TEST-TIME ADAPTATION FOR VISUAL DOCUMENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

For visual document understanding (VDU), self-supervised pretraining has been shown to successfully generate transferable representations, yet, effective adaptation of such representations to distribution shifts at test-time remains to be an unexplored area. We propose DocTTA, a novel test-time adaptation method for documents, that does source-free domain adaptation using unlabeled target document data. DocTTA leverages cross-modality self-supervised learning via masked visual language modeling as well as pseudo labeling to adapt models learned on a *source* domain to an unlabeled *target* domain at test time. We also introduce new benchmarks using existing public datasets for various VDU tasks including entity recognition, key-value extraction, and document visual question answering tasks, at which DocTTA improves the source model performance up to 1.79% in (F1 score), 3.43% (F1 score), and 17.68% (ANLS score), respectively.

1 INTRODUCTION

Visual document understanding (VDU) is on extracting structured information from document pages represented in various visual formats. It has a wide range of applications: including tax/invoice/mortgage/claims processing, identity/risk/vaccine verification, medical records understanding, compliance management, etc. These applications affect operations of businesses from major industries and lives of the general populace. Overall, it is estimated that there are trillions of documents in the world.

Machine learning solutions for VDU should rely on overall comprehension of the document content, extracting the information from text, image, and layout modalities. Most VDU tasks including key-value extraction, form understanding, document visual question answering (VQA) are often tackled by self-supervised pretraining, followed by supervised fine-tuning using human-labeled data (Appalaraju et al., 2021; Gu et al., 2021; Xu et al., 2020b;a; Lee et al., 2022; Huang et al., 2022). This paradigm uses unlabeled data in a task-agnostic way during the pretraining stage and aims to achieve better generalization at various downstream tasks. However, once the pretrained model is fine-tuned with labeled data on *source* domain, a significant performance drop might occur if these models are directly applied to a new unseen *target* domain – a phenomenon known as *domain shift* (Quiñonero-Candela et al., 2008a;b; Moreno-Torres et al., 2012).

The domain shift problem is commonly encountered in real-world VDU scenarios where the training and test-time distributions are different, a common scenario due to the tremendous diversity observed for document data. Fig. 1 exemplifies this, for key-value extraction task across visually different document templates and for visual question answering task on documents with different contents (figures, tables, letters etc.) for information. The performance difference due to this domain shift might reduce the stability and reliability of VDU models. This is highly undesirable for widespread adoption of VDU, especially given that the common use cases are for high-stakes applications from finance, insurance, healthcare, or legal. Thus, the methods to robustly guarantee high accuracy in the presence of distribution shifts would be of significant impact. Despite being a critical issue, to the best of our knowledge, no prior work has studied post-training domain adaptation for VDU.

Unsupervised domain adaptation (UDA) methods attempt to mitigate the adverse effect of data shifts, often by training a joint model on the labeled source and unlabeled target domains that map both domains into a common feature space. However, simultaneous access to data from source and target domains may not be feasible for VDU due to privacy concerns associated with source data access,

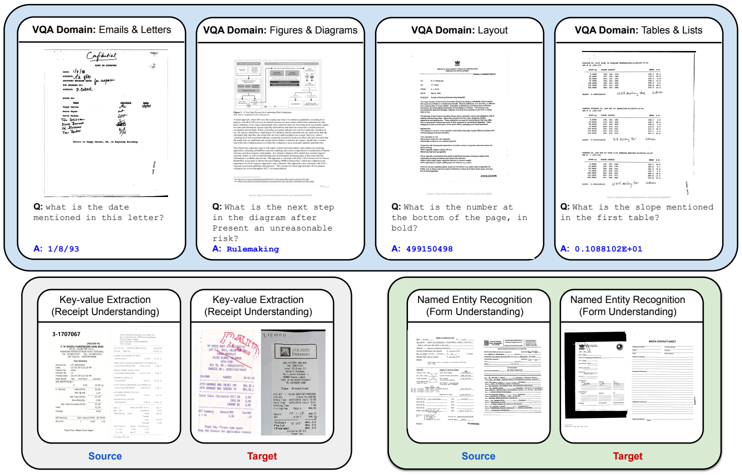
VQA Domain: Emails & Letters
Q: what is the date mentioned in this letter?
A: 1/8/93
VQA Domain: Figures & Diagrams
Q: What is the next step in the diagram after Present an unreasonable risk?
A: Rulemaking
VQA Domain: Layout
Q: What is the number at the bottom of the page, in bold?
A: 499150498
VQA Domain: Tables & Lists
Q: What is the slope mentioned in the first table?
A: 0.1088102E+01
Key-value Extraction (Receipt Understanding)
Key-value Extraction (Receipt Understanding)
Source
Target
Named Entity Recognition (Form Understanding)
Named Entity Recognition (Form Understanding)
Source
Target

Figure 1: Distribution shift examples for document samples from the proposed benchmark, DocVQA-TTA. **Top row:** shows documents from four domains: (i) Emails & Letters, (ii) Figures & Diagrams, (iii) Layout, (iv) Tables & Lists, from our VQA benchmark derived from DocVQA dataset (Mathew et al., 2021). **Bottom left:** documents from source and target domains for key-value information extraction task from SROIE (Huang et al., 2019) receipt dataset. **Bottom right:** documents from source and target domains for named entity recognition task from FUNSD (Jaume et al., 2019) dataset.

given legal, technical, and contractual constraints. In addition, the training and serving may be done in different computational environments, and thus, the expensive computational resources used for training may not be available. Test-time adaptation (TTA) (or source-free domain adaptation) has been introduced to adapt a model that is trained on the source to unseen target data, without using any source data (Liang et al., 2020; Wang et al., 2021b; Sun et al., 2020; Wang et al., 2021a; Chen et al., 2022; Huang et al., 2021). Existing TTA methods have mainly focused on image classification and semantic segmentation tasks, while VDU remains unexplored, despite the clear motivations of the distribution shift besides challenges for the employment of standard UDA.

Since VDU significantly differs from other computer vision (CV) tasks, applying existing TTA methods in a straightforward manner is suboptimal. First, in VDU, information is extracted from multiple modalities (including image, text, and layout) unlike other CV tasks. Therefore, a TTA approach proposed for VDU should leverage cross-modal information for better adaptation. Second, multiple outputs (e.g. entities or questions) are obtained from the same document, creating the scenario that their similarity in some aspects (e.g. in format or context) can be used. However, this may not be utilized in a beneficial way with direct application of popular pseudo labeling or self training-based TTA approaches (Lee et al., 2013), which have gained a lot of attention in CV (Liang et al., 2020; 2021; Chen et al., 2022; Wang et al., 2021a). Pseudo labeling uses predictions on unlabeled target data for training. However, in VDU, naive pseudo labeling can result in accumulation of errors due to generation of multiple outputs at the same time that are possibly wrong in the beginning, as each sample can contain a long sequence of words. Third, commonly-used self-supervised contrastive-based TTA methods in CV (He et al., 2020; Chen et al., 2020b;a; Tian et al., 2020) (that are known to increase generalization) employ a rich set of image augmentation techniques, while proposing data augmentation is much more challenging for general VDU.

In this paper, we propose DocTTA, a novel TTA method for VDU that utilizes self-supervised learning on text and layout modalities using masked visual language modeling (MVLM) while jointly optimizing with pseudo labeling. We introduce a new uncertainty-aware per-batch pseudo labeling selection mechanism, which makes more accurate predictions compared to the commonly-used pseudo labeling techniques in CV that use no pseudo-labeling selection mechanism (Liang et al., 2020) in TTA or select pseudo labels based on both uncertainty and confidence (Rizve et al., 2021) in semi-supervised learning settings. To the best of our knowledge, this is the first method that employs a self-supervised objective function that combines visual and language representation learning as a key differentiating factor compared to TTA methods proposed for image or text data. While our main focus is the TTA setting, we also showcase a special form of DocTTA where access to source data is granted at test time, extending our approach to be applicable for unsupervised domain adaptation, named DocUDA. Moreover, in order to evaluate DocTTA diligently and facilitate future research in this direction, we introduce new benchmarks for various VDU tasks including key-value extraction, entity recognition, and document visual question answering (DocVQA) using publicly available datasets by modifying them to mimic real-world adaptation scenarios. We show DocTTA significantly improves source model performance at test-time on all VDU tasks without any supervision. To our knowledge, our paper is first to demonstrate TTA and UDA for VDU applications, showing the significant accuracy gain potential via adaptation. We expect our work to open new horizons for future research in VDU and real-world deployment in applications.

2 Related work

Unsupervised domain adaptation aims to improve the performance on a different target domain, for a model trained on the source domain. UDA approaches for closed-set adaptation (where classes fully overlap between the source and target domains) can be categorized into four categories: (i) distribution alignment-based, (ii) reconstruction-based, and (iii) adversarial based, and (iv) pseudo-labeling based. Distribution alignment-based approaches feature aligning mechanisms, such as moment matching (Peng et al., 2019) or maximum mean discrepancy (Long et al., 2015; Tzeng et al., 2014). Reconstruction-based approaches reconstruct source and target data with a shared encoder while performing supervised classification on labeled data (Ghifary et al., 2016), or use cycle consistency to further improve domain-specific reconstruction (Murez et al., 2018; Hoffman et al., 2018). Inspired by GANs, adversarial learning based UDA approaches use two-player games to disentangle domain invariant and domain specific features (Ganin & Lempitsky, 2015; Long et al., 2018; Shu et al., 2018). Pseudo-labeling (or self-training) approaches jointly optimize a model on the labeled source and pseudo-labeled target domains for adaptation (Kumar et al., 2020; Liu et al., 2021; French et al., 2017). Overall, all UDA approaches need to access both labeled source data and unlabeled target data during the adaptation which is a special case for the more challenging setting of TTA and we show how our approach can be modified to be used for UDA.

Test-time adaptation corresponds to source-free domain adaptation, that focuses on the more challenging setting where only source model and unlabeled target data are available. The methods often employ an unsupervised or self-supervised cost function. TENT (Wang et al., 2021b) utilizes entropy minimization for fully test-time adaptation which encourages the model to become more "certain" on target predictions regardless of their correctness. In the beginning of the training when predictions tend to be inaccurate, entropy minimization can lead to error accumulation since VDU models create a long sequence of outputs per every document resulting in a noisy training. SHOT (Liang et al., 2020) combines mutual information maximization with offline clustering-based pseudo labeling. However, similar to TENT, using simple offline pseudo-labeling can lead to noisy training and poor performance when the distribution shifts are large (Chen et al., 2022; Liu et al., 2021; Rizve et al., 2021; Mukherjee & Awadallah, 2020). We also use pseudo labeling in DocTTA but we propose online updates per batch for pseudo labels, as the model adapts to test data. Besides, we equip our method with a pseudo label rejection mechanism using uncertainty, to ensure the negative effects of predictions that are likely to be inaccurate. Most recent TTA approaches in image classification use contrastive learning combined with extra supervision (Xia et al., 2021; Huang et al., 2021; Wang et al., 2021a; Chen et al., 2022). In contrastive learning, the idea is to jointly maximize the similarity between representations of augmented views of the same image, while minimizing the similarity between representations of other samples). All these methods rely on self-supervised learning that utilize data augmentation techniques, popular in CV while not yet being as effective for VDU. While we advocate for using SSL during TTA, we propose to employ multimodal SSL with pseudo labeling for the first time which is imore effective for VDU.

Self-supervised pretraining for VDU aims to learn generalizable representations on large scale unlabeled data to improve downstream VDU accuracy (Appalaraju et al., 2021; Gu et al., 2021; Xu et al., 2020b;a; Lee et al., 2022; Huang et al., 2022). LayoutLM (Xu et al., 2020b) jointly models interactions between text and layout information using a masked visual-language modeling objective and performs supervised multi-label document classification on IIT-CDIP dataset (Lewis et al., 2006). LayoutLMv2 (Xu et al., 2020a) extends it by training on image modality as well, and optimizing text-image alignment and text-image matching objective functions. DocFormer (Appalaraju et al.,

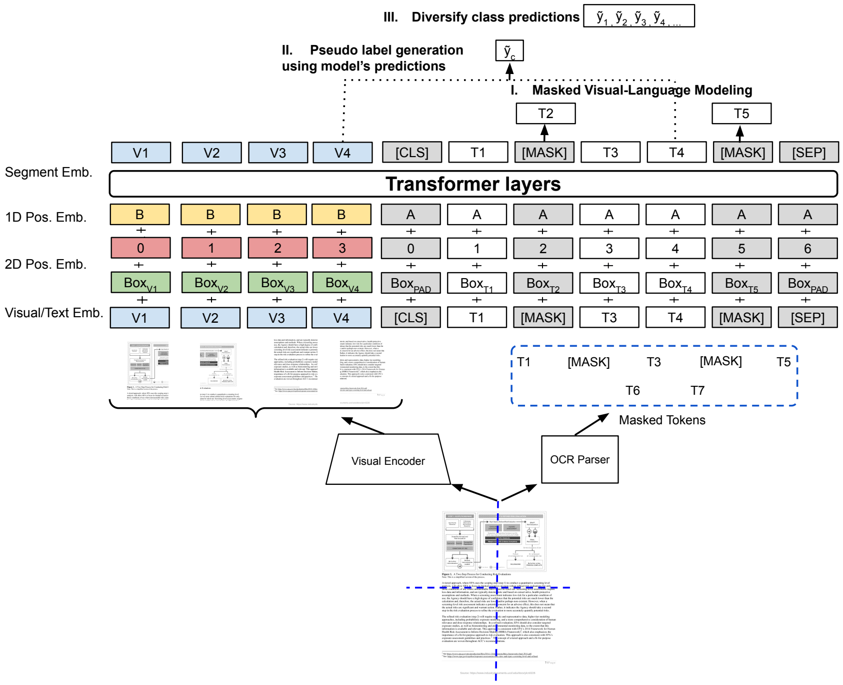

Figure 2: Illustration of how our approach, DocTTA, leverages unlabeled target data at test time to i) learn how to predict masked language given visual cues, ii) generate pseudo labels to supervise the learning, and iii) maximize the diversity of predictions to generate enough labels from all classes.

2021) is another multi-modal transformer based architecture that uses text, vision and spatial features and combines them using multi-modal self-attention with a multi-modal masked language modeling (MM-MLM) objective (as a modified version of MLM in BERT (Devlin et al., 2018)), an image reconstruction loss, and a text describing image loss represented as a binary cross-entropy to predict if the cut-out text and image are paired. FormNet (Lee et al., 2022) is a structure-aware sequence model that combines a transformer with graph convolutions and proposed *rich attention* that uses spatial relationship between tokens. UniDoc (Gu et al., 2021) is another multi-modal transformer based pretraining method that uses masked sentence modeling, visual contrastive learning, and visual language alignment objectives which unlike other methods, does not have a fixed document object detector (Li et al., 2021; Xu et al., 2020a). In this work, we focus on a novel TTA approach for VDU, that can be integrated with any pre-training method. We demonstrate DocTTA using the publicly available LayoutLMv2 architecture pretrained on IIT-CDIP dataset.

3 DocTTA: Test-time adaptation for documents

In this section, we introduce DocTTA, a test-time adaptation framework for VDU tasks including key-value extraction, entity recognition, and document visual question answering (VQA).

3.1 DocTTA framework

We define a *domain* as a pair of distribution $\mathcal{D}$ on inputs $\mathcal{X}$ and a labeling function $l : \mathcal{X} \rightarrow \mathcal{Y}$. We consider *source* and *target* domains. In the source domain, denoted as $\langle D_s, l_s \rangle$, we assume to have a model denoted as f_s and parameterized with θ_s to be trained on source data $\{x_s^{(i)}, y_s^{(i)}\}_{i=1}^{n_s}$, where $x_s^{(i)} \in \mathcal{X}_s$ and $y_s^{(i)} \in \mathcal{Y}_s$ are document inputs and corresponding labels, respectively and n_s is the number of documents in the source domain. Given the trained source model f_s and leaving $\mathcal{X}_s$ behind, the goal of TTA is to train f_t on the target domain denoted as $\langle \mathcal{D}_t, l_t \rangle$ where f_t is parameterized with θ_t and is initialized with θ_s and $\mathcal{D}_t$ is defined over $\{x_t^{(i)}\}_{i=1}^{n_t} \in \mathcal{X}_t$ without any ground truth label. Algorithm 1 overviews our proposed DocTTA procedure.

Unlike single-modality inputs commonly used in computer vision, documents are images with rich textual information. To extract the text from the image, we consider optical character recognition (OCR) is performed and use its outputs, characters, and their corresponding bounding boxes (details are provided in Appendix). We construct our input $\mathcal{X}$ in either of the domains composed of three components: text input sequence X^T of length n denoted as $(x_1^T, \cdots, x_n^T) \in \mathbb{R}^{(n \times d)}$, image $X^I \in \mathbb{R}^{3 \times W \times H}$, and layout X^B as a 6-dimensional vector in the form of $(x_{min}, x_{max}, y_{min}, y_{max}, w, h)$ representing a bounding box associated with each word in the text input sequence. For the entity recognition task, labels correspond to the set of classes that denote the extracted text; for the key-value extraction task, labels are values for predefined keys; and for the VQA task, labels are the starting and ending positions of the answer presented in the document for the given question. We consider the closed-set assumption: the source and target domains share the same class labels $\mathcal{Y}_s = \mathcal{Y}_t = \mathcal{Y}$ with $|\mathcal{Y}| = C$ being the total number of classes.

Algorithm 1 DocTTA for closed-set TTA in VDU

```
Input: Source model weights $\theta_s$, target documents $\{x_t^i\}_{i=1}^{n_t}$, test-time training epochs $n_e$, test-time training learning rate $\alpha$, uncertainty threshold $\gamma$
Initialization: Initialize target model $f_{\theta_t}$ with $\theta_s$ weights.
for $epoch = 1$ to $n_e$ do
    Perform masked visual-language modeling in Eq. 1
    Generate pseudo labels and accept a subset using criteria in Eq. 3 and fine-tune with Eq. 2
    Maximize diversity in pseudo label predictions Eq. 4
    $\theta_t \leftarrow \theta_t - \alpha \nabla \mathcal{L}_{\text{DocTTA}}$        ▷ Update $\theta_t$ via total loss in Eq. 5
end for
```

3.2 DocTTA objective functions

In order to adapt f_t in DocTTA, we propose three objectives to optimize on the unlabeled target data:

Objective I: masked visual language modeling (MVLM). Inspired by masked language modeling in BERT (Devlin et al., 2018) and MVLM used in (Xu et al., 2020a) to perform self-supervised pretraining, we propose to employ MVLM at test time to encourage the model to learn better the text representation of the test data given 2D positions and other text tokens. The intuition behind using this objective for TTA is to enable the target model to learn the language modality of the new data given visual cues and thereby bridging the gap between the different modalities on the target domain. To do so, we randomly mask 15% of input text tokens among which 80% are replaced by a special token `[MASK]` and the remaining tokens are replaced by a random word from the entire vocabulary. The model is then trained to recover the masked tokens while the layout information remains fixed. To do so, the output representations of masked tokens from the encoder are fed into a classifier which outputs logits over the whole vocabulary, to minimize the negative log-likelihood of correctly recovering masked text tokens x_m^T given masked image tokens x^I and masked layout x^B:

$$\mathcal{L}_{MVLM}(\theta_t) = -\mathbb{E}_{x_t \in \mathcal{X}_t} \sum_m \log p_{\theta_t}(x_{t_m}^T | x_t^I, x_t^B). \quad (1)$$

Objective II: self training with pseudo labels. While optimizing MVLM loss during the adaptation, we also generate pseudo labels for the unlabeled target data in an online way and treat them as ground truth labels to perform supervised learning on the target domain. Unlike previous pseudo labeling-based approaches for TTA in image classification which update pseudo labels only after each epoch (Liang et al., 2020; 2021; Wang et al., 2021a), we generate hard pseudo labels per batch aiming to use the latest version of the model for predictions. In addition, unlike prior works, we do not use a clustering mechanism to generate pseudo labels as they will be computationally expensive for documents. Instead, we directly use predictions by the model. However, simply using all the predictions would lead to noisy pseudo labels. Inspired by (Rizve et al., 2021), in order to prevent noisy pseudo labels, we employ an uncertainty-aware selection mechanism to select the subset of pseudo labels with low uncertainty. Note that in (Rizve et al., 2021), pseudo labeling is used as a semi-supervised learning approach and the selection criteria is based on both thresholding confidence and using MC-Dropout (Gal & Ghahramani, 2016) as the measure of uncertainty. We empirically observe that raw confidence values (when taken as the posterior probability output from the model) are overconfident despite being right or wrong. Setting a threshold on pseudo labels' confidence only introduces a new hyperparameter without a performance gain (see Sec. 5.1). Instead, to select the predictions we propose to only use uncertainty, in the form of Shannon's entropy (Shannon, 2001). We also expect this selection mechanism leads to reducing miscalibration due to the direct relationship between the ECE[1] and output prediction uncertainty, i.e. when more certain predictions are selected, ECE is expected to reduce for the selected subset of pseudo labels. Assume $\mathbf{p}^{(i)}$ be the output probability vector of the target sample $x_t^{(i)}$ such that $p_c^{(i)}$ denotes the probability of class c being the correct class. We select a pseudo label $\tilde{y}_c^{(i)}$ for $x_t^{(i)}$ if the uncertainty of the prediction $u(p_c^{(i)})$, measured with Shannon's entropy, is below a specific threshold γ and we update θ_t weights with a cross-entropy loss:

$$\tilde{y}_c^i = \mathbb{1}\left[u(p_c^{(i)}) \leq \gamma\right], \tag{2}$$

$$\mathcal{L}_{CE}(\theta_t) = -\mathbb{E}_{x_t \in \mathcal{X}_t} \sum_{c=1}^{C} \tilde{y}_c \log \sigma(f_t(x_t)), \tag{3}$$

where $\sigma(\cdot)$ is the softmax function.

Objective III: diversity objective. To prevent the model from indiscriminately being dominated by the most probable class based on pseudo labels, we encourage class diversification in predictions by minimizing the following objective:

$$\mathcal{L}_{DIV} = \mathbb{E}_{x_t \in \mathcal{X}_t} \sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c, \tag{4}$$

where $\bar{p} = \mathbb{E}_{x_t \in \mathcal{X}_t} \sigma(f_t(x_t))$ is the output embedding of the target model averaged over target data. By combining Eqs. 1, 2, and 4, we obtain the full objective function in DocTTA as below:

$$\mathcal{L}_{\text{DocTTA}} = \mathcal{L}_{MVLM} + \mathcal{L}_{CE} + \mathcal{L}_{DIV}. \tag{5}$$

3.3 DocTTA vs. DocUDA

The proposed DocTTA framework can be extended as a UDA approach, which we refer to as DocUDA (see Appendix for the algorithm and details), by enabling access to source data during adaptation to the target. In principle, the availability of this extra information of source data provides can provide an advantage over TTA, however, as we show in our experiments, their difference is small in most cases, and even TTA can be superior when source domain is significantly smaller than the target domain and the distribution gap is large, highlighting the efficacy of our DocTTA approach in adapting without relying on already-seen source data. The UDA version comes with fundamental drawbacks. From the privacy perspective, there would be concerns associated with accessing or storing source data in deployment environments, especially given that VDU applications are often from privacy-sensitive domains like legal or finance. From the computational perspective, UDA would yield longer convergence time and higher memory requirements due to joint learning from source data. Especially given that the state-of-the-art VDU models are large in size, this may become a major consideration.

4 DocTTA benchmarks

To better highlight the impact of distribution shifts and to study the methods that are robust against them, we introduce new benchmarks for VDU. Our benchmark datasets are constructed from existing popular and publicly-available VDU data to mimic real-world challenges. We have attached the training and test splits for all our benchmark datasets in the supplementary materials.

4.1 FUNSD-TTA: Entity recognition adaptation benchmark

We consider FUNSD (Jaume et al., 2019) dataset for this benchmark which is a noisy form understanding collection with 9707 semantic entities and 31,485 words with 4 categories of entities `question`, `answer`, `header`, and `other`, where each category (except `other`) is either the beginning or the intermediate word of a sentence. Therefore, in total, we have 7 classes. We first

[1] Expected calibration error (Naeini et al., 2015) which is a metric to measure calibration of a model

combine the original training and test splits and then manually divide them into two groups. We set aside 149 forms that are filled with more texts for the source domain and put 50 forms that are sparsely filled for the target domain. We randomly choose 10 out of 149 documents for validation, and the remaining 139 for training. Fig. 1 (bottom row on the right) shows examples from the source and target domains.

4.2 SROIE-TTA: Key-value extraction adaptation benchmark

We use SROIE (Huang et al., 2019) dataset with 9 classes in total. Similar to FUNSD, we first combine the original training and test splits. Then, we manually divide them into two groups based on their visual appearance – source domain with 600 documents contains standard-looking receipts with proper angle of view and clear black ink color. We use 37 documents from this split for *validation*, which we use to tune adaptation hyperparameters. Note that the validation split does not overlap with the target domain, which has 347 receipts with slightly blurry look, rotated view, colored ink, and large empty margins. Fig. 1 (bottom row on the left) exemplifies documents from the source and target domains.

4.3 DocVQA-TTA: Document VQA adaptation benchmark

We use DocVQA (Mathew et al., 2021), a large-scale VQA dataset with nearly 20 different types of documents including scientific reports, letters, notes, invoices, publications, tables, etc. The original training and validation splits contain questions from all of these document types. However, for the purpose of creating an adaptation benchmark, we select 4 *domains* of documents: i) *Emails & Letters* (**E**), ii) *Tables & Lists* (**T**), iii) *Figure & Diagrams* (**F**), and iv) *Layout* (**L**). Since DocVQA doesn't have public meta-data to easily sort all documents with their questions, we use a simple keyword search to find our desired categories of questions and their matching documents. We use the same words in domains' names to search among questions (i.e., we search for the words of "email" and "letter" for *Emails & Letters* domain). However, for *Layout* domain, our list of keywords is ["top", "bottom", "right", "left", "header", "page number"] which identifies questions that are querying information from a specific location in the document. Among the four domains, **L** and **E** have the shortest gap because emails/letters have structured layouts and extracting information from them requires understanding relational positions. For example, the name and signature of the sender are usually at the bottom, while the date usually appears at top left. However, **F** and **T** domains seem to have larger gaps with other domains, that we attributed to that learning to answer questions on figures or tables requires understanding local information withing the list or table. Fig. 1 (top row) exemplifies some documents with their questions from each domain. Document counts in each domain are provided in Appendix.

5 Experiments

Evaluation metrics: For entity recognition and key-value extraction tasks, we use entity-level F1 score as the evaluation metric, whereas for the document VQA task, we use Average Normalized Levenshtein Similarity (ANLS) introduced by (Biten et al., 2019) (as it is recognized as a better measure compared to accuracy since it doesn't penalize minor text mismatches due to OCR errors).

Model architecture: In all experiments, we use LayoutLMv2$_{BASE}$ architecture which has a 12-layer 12-head transformer encoder with a hidden size of 768. Its visual backbone is based on ResNeXt101-FPN, similar to that of MaskRCNN (He et al., 2017). Overall, it has $\sim$200M parameters. We note that our approach is architecture independent and hence applicable to any attention-based VDU model. Details on training and hyper parameter tuning are provided in Appendix.

Baselines: As our method is the first TTA approach proposed for VDU tasks, there is no baseline to compare directly. Thus, we adopt TTA and UDA approaches from image classification, as they can be applicable for VDU given that they do not depend on augmentation techniques, contrastive learning, or generative modeling. For UDA, we use the baselines **DANN** (Ganin & Lempitsky, 2015) and **CDAN** (Long et al., 2018) and for TTA, we use the baselines batch normalization **BN** (Ioffe & Szegedy, 2015; Nado et al., 2020), **TENT** (Wang et al., 2021b), and **SHOT** (Liang et al., 2020). We also provide **source-only**, where the model trained on source and evaluated on target without any adaptation mechanism, and **train-on-target**, where the model is trained (and tested) on target domain

Table 1: F1 score results for adapting source to target in **FUNSD-TTA** and **SROIE-TTA** benchmarks. Availability of the labeled/unlabeled data from source/target domains during *adaptation* in UDA and TTA settings and *training* phase in source-only and train-on-target settings are marked. Standard deviations are in parentheses.

DA category	Methods	Labeled source data	Labeled target data	Unlabeled target data	FUNSD-TTA	SROIE-TTA
-	Source-only	✓	×	×	80.80 (0.12)	92.45 (0.08)
UDA	DANN	✓	×	✓	82.54 (0.14)	92.89 (0.13)
	CDAN	✓	×	✓	83.72 (0.61)	93.36 (0.18)
	DocUDA (ours)	✓	×	✓	**89.76** (0.09)	**97.38** (0.15)
TTA	BN	×	×	✓	80.84 (0.93)	92.41 (0.45)
	TENT	×	×	✓	79.78 (1.28)	92.42 (0.87)
	SHOT	×	×	✓	80.89 (1.03)	92.78 (0.65)
	DocTTA (ours)	×	×	✓	**84.23** (0.88)	**94.34** (0.43)
-	Train-on-target	×	✓	×	99.89 (0.05)	100.0 (0.00)

using the exact same hyperparameters used for TTA (which are found on the validation set and might not be the most optimal values for target data). While these two baselines don't adhere to any domain adaptation setting, they can be regarded as the ultimate lower and upper bound for performance.

5.1 Results and Discussions

FUNSD-TTA: Table 1 shows the comparison between DocTTA and DocUDA with their corresponding TTA and UDA baselines. For UDA, DocUDA outperforms all other UDA baselines by a large margin, and improves 8.96% over the source-only. For the more challenging setting of TTA, DocTTA improves the F1 score of the source-only model by 3.43%, whereas the performance gain by all the TTA baselines is less than 0.5%. We also observe that DocTTA performs slightly better than other UDA baselines DANN and CDAN, which is remarkable given that unlike those, DocTTA does not have access to the source data at test time.

SROIE-TTA: Table 1 shows the comparison between UDA and TTA baselines vs. DocUDA and DocTTA on SROIE-TTA benchmark. Similar to our findings for FUNSD-TTA, DocUDA and DocTTA outperform their corresponding baselines, where DocTTA can even surpass DANN and CDAN (which use source data at test time). Comparison of DocUDA and DocTTA shows that for small distribution shifts, UDA version of our framework results in better performance.

DocVQA-TTA: Table 2 shows the results on our DocVQA-TTA benchmark, where the ANLS scores are obtained by adapting each domain to all the remaining ones. The distribution gap between domains on this benchmark is larger compared to FUNSD-TTA and SROIE-TTA benchmarks. Hence, we also see a greater performance improvement by using TTA/UDA across all domains and methods. For the UDA setting, DocUDA consistently outperforms adversarial-based UDA methods by a large margin, underlining the superiority of self-supervised learning and pseudo labeling in leveraging labeled and unlabeled data at test time. Also in the more challenging TTA setting, DocTTA consistently achieves the highest gain in ANLS score of at least 2.57% increase on $\mathbf{E} \rightarrow \mathbf{F}$ and up to 17.68% on $\mathbf{F} \rightarrow \mathbf{E}$. Moreover, DocTTA significantly outperforms DANN on all domains and CDAN on 11 out of 12 adaptation scenarios, even though it does not utilize source data at test time. This demonstrates the efficacy of joint online pseudo labeling with diversity maximization and masked visual learning. Between DocUDA and DocTTA, it is expected that DocUDA performs better than DocTTA due to having extra access to source domain data. However, we observe three exceptions where DocTTA surpasses DocUDA by 1.13%, 0.79%, and 2.16% in ANLS score on $\mathbf{E} \rightarrow \mathbf{F}$ and $\mathbf{T} \rightarrow \mathbf{F}$, and $\mathbf{L} \rightarrow \mathbf{F}$, respectively. We attribute this to: i) target domain (**F**) dataset size being relatively small, and ii) large domain gap between source and target domains. The former can create an imbalanced distribution of source and target data, where the larger split (source data) dominates the learned representation. This effect is amplified due to (ii) because the two domains aren't related and the joint representation is biased in favor of the labeled source data. Another finding on this benchmark is that a source model

Table 2: ANLS scores for adapting between domains in **DocVQA-TTA** benchmark. Standard deviations are shown in Appendix.

Source:	Emails&Letters (**E**)			Figures&Diagrams (**F**)			Tables&Lists (**T**)			Layout (**L**)		
Target:	**F**	**T**	**L**	**E**	**T**	**L**	**E**	**F**	**L**	**E**	**F**	**T**
Source-only	37.79	25.59	38.25	5.23	7.03	3.65	13.66	20.48	14.58	53.55	33.36	33.43
DANN	38.94	27.22	40.23	15.43	9.34	7.45	17.67	22.19	17.67	54.55	33.87	33.58
CDAN	39.08	29.33	41.29	16.99	11.32	10.23	27.87	25.23	27.66	56.82	34.27	34.81
DocUDA (ours)	**39.23**	**43.54**	**57.99**	**24.21**	**15.76**	**20.45**	**53.19**	**29.91**	**47.81**	**61.09**	**34.85**	**41.80**
BN	38.10	26.89	38.23	7.32	8.56	9.35	15.13	22.24	15.65	53.23	33.67	33.55
TENT	38.34	26.42	40.45	12.38	7.34	11.29	16.01	20.23	15.02	53.34	33.59	34.55
SHOT	38.98	27.55	39.15	14.34	10.10	13.21	22.56	24.33	19.15	56.23	34.56	35.65
DocTTA (ours)	**40.36**	**35.28**	**49.35**	**22.91**	**15.67**	**16.01**	**35.67**	**30.70**	**26.32**	**59.84**	**37.01**	**39.10**
Train-on-target	95.28	93.54	95.01	39.70	24.77	38.59	84.59	70.66	83.73	92.32	91.36	93.41

Table 3: Ablation analysis on adapting from E to F, T, L in our DocVQA-TTA benchmark with different components including pseudo labeling, $\mathcal{L}_{MVLM}$, $\mathcal{L}_{DIV}$, and pseudo label selection mechanism using confidence only or together with uncertainty. Standard deviations are in parentheses.

Method	F	T	L
Source-only	37.79 (1.30)	25.59 (1.78)	38.25 (0.92)
DocTTA, conf.	32.67 (1.68)	21.50 (1.52)	6.71 (3.21)
DocTTA, conf. & unc.	39.45 (0.87)	28.47 (0.72)	47.50 (0.51)
DocTTA, no $\mathcal{L}_{MVLM}$	35.66 (0.46)	25.72 (0.55)	45.88 (0.34)
DocTTA, no $\mathcal{L}_{DIV}$	34.32 (0.53)	25.17 (0.49)	46.36 (0.21)
DocTTA, no pseudo labeling	33.61 (1.65)	23.43 (0.87)	15.89 (1.35)
DocTTA	40.36 (0.53)	35.28 (0.76)	49.35 (1.20)

trained on a domain with a small dataset generalizes less compared to the one with sufficiently-large dataset but has a larger domain gap with the target domain. Results for train-on-target on each domain can shed light on this. When we use the domain with the smallest dataset (**F**) as the source, each domain can only achieve its lowest ANLS score (39.70% on **E**, 24.77% on **T**, and 38.59% on **L**) whereas with **T**, second smallest domain in dataset size in our benchmark (with 657 training documents), the scores obtained by train-on-target on **E** and **L** increases to 84.59% and 83.73%, respectively. Thus, even if we have access to entire target labeled data, the limitation of source domain dataset size is still present.

Ablation studies: We compare the impact of different constituents of our methods on DocVQA-TTA benchmark, using a model trained on Emails&Letters domain and adapted to other three domains. Table 3 shows that pseudo labeling selection mechanism plays an important role and using confidence scores to accept pseudo labels results in the poorest performance, much below the source-only ANLS values and even worse than not using pseudo labeling. On the other hand, using uncertainty and raw confidence together to select pseudo labels yields the closest performance to that of the full (best) method (details are provided in Appendix). MVLM loss and diversity maximization criteria have similar impact on DocTTA's performance.

6 Conclusions

We introduce TTA for VDU for the first time, with our novel approach DocTTA, along with new realistic adaptation benchmarks for common VDU tasks such as entity recognition, key-value extraction, and document VQA. DocTTA starts from a pretrained model on the source domain and uses online pseudo labeling along with masked visual language modeling and diversity maximization on the unlabeled target domain. We propose an uncertainty-based online pseudo labeling mechanism that generates significantly more accurate pseudo labels in a per-batch basis. Overall, novel TTA approaches result in surpassing the state-of-the-art TTA approaches adapted from computer vision.

7 Reproducibility

We have included the details of our experimental setup such as the utilized compute resources, pretrained model, optimizer, learning rate, batch size, and number of epochs, etc. in Section A.2 of the Appendix. We have provided the details of our hyper parameter tuning and our search space in Section A.2.2 of the Appendix. For our introduced benchmark datasets, the statistics of each dataset is detailed in Section A.1.2. List of the all the training and validation splits for our proposed benchmarks are also located at `Supplemental/TTA_Benchmarks` as `json` files. To ensure full reproducibility, we will also release our code upon acceptance.

References

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*, pp. 993–1003, 2021.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1563–1570. IEEE, 2019.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pp. 597–613. Springer, 2016.

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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 39–50. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/0084ae4bc24c0795d1e6a4f58444d39b-Paper.pdf`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hoffman18a.html.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

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

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

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, pp. 1–6. IEEE, 2019.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

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.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

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*, pp. 665–666, 2006.

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*, pp. 5652–5660, 2021.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039, 2020.

Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. In Press.

Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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*, pp. 2200–2209, 2021.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for text classification with few labels. *arXiv preprint arXiv:2006.15315*, 2020.

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4500–4509, 2018.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and N Lawrence. Covariate shift and local learning by distribution matching, 2008a.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008b.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-ODN6SbiUU.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1q-TM-AW.

Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. *arXiv preprint arXiv:2109.01087*, 2021a.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9010–9019, 2021.

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

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200, 2020b.