set the layout (bounding box) of sentinel token, e.g., <text_0>, and layout token, e.g., *<0><10><2><20>*, to (0,0,0,0). This objective helps model learn joint vision-text embedding by understanding vision-text correspondence.

(4) Masked Image Reconstruction with Text and Layout aims to reconstruct image with text and layout as shown in Figure 3. We adopt the MAE objective [14] for vision self-supervised learning. Originally, MAE masks a percentage of the image patches and feed non-masked patches into a vision encoder. It then feeds encoder outputs to a vision decoder to reconstruct masked patches. MAE uses mean squared error and apply loss only on masked patches. We make the following modifications to the MAE decoding process to customize it for document image generation and our task unification framework:

(4.a) Cross-Attention with Character Embeddings. In document, the textual content mostly consists of alphabetic characters, numbers and punctuation. The character-level composition of text tokens should be helpful for the vision generation. We add cross-attention in the vision decoder that it attends to both the text token encoder features and embeddings of characters in the token (Figure 3 left upper). These characters embeddings are trainable parameters and not encoded by the encoder. This cross-attention with characters only adds linear computation complexity but considerably improves the image generation quality.

(4.b) Image Decoding. Next, we describe the MAE decoding process. For UDOP, we cannot directly feed the unified encoder output to the vision decoder, since the joint vision-text embedding only contains non-masked image patches to the unified encoder (Section 3.1), and image patches are fused with text tokens. Therefore, we propose that the vision decoder takes in a sequence of trainable placeholder embeddings. The length and order of the placeholder sequence is same as the patches of target image. We use two types of placeholder embeddings to indicate whether the image patch is masked in the input document image. The vision decoder attends to encoder vision-text output AND character embeddings via cross-attention. The above process is illustrated in Figure 3. We show the high quality generation visualization in Section 6.1.

4.2. Supervised Pretraining Tasks

Self-supervised tasks leverage large-scale unlabeled data to learn robust representations. On the other hand, supervised tasks use labeled data for fine-grained model supervision. We include the following supervised tasks in pretraining: document classification, layout analysis, information extraction, question answering, and document natural language inference. Details of the following supervised dataset are in Appendix E. Note that we do not conduct self-supervised tasks on the supervised datasets since we already have large-scale and diverse unlabeled data. Note that the validation or test set of downstream tasks is not used in supervised pretraining.

Classification. The task is to predict the document type. The task prompt is "*Document Classification on (Dataset Name)*" like "*Document Classification on RVLCDIP*", then followed by text tokens. The target is the document class. We use RVL-CDIP [13] with 16 document categories.

Layout Analysis. This task is to predict locations of an entity in the document like title, paragraph, etc. The task prompt is "*Layout Analysis on (Dataset Name)*", then followed by the entity name. The target are all bounding boxes that cover the given entity. We use PubLayNet [57].

Information Extraction. This task predict the entity type and location of a text query (e.g., the abstract paragraph). The task prompt is "*Information Extraction on (Dataset Name) (Text Query)*". The target is the entity label and the bounding box of each token of the query. We use DocBank [28], Kleister Charity (KLC) [41], PWC [19], and DeepForm [43].

Question Answering. The task is to answer a given question associated with the document image. The task prompt is "*Question Answering on (Dataset Name)*", then followed by the question and all document tokens. The target is the answer. We use WebSRC [3], VisualMRC [45], DocVQA [33], InfographicsVQA [32], and WTQ (WikiTableQuestions) [35].

Document NLI. Document Natural Language Inference predicts the entailment relationship between two sentences in a document. The prompt is "*Document Natural Language Inference on (Dataset Name)*", then followed by the sentence pair. The target is the "Entailment" or "Not Entailment". We use TabFact [5] for this task.

5. Experimental Setup

5.1. Model Pretraining

Model Configuration. In UDOP, the unified encoder and text-layout decoder follows the encoder-decoder architecture of T5-large [39]. The vision decoder is MAE-large decoder [14]. Overall UDOP has 794M trainable parameters. For tokenizer, we use T5 tokenizer and embedding from Hugging Face Transformers [51]. We also extend the vocabulary to accommodate special tokens (e.g., new sentinel and layout tokens).

Data. For self-supervised learning, we use IIT-CDIP Test Collection 1.0 [25], a large-scale document collections commonly-used in previous works [16, 53, 55]. It contain 11 million scanned document with contains text and token-level bounding boxes extracted by OCR. Supervised datasets are as introduced in Section 4.2.

Curriculum Learning. We use large image resolution, 1024, in our final settings since low resolution makes document text unidentifiable for both detection and generation. It will result in $(1024/16)^2 = 4096$ image patch sequence length which takes longer training time than small image resolution, e.g., 224. Therefore, we use curriculum learning to start from a relatively small resolution and gradually scale up to 1024 resolution. In practice, we use scale with 3 resolutions during the pretraining 224 $\rightarrow$ 512 $\rightarrow$ 1024. We show the performance of the 3 stages in Appendix I.

Training. We use Adam [23] optimizer with learning rate 5e-5, 1000 warmup steps, batch size 512, weight decay of 1e-2, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. For each curriculum learning stage, we train for 1 epoch.

5.2. Downstream Evaluations

We report the results on FUNSD [18], CORD [34], RVL-CDIP [13], and DocVQA [33] in Table 3 and describe their respective settings in below. We also report the results on 7 datasets of DUE-Benchmark [2] in Table 2. Finetuning training details are available in Appendix E.6 and performance variance is available in Table 7 and Table 8. Note that for all downstream tasks, we use the original OCR annotations provided in the datasets. We include their details in Appendix C.

Results. Pretrained models are finetuned on each evaluation dataset. As shown in Table 2, our models UDOP achieve SOTA performance on all 7 tasks of DUE-Benchmark, ranking the 1st place on the leaderboard as of November 11, 2022. It also sets SOTA on CORD and (Table 3). It is worth noting that UDOP is an **open-vocabulary generative model** and uses **one single model for all tasks**. In comparison, most baselines leverage task-specific network for each dataset and are classification-based models. Nonetheless, UDOP still exhibits better results than those models.

Curriculum learning on image resolution (appendix Ta-

Table 2. Comparison with existing published models on the DUE-Benchmark. Modality T, L, V denote text, layout, or vision. Results with * are obtained by training with auxiliary data following TILT [36] (details are in Appendix H).

Model	Modality	Question Answering		Information Extraction			Table QA/NLI		Avg.
		DocVQA	InfoVQA	KLC	PWC	DeepForm	WTQ	TabFact	
Donut [21]	V	72.1	-	-	-	-	-	-	-
BERT$_{\text{large}}$ [9]	T	67.5	-	-	-	-	-	-	-
T5$_{\text{large}}$ [39]	T	70.4	36.7	74.3	25.3	74.4	33.3	58.9	50.7
T5$_{\text{large}}$+U [36]	T	76.3	37.1	76.0	27.6	82.9	38.1	76.0	56.5
T5$_{\text{large}}$+2D [36]	T+L	69.8	39.2	72.6	25.7	74.0	30.8	58.0	50.4
T5$_{\text{large}}$+2D+U [36]	T+L	81.0	46.1	75.9	26.8	83.3	43.3	78.6	59.8
LAMBERT [10]	T+L	-	-	81.3	-	-	-	-	-
StructuralLM$_{\text{large}}$ [26]	T+L	83.9	-	-	-	-	-	-	-
LayoutLMv2$_{\text{large}}$ [55]	V+T+L	78.8	-	-	-	-	-	-	-
LayoutLMv3$_{\text{large}}$ [16]	V+T+L	83.4	45.1	77.1	26.9	84.0	45.7	78.1	62.9
UDOP	V+T+L	**84.7 (87.8*)**	**47.4 (63.0*)**	**82.8**	**28.0**	**85.5**	**47.2**	**78.9**	**64.8**

Table 3. Performance on FUNSD, CORD, and RVL-CDIP datasets. Modality V, T, L denote vision, text and layout.

Model	Modality	Info Ext.		Classification
		FUNSD	CORD	RVL-CDIP
Donut [21]	V	-	91.6	95.3
BERT$_{\text{large}}$ [9]	T	65.63	90.25	89.92
BROS$_{\text{large}}$ [15]	T+L	84.52	97.40	-
StructuralLM$_{\text{large}}$ [26]	T+L	85.14	-	**96.08**
LiLT [48]	T+L	88.41	96.07	95.68
FormNet [24]	T+L	84.69	97.28	-
LayoutLM$_{\text{large}}$ [53]	T+L	77.89	-	91.90
SelfDoc [29]	V+T+L	83.36	-	92.81
UniDoc [11]	V+T+L	87.93	96.86	95.05
DocFormer$_{\text{large}}$ [1]	V+T+L	84.55	96.99	95.50
TILT$_{\text{large}}$ [36]	V+T+L	-	96.33	95.52
LayoutLMv2$_{\text{large}}$ [55]	V+T+L	84.20	96.01	95.64
LayoutLMv3$_{\text{large}}$ [16]	V+T+L	**92.08**	97.46	95.93
UDOP	V+T+L	91.62	**97.58**	96.00

ble 6) shows that with larger resolution, UDOP steadily gains stronger performance. E.g., UDOP average performance on DUE-Benchmark with 224, 512 and 1024 resolution is 63.9, 64.3 and 65.1 respectively. Note our model with 224 resolution already outperform previous best models (e.g., average 62.9 on DUE-Benchmark). We then train UDOP only with self-supervised objectives (224 resolution). Its performance (Table 4) also surpasses baselines, which shows the effectiveness of the unified representations, TVL transformer and the proposed self-supervised objectives.

6. Analysis

6.1. Visualization Analysis

Masked Image Reconstruction. Figure 5 presents masked image reconstruction. Even with high masking ratio, the model can reconstruct the document image from text and layout signals with high quality: reconstructed contents are clear, consistent, and almost identical with the original image (all demonstrations are conducted on unseen documents.).

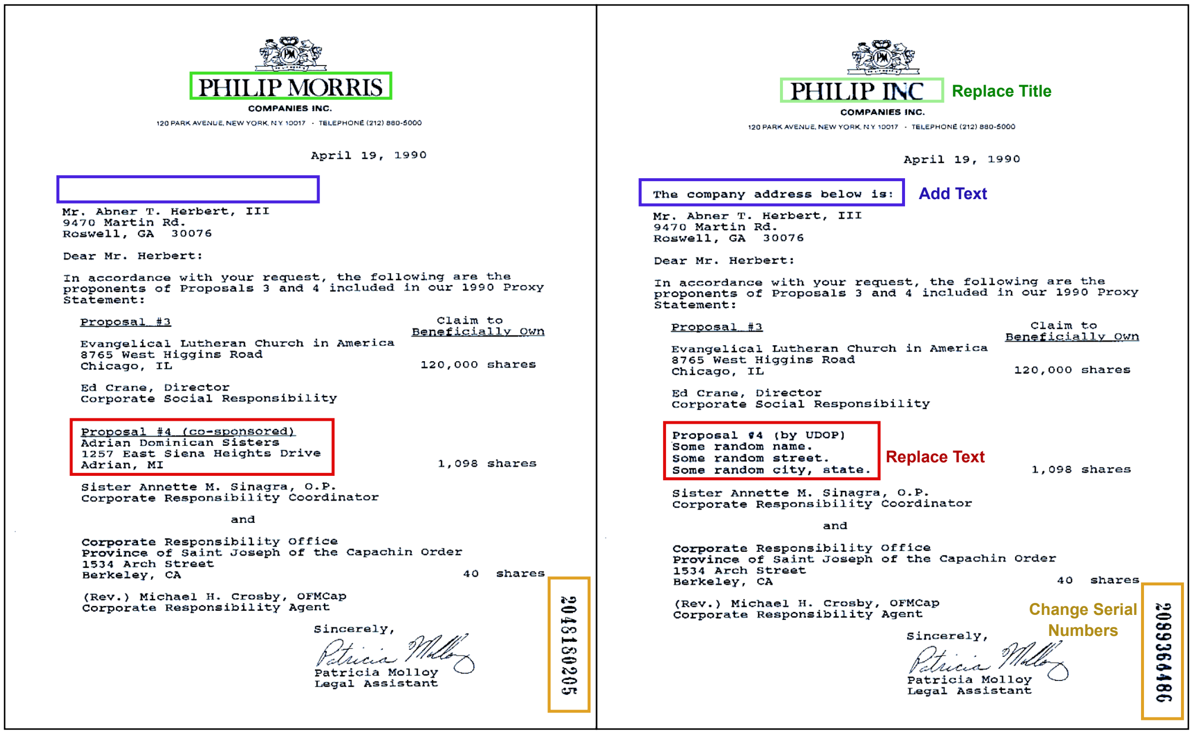
Document Generation & Editing. For the first time in Document AI, UDOP achieves controllable high-quality document generation and editing. As shown in Fig. 4), one can edit and add to the document image content with customized contents. The generated content is of high resolution and is consistent with the context in font, size, style and orientation (e.g., vertical numbers in Fig. 4). More generation examples are available in Appendix B. This is done by masking the regions to edit in the document image, and specifying the customized content in the text input, and their positions through layout embeddings. This novel functionality can generate augmentation document data for future research.

6.2. Ablation Analysis

Pretraining Objectives. Table 4 presents the ablation study of pretraining objectives on DocVQA and RVL-CDIP validation sets. We first develop a MLM (Masked Language Modeling) baseline that is a UDOP model pre-trained only on the BERT's MLM [9] that masks 15% of the input tokens. UDOP models (224 image resolution) pretrained with layout/text self-supervised objectives ("Layout Modeling", "Visual Text Dataition", and "Joint Text-Layout Reconstruction") outperforms the one trained with masked language modeling (MLM), confirming their effectiveness. Table 4 also shows relative effectiveness of each pretraining task. Layout modeling improves upon Joint Text-Layout Modeling; Masked Image Reconstruction improves on text-based pretraining tasks. Adding vision self-supervised learning (masked image reconstruction) and supervised learning further improves the performance.

Table 4. Ablation study on pre-training objectives.

Pretrain Objectives	#Pretrain Data	DocVQA	RVL-CDIP
MLM	11.0M	79.7 ± 0.4	95.3 ± 0.3
Joint Text-Layout	11.0M	82.8 ± 0.1	95.4 ± 0.3
+ Visual Text Recognition	11.0M	83.3 ± 0.2	95.4 ± 0.2
+ Layout Modeling	11.0M	84.0 ± 0.3	95.6 ± 0.2
+ Image Reconstruction	11.0M	84.4 ± 0.2	96.2 ± 0.2
+ Supervised	12.8M	**85.0**± 0.2	**96.3** ± 0.1



PHILIP MORRIS
COMPANIES INC.
120 PARK AVENUE, NEW YORK, N.Y 10017 · TELEPHONE (212) 880-5000

April 19, 1990

Mr. Abner T. Herbert, III
9470 Martin Rd.
Roswell, GA 30076

Dear Mr. Herbert:

In accordance with your request, the following are the proponents of Proposals 3 and 4 included in our 1990 Proxy Statement:

Proposal #3 — Claim to Beneficially Own

Evangelical Lutheran Church in America
8765 West Higgins Road
Chicago, IL — 120,000 shares

Ed Crane, Director
Corporate Social Responsibility

Proposal #4 (co-sponsored)
Adrian Dominican Sisters
1257 East Siena Heights Drive
Adrian, MI — 1,098 shares

Sister Annette M. Sinagra, O.P.
Corporate Responsibility Coordinator

and

Corporate Responsibility Office
Province of Saint Joseph of the Capachin Order
1534 Arch Street
Berkeley, CA — 40 shares

(Rev.) Michael H. Crosby, OFMCap
Corporate Responsibility Agent

Sincerely,
Patricia Molloy
Patricia Molloy
Legal Assistant

2048180205

PHILIP INC
COMPANIES INC.
120 PARK AVENUE, NEW YORK, N.Y 10017 · TELEPHONE (212) 880-5000

April 19, 1990

The company address below is:
Mr. Abner T. Herbert, III
9470 Martin Rd.
Roswell, GA 30076

Dear Mr. Herbert:

In accordance with your request, the following are the proponents of Proposals 3 and 4 included in our 1990 Proxy Statement:

Proposal #3 — Claim to Beneficially Own

Evangelical Lutheran Church in America
8765 West Higgins Road
Chicago, IL — 120,000 shares

Ed Crane, Director
Corporate Social Responsibility

Proposal #4 (by UDOP)
Some random name.
Some random street.
Some random city, state. — 1,098 shares

Sister Annette M. Sinagra, O.P.
Corporate Responsibility Coordinator

and

Corporate Responsibility Office
Province of Saint Joseph of the Capachin Order
1534 Arch Street
Berkeley, CA — 40 shares

(Rev.) Michael H. Crosby, OFMCap
Corporate Responsibility Agent

Sincerely,
Patricia Molloy
Patricia Molloy
Legal Assistant

2089364486

Figure 4. Document generation with customized content (right). Left is the original document. We show four document edits within the same figure including title replacement, text addition, text replacement, and tilted text replacement. All edits are done with one model run.

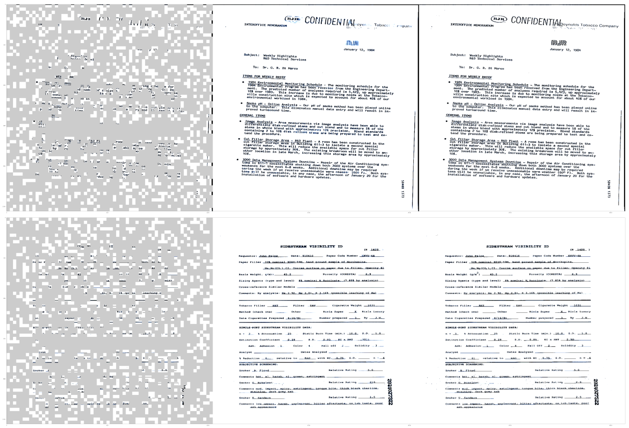


Figure 5. MAE demonstrations with 75% masking. Middle: reconstruction, Right: original.

Table 5. Ablations on model architecture.

Model	Question Answering		Information Extraction			Table QA/NLI		Avg.
	DocVQA	InfoVQA	KLC	PWC	DeepForm	WTQ	TabFact	
UDOP-Dual	84.4	47.1	81.9	28.0	85.2	46.7	**79.5**	64.6
UDOP	**84.7**	**47.4**	**82.8**	**28.0**	**85.5**	**47.2**	78.9	**64.8**

Modality-Specific Model Variant. In the field of multi-modal learning, a common model architecture is the two-tower model, where vision and text are encoded by two modality-specific encoders respectively [38, 56]. Therefore, we explore an variant of UDOP such that instead of having one unified encoder, we separately use a text encoder (to encode both text and layout tokens) and a vision encoder. Position bias are used in both encoders to represent layout information following previous works. We name this variant UDOP-Dual. For UDOP-Dual, the text-layout encoder-decoder follows T5-large, and the vision encoder-decoder has the same configuration as MAE-large. It has in total 1098M trainable parameters. As shown in Table 5 and Table 9, using one unified encoder is better than having separated encoders in most datasets. The exceptions are WTQ and RVL-CDIP on which UDOP-Dual achieves SOTA.

7. Conclusion

In this work, we propose UDOP, a foundation model for document AI. UDOP unifies the vision, text and layout modalities of documents by utilizing their strong spatial correlations through layout-induced vision-text representations and Vision-Text-Layout transformer. It also unites all self-supervised and supervised document tasks with a generative framework. UDOP achieves SOTA on 8 tasks and currently ranks the 1st place on the Document Understanding Benchmark Leaderboard. For the first time in document AI, UDOP achieves customizable realistic document generation and editing. We discuss the limitations and societal impact of our work in the appendix.

References

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 1, 7, 14, 15

[2] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 6

[3] Lu Chen, Xingyu Chen, Zihan Zhao, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021. 6, 13

[4] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations*, 2022. 4

[5] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 6, 12, 13

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 3

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018. 4, 7, 14, 15

[10] Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: layout-aware language modeling for information extraction. In *International Conference on Document Analysis and Recognition*, pages 532–547. Springer, 2021. 1, 3, 7, 14

[11] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021. 1, 3, 7, 15

[12] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 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, 2022. 1, 3

[13] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 1, 2, 6, 12

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 4, 5, 6

[15] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10767–10775, Jun. 2022. 1, 3, 7, 14, 15

[16] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 1, 3, 4, 6, 7, 14, 15

[17] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 1

[18] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019. 2, 6, 12

[19] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. Axcell: Automatic extraction of results from machine learning papers. *arXiv preprint arXiv:2004.14356*, 2020. 6, 12, 13

[20] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3

[21] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 2021. 7, 15

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, 2021. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6, 14

[24] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *arXiv preprint arXiv:2203.08411*, 2022. 1, 7, 14, 15

[25] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson 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*, pages 665–666, 2006. 6

[26] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*, 2021. 1, 7, 14, 15

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1

[28] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020. 1, 6, 12

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 1, 3, 7, 15

[30] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920, 2021. 1

[31] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[32] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6, 12, 13

[33] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 6, 12, 13

[34] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 2, 6, 12

[35] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 6, 12, 13

[36] 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. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer, 2021. 1, 3, 4, 7, 14, 15

[37] Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. *arXiv preprint arXiv:2009.14457*, 2020. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 8

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 6, 7, 14

[40] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016. 15

[41] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 6, 12

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 3

[43] S Svetlichnaya. Deepform: Understand structured documents at scale, 2020. 6, 12, 13

[44] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1

[45] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13878–13888, May 2021. 1, 6, 13

[46] Zineng Tang, Jaemin Cho, Jie Lei, and Mohit Bansal. Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention. *arXiv preprint arXiv:2211.11701*, 2022. 1

[47] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, 2021. 1

[48] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757, 2022. 1, 7, 14, 15

[49] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3, 4

[50] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 3

[51] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 6

[52] Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. Lampret: Layout-aware multimodal pretraining for document understanding. *arXiv preprint arXiv:2104.08405*, 2021. 1

[53] 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*, pages 1192–1200, 2020. 1, 3, 6, 7, 15

[54] 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. *arXiv preprint arXiv:2104.08836*, 2021. 1

[55] 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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, 2021. 1, 2, 3, 6, 7, 14, 15

[56] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-code: An integrative and composable multimodal learning framework. *arXiv preprint arXiv:2205.01818*, 2022. 3, 8

[57] 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)*, pages 1015–1022. IEEE, 2019. 1, 6, 12