RegCLR: A Self-Supervised Framework for Tabular Representation Learning in the Wild

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in self-supervised learning (SSL) using large models to learn visual representations from natural images are rapidly closing the gap between the results produced by fully supervised learning and those produced by SSL on downstream vision tasks. Inspired by this advancement and primarily motivated by the emergence of tabular and structured document image applications, we investigate which self-supervised pretraining objectives, architectures, and fine-tuning strategies are most effective. To address these questions, we introduce RegCLR, a new self-supervised framework that combines contrastive and regularized methods and is compatible with the standard Vision Transformer (Dosovitskiy et al., 2021) architecture. Then, RegCLR is instantiated by integrating masked autoencoders (He et al., 2022) as a representative example of a contrastive method and enhanced Barlow Twins as a representative example of a regularized method with configurable input image augmentations in both branches. Several real-world table recognition scenarios (e.g., extracting tables from document images), ranging from standard Word and Latex documents to even more challenging electronic health records (EHR) computer screen images, have been shown to benefit greatly from the representations learned from this new framework, with detection average-precision (AP) improving relatively by 4.8% for Table, 11.8% for Column, and 11.1% for GUI objects over a previous fully supervised baseline on real-world EHR screen images.

1 Introduction

Many self-supervised learning (SSL) methods for learning visual representation without supervision (or labels) have been proposed in recent years (Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018; Chen et al., 2020; Grill et al., 2020; Zbontar et al., 2021; He et al., 2022). Indeed, the potentially practical benefits of avoiding the high cost of human annotations and moving away from human-defined and language-dependent label categories strongly motivate this direction. In particular, the most common framework is to use abundant *unlabeled* images to pretrain rich visual representations and then transfer them to supervised downstream tasks with much fewer annotated ones.

Tables are the most prevalent means of representing and communicating structured data in a wide range of document images, such as financial statements, scientific papers, and electronic medical health documents. Despite the explosive growth of the number of these document images (Vashisth et al., 2022), most SSL approaches have so far been proposed in the context of the natural image domain, with little attention paid to tabular and structured document image domains. Even the most advanced systems (e.g., Prasad et al. (2020); Agarwal et al. (2021)) are largely dependent on object detection models that have been trained on human-labeled natural images.

Table objects are a compact representation that humans can easily understand. However, this is not true for machines since, unlike classic object detection classes, they might have widely disparate sizes, types, styles, and aspect ratios. In other words, table structure varies greatly between document domains (e.g., Word vs GUI screen), and a large variety of table styles are feasible even within the same document format (e.g., borderless vs bordered).

More importantly, object detection models trained for natural images (e.g., Faster R-CNN or YOLO) retrained for table can work well locally while ignoring the overall style of tables that is important to the table (Burdick et al., 2020). In that regard, very few works have started exploring SSL approaches to the problem of tabular rich document image domains. DiT (Li et al., 2022), for example, directly employs BERT-style BEiT (Bao et al., 2021) to pretrain in a self-supervised manner on IIT-CDIP dataset (Lewis et al., 2006) of 42 million document images and then fine-tune on a couple of classification and detection tasks.

However, to the best of our knowledge, there are several open research problems in developing SSL methods, particularly for tabular and structured document image domains, that require further study. Finding simple but effective self-supervised pretraining objectives, architectures, and fine-tuning strategies are examples of this. Keeping those questions in mind, our paper makes two main contributions:

- First, we introduce REGCLR, a new self-supervised framework that combines contrastive-based masked image modeling (MIM)[1] and regularized methods and is Vision Transformer (ViT) compatible. It is then instantiated by coupling masked autoencoders (MAE) and enhanced Barlow Twins (eBT) with configurable image augmentations; in principle, the MAE loss is augmented with the eBT loss as a regularization (prior) term. In fact, this simple SSL training loss objective is highly effective in reducing the capacity of learned/trained latent representation, especially in the balanced direction of MAE and eBT loss. MAE loss chooses to ignore irrelevant details, while eBT loss pushes relevant details to latent vectors (i.e., minimizing the ignored details), which is highly important for tabular representation learning.

- Second, we pretrain REGCLR on two distinct domains of document image datasets with rich tabular structures, the publicly accessible TableBank (Li et al., 2019) and more challenging internal EHRBank, and then validate the higher sample efficiency of the learned tabular representations in finetuning with TableBank and the superior detection performance with EHRBank in various practical downstream settings. A recent MIMDet (Fang et al., 2022) framework is used to extract multiscale features from the pretrained ViT encoder and decoder to use as a detection backbone without the need for architecture redesign or changes. By leveraging MIMDet, REGCLR outperforms the latest and purely supervised convolutional neural network (CNN) baseline by large margins.

2 REGCLR DESCRIPTION

2.1 INFORMATION THEORETIC FORMULATION

Problem Motivation and Notations In order to build a new unified SSL framework that incorporates the best of both worlds of MIM-based contrastive and regularized self-supervised approaches, we adopt an information-theoretic perspective to pose the problem. The main question is well expressed in the following quote from Yan Lecun's recent positional paper (LeCun, 2022, p. 23).

> [LeCun (2022)] —It is important to note that contrastive and regularized methods are not incompatible with each other, and can be used *simultaneously* on the same model. How would regularized methods apply to the standard architectures?—

To begin with, let X be the input image, (X_1, X_2, X_3) be the augmented or varied views of the image, and (Y_1, Y_2, Y_3) be the corresponding hidden/latent representation. Our goal is to learn the best representation Y that retains as much information about the input image X as possible given a reasonable representation constraint. The information bottleneck (IB) framework (Slonim et al., 2006) can be a good constraint to use in this case. To use the IB principle, denote n be the number of samples, d be the feature dimension, and $Z_2, Z_3 \in \mathbb{R}^{n \times d}$ be the projected features.

The role of the IB constraint is to compress X_2 into Z_2 while preserving the information about Z_3, and similarly X_3 into Z_3 while keeping the information about Z_2. To put it another way, the constraint

---

[1]We follow Yan LeCun's broader classification of contrastive methods introduced in (LeCun, 2022), which includes recent standard approaches (e.g., SimCLR (Chen et al., 2020)), MIM family (e.g., MAE (He et al., 2022)), and even Generative Adversarial Networks.
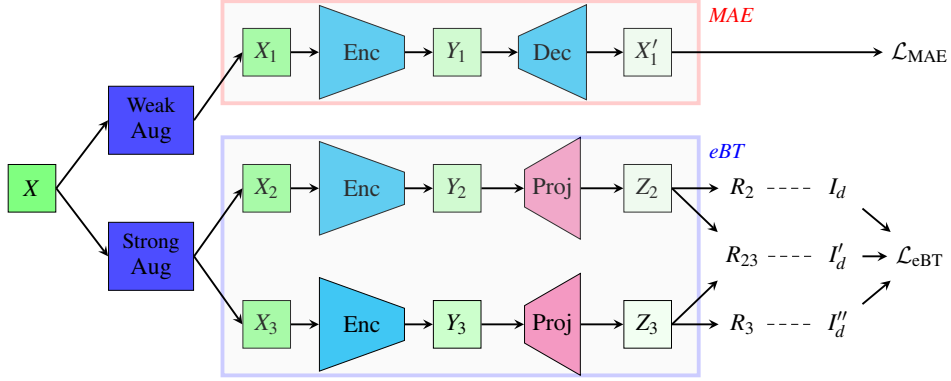
Figure 1: **REGCLR's self-supervised pretraining**. REGCLR is consisted of two branches: MAE and eBT. MAE branch uses weak augmentation and masking to obtain X_1 and then follows procedure introduced in MAE (He et al., 2022) to compute reconstructed loss for unmasked patches as $\mathcal{L}_{\text{MAE}}$ to obtain X_1'. In the eBT branch, strong augmentation is applied twice to get X_2 and X_3. Through the same ViT encoder used in the MAE branch, the encoded images are then projected into features Z_2 and Z_3. The proposed loss $\mathcal{L}_{\text{eBT}}$ computes three correlation matrices between Z_2 and Z_3, attempting to make each matrix near to an identity matrix. REGCLR's overall self-supervised pretraining is performed by jointly minimizing $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{eBT}}$ through optimizing both ViT encoder and decoder.

seeks to compress X_2 and X_3 as much as possible, while also making Z_2 and Z_3 as informative about each other as possible. For simplicity, we assume that Z_2 and Z_3 are two jointly multivariate Gaussian variables with zero means and covariance matrices of $K_2, K_3 \in \mathbb{R}^{d \times d}$ that are full rank. Note that $|K|$ denotes the determinant of K, $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of K, $||\cdot||_F$ is the Frobenius norm, and h is the differential entropy.

Training Loss Design From an information theoretic perspective, the problem can be written as an optimization problem in particular as:

$$\arg\max_{\theta} I(X_1; Y_1|\theta) - I(Z_2; X_2|\theta) - I(Z_3; X_3|\theta) + \alpha I(Z_2; Z_3|\theta). \tag{1}$$

Equivalently, if we omit θ for brevity, the target function $\mathcal{L}$ to minimize can be expressed as:

$$\mathcal{L} = \underbrace{h(X_1|Y_1)}_{\mathcal{L}_1} + \underbrace{(1-\alpha)h(Z_2) + (1-\alpha)h(Z_3) + \alpha h(Z_2, Z_3)}_{\mathcal{L}_2}, \tag{2}$$

this equality follows from the fact that $h(X_1)$ is a constant when the input distribution is fixed and $h(Z_2|X_2) = h(Z_3|X_3) = 0$. It has previously been shown (Vincent et al., 2008) that minimizing the expected reconstruction error equates to maximizing a lower bound on the first loss term $\mathcal{L}_1$ in (2). This holds true even when Y is a function of a corrupted input (e.g., masked out image). Therefore, for example, the mean squared error (MSE) between the augmented image X_1 and the reconstructed version X_1' denoted as $\mathcal{L}_{\text{MAE}}$ can be used as the equivalent loss function of $\mathcal{L}_1$.

We can further express the remaining three loss terms $\mathcal{L}_2$ in (2) as[2]

$$\log\left(|K_2|\right) + \log\left(|K_3|\right) + \beta \log\left(|K_{23}|\right), \tag{3}$$

where $\beta \triangleq \left(\frac{\alpha}{1-\alpha}\right)$. Then, $\log\left(|K|\right)$ can be upper bounded to be computationally convenient to use as follows:

$$\log\left(|K|\right) = \log\left(\prod_{i=1}^{n} \lambda_i\right) = \sum_{i=1}^{n} \log \lambda_i < \sum_{i=1}^{n} \lambda_i^2 = ||K||_F^2. \tag{4}$$

Combining (4) with (3), we obtain the further simplified version of $\mathcal{L}_2$ as

$$||K_2||_F^2 + ||K_3||_F^2 + \beta ||K_{23}||_F^2. \tag{5}$$

---

[2]Recall that the entropy of an d dimensional Gaussian variable is $h(Z) = \frac{1}{2}\log\left((2\pi e)^d |K|\right)$.
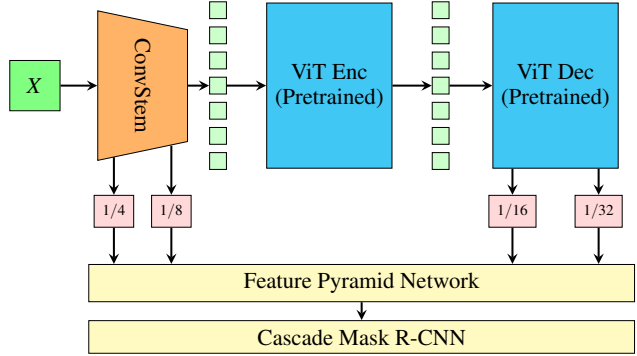
Figure 2: **REGCLR's detection via ViT backbone**. The MIMDet architecture is leveraged by REGCLR, as described in Section 2.2, to transfer both the pretrained ViT encoder and decoder to detection with Cascade Mask R-CNN without requiring the ViT network redesign. In MIMDet, a randomly initialized ConvStem replaces the pretrained large kernel PatchStem, and the ConvStem's intermediate features can directly be used as higher resolution inputs for a standard feature pyramid network.
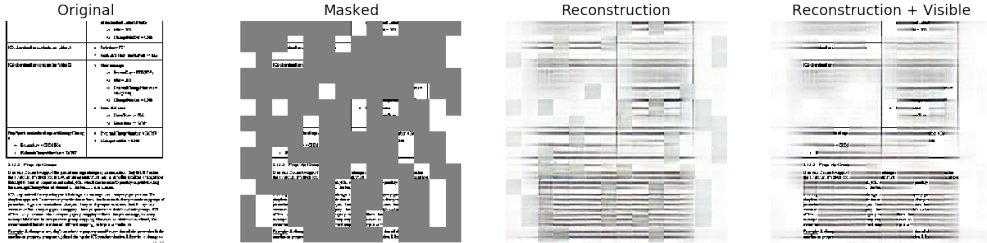

Figure 3: **Visualization of the MAE branch's reconstruction** for a TableBank Word test set example (with a masking ratio of 75%). The MAE arm reconstructs the table's structure as well as the most of the masked boundaries, demonstrating that it can learn tabular semantic information from only unmasked patches while ignoring irrelevant document details (the eBT arm, on the other hand, is anticipated to keep the ignored details to a minimum).

Since Z_2 and Z_3 are zero means, the cross-covariance matrix K_{23} becomes the cross-correlation matrix R_{23} assuming normalization without loss of generality. Similarly, auto-covariance matrices K_{Z_2} and K_3 become the auto-correlation matrices R_2 and R_3, respectively. Because there is a trivial solution of zero matrices when minimizing (5), we can circumvent this trivial solution by urging the diagonal terms to be 1 and all the off-diagonal terms to be zero. Hence, the resulting objective denoted as $\mathcal{L}_{\text{eBT}}$ is given by

$$
\begin{aligned}
\mathcal{L}_{\text{eBT}} = {} & \nu_1 \sum_i \left(1 - R_{2,ii}\right)^2 + \nu_1 \sum_i \left(1 - R_{3,ii}\right)^2 + \nu_2 \sum_i \left(1 - R_{23,ii}\right)^2 \\
& + \mu_1 \sum_i \sum_{j \neq i} \left(R_{2,ij}\right)^2 + \mu_1 \sum_i \sum_{j \neq i} \left(R_{3,ij}\right)^2 + \mu_2 \sum_i \sum_{j \neq i} \left(R_{23,ij}\right)^2,
\end{aligned}
\tag{6}
$$

where ν_1, ν_2, μ_1 and μ_2 are hyper-parameters controlling diagonal and off-diagonal terms of each matrix. We use $\mu_1 = 0.5$, $\mu_2 = 1.0$, $\nu_1 = 0.00255$ and $\nu_2 = 0.0051$ to balance between auto-correlation and cross-correlation terms during training. After introducing $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{eBT}}$, the overall training loss is now straightforward to compose as:

$$
\mathcal{L}_{\text{REGCLR}} = \mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{eBT}}.
\tag{7}
$$

2.2 REGCLR FRAMEWORK

Self-Supervised Pretraining One possible instantiation is to use the new objective, $\mathcal{L}_{\text{REGCLR}}$ in (7), for self-supervised pretraining, and build the model with two branches, MAE and eBT, as shown

Table 1: **TableBank and EHRBank dataset statistics**. TableBank includes Word and Latex document image sets, while internally curated EHRBank has EHR screen images in the form of tables, table columns, and GUI elements sets. Because labeling all GUI elements in the EHR images is more costly, we have fewer samples in the GUI dataset. Note that the *unlabeled* EHRBank Screenshot dataset (shown in the last row) is used only for pretraining.

	Train	Val	Test
TableBank Word	73,383	2,735	2,281
TableBank Latex	187,199	7,265	5,719
EHRBank Table	1,917	411	209
EHRBank Column	1,194	255	208
EHRBank GUI	157	38	157
EHRBank Screenshot (*unlabeled*)	28,121		

in Figure 1. For the MAE branch, we choose to apply weak augmentation to the input image and then randomly mask out the selected patches. Only unmasked patches are fed into the ViT encoder, and subsequently the masked patches are reconstructed by the Vit decoder using the MSE calculated over the masked patches as the loss function $\mathcal{L}_{\text{MAE}}$ per the original MAE design (see Figure 3 for an example of image visualization).

Secondly, the eBT branch operates on the cross embedding of the input image's two strongly augmented versions X_2 and X_3. In contrast to the previous instantiation of the IB principle presented in BT loss Zbontar et al. (2021), which computes only the cross-correlation matrix, $\mathcal{L}_{\text{eBT}}$ computes a cross-correlation matrix as well as two auto-correlation matrices bringing these three matrices close to the identity ones. We believe that this is a more natural and fundamental improvement over BT loss because augmented images typically carry a significant amount of input image information, implying that the conditional independence assumption of the BT loss (Zbontar et al., 2021, Figure 6) that leads to minimizing $I(Z_2; X_2|\theta)$ while maximizing $I(Z_3; X_3|\theta)$ is suboptimal.

Detection via ViT Backbone For detection, as illustrated in Figure 2, we combine a ViT encoder and decoder pretrained in a self-supervised manner with MIMDet (Fang et al., 2022) to serve as the detection backbone and leverage Cascade Mask R-CNN (Cai & Vasconcelos, 2019), which is the common architecture in supervised state-of-the-art systems (e.g., Prasad et al. (2020); Agarwal et al. (2021)). Compared to previous representative approaches of adapting vanilla ViT for object detection (e.g. Li et al. (2021)), MIMDet replaces the pretrained patchify stem (PatchStem) with a compact convolutional stem (ConvStem) without further upsampling or redesigns, resulting in a ConvNet-ViT hybrid multi scale feature extractor that requires far fewer epochs in the fine-tuning procedure.

3 EXPERIMENTS

3.1 DATASETS

TableBank Dataset TableBank is a publicly available image-based table dataset for Word and Latex documents on the Internet. It contains 417K high-quality tables labeled with weak supervision split across Word and Latex sets. Note that the Word document set contains English, Chinese, Japanese, Arabic, and other languages, whereas the Latex document set is primarily in English. Details of the data split can be found in Table 1 and one can refer to the top row of Figure 5 for image samples with annotated bounding boxes.

EHRBank Dataset Beyond the TableBank dataset, we systematically curated an internal dataset of screen images from real-world EHR systems, which consists of screenshots collected by bots as they navigate the EHR systems of ten US health systems from various EHR providers.[3] These screens were then carefully selected to include essential template screens and labeled by an internal team of labelers, including tables, table columns, and other GUI elements[4]. Details of the dataset's

---

[3]The collection and use of this EHRBank dataset required HIPPA compliance training.
[4]Appendix A contains brief descriptions of all twelve classes of GUI elements labeled.

composition can be found in Table 1. For the Column dataset, we further cropped these images to just tables, removing examples that were occluded from overlays. Visual examples of these datasets with human-annotated bounding boxes can be found in the first column of Figure 6.

In addition to the aforementioned supervised dataset with labeled objects, we also collected a *unlabeled* Screenshot dataset from the worklogger system of a single US health system, which records videos of hospital staff interacting with the EHR system. We randomly sampled screen recordings from the database, then the randomly sampled frames from the randomly sampled videos. The frames were then split into two because most staff members had a dual monitor set up, and these frames were labeled by the same internal team of labelers, and frames without tables or columns or corrupted due to the splitting process were removed. This dataset contains 28,121 PNG images with a resolution of 1920x1080, which represent 10.8% of the volume of the TableBank Train set.

3.2 SETTINGS

Pretraining For TableBank experiments, we pretrain REGCLR on the combined train set of 260,582 document images from the Word and Latex sets. For EHRBank experiments, we have an unsupervised screenshot set of 28,121 images to be used for pretraining. For both experiments, we used adaptive binarization, random resize cropping, and random horizontal flipping as a weak augmentation to train the MAE branch. For a strong augmentation to train the eBT arm, we further add smudge transformation and dilation, which are introduced in (Prasad et al., 2020). The input images are resized to 224 by 224 in size, and the pretraining phase takes 500 epochs to complete with a batch size of 80.

Table 2: **Results of table detection on the TableBank test set** (in AP and AP_{75}). MAE denotes the representative SSL pretraining baseline, while ResNet (He et al., 2016) stands for the purely supervised baseline using the ResNet-152 backbone, with Cascade Mask R-CNN. The bold value represents the best (highest) value for each column metric. All baselines are outperformed by the proposed REGCLR.

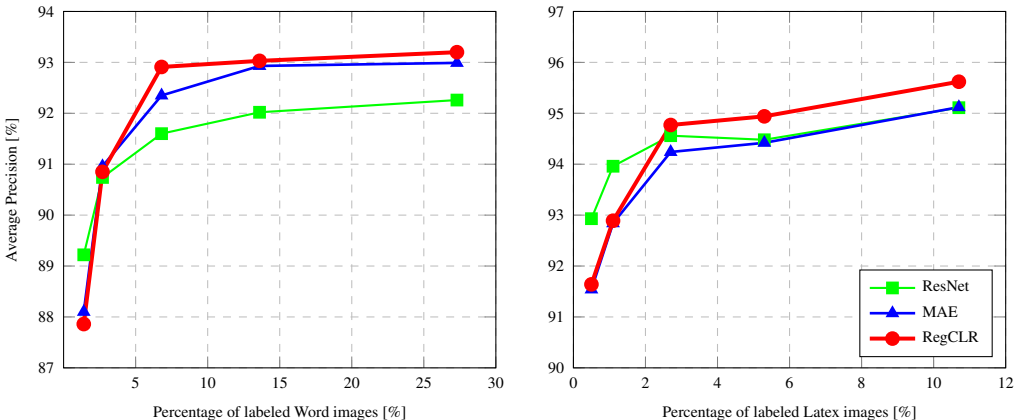|	Word		Latex	
	AP	AP_{75}	**AP**	AP_{75}
ResNet (supervised baseline)	95.42	95.78	97.32	98.62
MAE (self-supervised baseline)	95.94	96.16	97.63	98.70
REGCLR (our method)	**96.03**	**96.22**	**97.68**	**98.75**

Figure 4: **Benefits of REGCLR in low labeled data regimes** (Left) Word set and (Right) Latex set. Each datapoint in the graph represents the number of labeled subsets of 1k, 2k, 5k, 10k, and 20k. REGCLR begins performing best after a 2-3% subset of labeled train data is provided for fine-tuning.
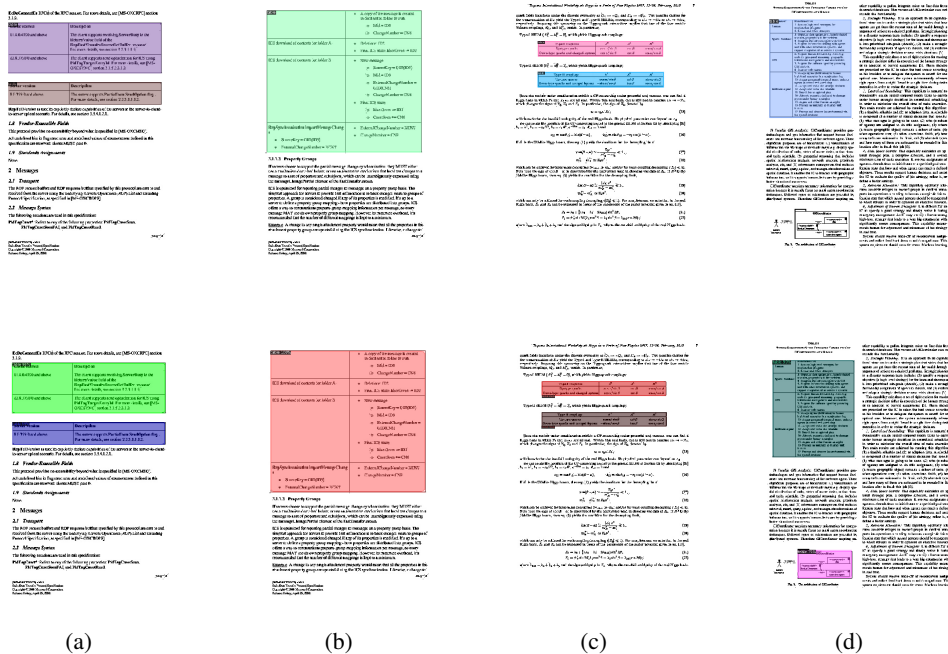
Figure 5: **TableBank test set examples** (a-b) Word test set and (c-d) Latex test set. The top row depicts examples with annotated boundaries, while the bottom row presents the corresponding detection predictions by REGCLR .

Our encoder and decoder model uses the ViT base backbone of a 12-layer Transformer with 12 self-attention heads. Our implementation of pretraining is built on the repo provided in MAE (He et al., 2022), and the projector is implemented as a 2 layer multilayer perceptron with 1024 dimensional output. Overall, pretraining on TableBank uses 4 Nvidia A100 GPUs for 20 hours, and pretraining on EHRBank uses 4 Nvidia V100 GPUs for 40 hours.

Fine-tuning For both experiments, we use MIMDet method to adapt pretrained ViT weights to detection. For all experiments except table detection in EHRBank dataset, the input is resized so that the shortest side is between 480 and 800 pixels and the longest side is no more than 1333 pixels. For the table detection in EHRBank dataset, the input is resized so that it has the shortest side between 307 and 512 pixels and the longest side no more than 853 pixels. We use adaptive binarization as preprocessing for input images, and our implementation for fine-tuning is based on the repo provided in MIMDet (Fang et al., 2022), with a 50% sampling ratio for both training and inference.

Baselines We carefully chose an MAE as a representative self-supervised pretraining baseline and ResNet with the ResNet-152 backbone as a fully supervised baseline. For a fair comparison, both baselines, like REGCLR, use the Cascade Mask R-CNN architecture. All methods for TableBank datasets are trained for 125,000 iterations with a batch size of 12. All methods for table detection in the EHRBank datasets are trained for 10,000 iterations with a batch size of 16 in the experiments. For other tasks in the EHRBank datasets, we train for 20,000 iterations with a batch size of 4.

3.3 RESULTS

TableBank Dataset The prediction results of the TableBank table detection are shown in Table 2 in AP (mAP @ IOU [0.50:0.95]) and AP$_{75}$ (mAP @ IOU 0.75) with the results of two baselines. We can see that our method REGCLR outperforms the other self-supervised and fully supervised baselines. The visualization of the table detection predictions by REGCLR can be found in the second row of Figure 5. We conduct further evaluations to better understand our method, REGCLR, especially how it performs when the number of labeled data decreases, particularly in low labeled data regimes.

Table 3: **Results of GUI elements detection on the EHRBank Table and Column test sets** (in AP and AP$_{75}$). REGCLR performs best when pretraining with the EHRBank Screenshot dataset, increasing AP scores relatively by 4.8% for Table and 11.8% for Column over the supervised baseline, as seen by comparing the first and second rows. Interestingly, despite pretraining with approximately 10% volumes of TableBank, RegCLR fast approaches the best cross-domain transfer results from TableBank to EHRBank in the last row.

Pretrain on	Method	Table		Column	
		AP	**AP$_{75}$**	**AP**	**AP$_{75}$**
N/A	**ResNet**	40.53	44.46	61.43	67.07
EHRBank Screenshot	**REGCLR**	**42.46**	**45.32**	**68.68**	**75.17**
	MAE	36.78	39.05	64.67	71.09
TableBank (cross-domain)	**REGCLR**	40.96	43.47	67.77	75.06
	MAE	**43.99**	**48.77**	**69.83**	**77.29**

Table 4: **Results of GUI elements detection on the EHRBank GUI test set** (Left) GUI elements detection AP and AP$_{75}$ in the GUI test set, and (Right) the detection AP of each GUI element class in the test set. As shown in the left table, REGCLR outperforms the baselines in both AP and AP$_{75}$ metrics on average. Specifically, REGCLR outperforms the baselines in eight of twelve categories, while MAE performs even worse than ResNet in six.

	ResNet	MAE	RegCLR
AP	43.37	45.69	**48.17**
AP$_{75}$	48.25	50.80	**54.10**

	ResNet	MAE	REGCLR
Button	**38.42**	33.02	33.13
Dropdown	**48.02**	44.33	43.14
Dropdown_group	43.25	41.71	**43.33**
Horizontal_scrollbar	25.12	36.88	**38.78**
Overlay	52.92	**65.58**	62.12
Tab	**37.62**	34.13	36.46
Tab_group	29.94	30.35	**32.77**
Table	62.72	66.10	**72.64**
Table_column	46.31	50.44	**54.74**
Text_box	55.03	54.56	**58.06**
Text_input_group	42.82	41.38	**44.87**
Vertical_scrollbar	37.28	47.32	**52.00**

To do so, we evaluate the test set when fine-tuning on the subsets of 1k, 2k, 5k, 10k, and 20k. Note that all methods are trained for 10,000 iterations with a batch size of 12 for this experiment. As shown in Figure 4, REGCLR outperforms the baselines soon after fine-tuning on a subset of 2-3% labeled train set, AP scores improve quickly even with considerably fewer labeled images (i.e., less than 10%) provided than the unlabeled set, and the gain decreases as more labeled data are added.

EHRBank Dataset We then evaluate REGCLR on internal EHRBank dataset which has higher values in our production scenario. As shown in Table 3, when pretrained on *unlabeled* EHRBank, our method outperforms the baselines in both Table and Column detection, increasing relative AP scores by 4.8% and 11.8% respectively over the supervised baseline, as seen by comparing the first and second rows. Furthermore, even though pretrained with only around 10% of the TableBbank volume, REGCLR quickly approaches the best cross-domain transfer performance from TableBank to EHRBank, as shown in the last row.

Additionally, it is worth noting that MAE performs worse than even ResNet on Table when pretrained on EHRBank (by comparing the first and third rows). It does, however, transfer better than REGCLR in scenarios involving cross-domain transfer from public TableBank to private EHRBank (by comparing the last two rows). In the future, we intend to investigate how quickly detection performance improves as the unlabeled data volume scales, as well as how effectively pretrained weights transfer across domains in the context of tabular rich images, so that they can be applicable to other document format datasets (e.g., Word to GUI and vice versa).

Table 4 presents the results in GUI elements detection, in which our method pretrained on EHRBank Screenshot again produces the highest overall detection scores compared to the base-

lines. This demonstrates that REGCLR can provide performance improvement even in more complicated scenarios with a larger number of classes and possible occlusions between different GUI elements. The performance variation across different GUI categories is also presented in the right table of Table 4. Visualization of the predictions in the EHRBank can be found in Figure 6.

4 RELATED WORK

SSL Methods for Natural Images Learning good visual representation in a self-supervised manner has been a rising paradigm in computer vision over the past few years. More recently, a few families of approaches have emerged. i) Contrastive learning is trained to extract meaningful features by mapping different images to distant points and the same image under various image augmentations to close points in the feature space. As an example, SimCLR (Chen et al., 2020) defines the feature distance by cosine similarity and utilizes InfoNCE proposed by (Oord et al., 2018) as a loss function. ii) Self-distillation serves as another powerful approach that achieves effective representation learning while alleviating the need for large contrastive batch sizes that are often required for contrastive methods. BYOL (Grill et al., 2020) simply minimizes the similarity loss between the encoder and its momentum target version to achieve this. iii) Feature regularization methods, alternatively, optimize over the auxiliary regularizing loss to directly learn features with desired properties. In this family, Barlow Twins (Zbontar et al., 2021) minimizes off-diagonal terms in the covariance matrix between two augmentations to obtain low redundancy representations.

Transformer Architecture for SSL Methods In addition to the advancement of the SSL algorithm, the Transformer (Vaswani et al., 2017) architecture also presents itself as a powerful general architecture in computer vision. It brings advantages over CNN, which has dominated the field as a de facto network structure since the emergence of deep learning. Using the ViT architecture, MAE (He et al., 2022), in particular, generalizes masked language modeling (MLM) popular in natural language processing to MIM in computer vision. MAE feeds unmasked patches to the transformer and predicts masked patches as a pretext task. Despite its simplicity, MAE shows strong performance on various computer vision tasks, and specifically, MAE has been found to produce features that better support the downstream object detection task (Li et al., 2021) that was the inspiration for our study.

SSL Methods for Document Images Beyond natural images, self-supervised pre-training with transformers is actively used in vision-based document AI tasks. In particular, DiT (Li et al., 2022), which uses BEiT (Bao et al., 2021), demonstrates MIM's potential to extract high-quality visual representations from document images. LayoutLMv3 (Huang et al., 2022) further extends pretraining to a multimodal setup with unified text and image masking, including MLM for texts, MIM for images, and a novel word-patch alignment that predicts whether the corresponding image patch for any given word is masked. In the direction of multimodal pretraining for document images, VL-CDoC (Bakkali et al., 2022) proposes using contrastive loss to align text and vision cues, producing strong performance when both text and image are available. In the context of the images of medical-health documents that we are interested in, DEXTER (PR et al., 2022) fine-tunes the TableBank from pretrained MS COCO weights, while no SSL method is employed.

5 CONCLUSION

In this paper, we present REGCLR, a novel self-supervised framework for learning tabular representation and downstream detections, which had received little attention than natural image domain. Our brand-new framework combines contrastive and regularized self-supervised methods and has been pretrained on both public and private domain tabular rich images. We demonstrate that REG-CLR outperforms previous self-supervised pretraining and fully supervised baselines by a large margin in various real-world contexts, with high sample efficiency for fine-tuning. It is left to future study how quickly the quality of the representation learned by our method improves as the volume of *unlabeled* images scales, which is a common and practical scenario in production. We believe that this study is an important step towards semantic comprehension of real-world document images, and it will be interesting to see how this vision-based framework can be expanded to include textual content without manual data annotation.

REFERENCES

Madhav Agarwal, Ajoy Mondal, and C. V. Jawahar. Cdec-net: Composite deformable cascade network for table detection in document images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. Vl-cdoc: Vision-language contrastive pre-training model for cross-modal document classification. *arXiv preprint arXiv:2205.12029*, 2022.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Douglas Burdick, Alexandre V Evfimievski, Yannis Katsis, Marina Danilevsky, and Nancy Wang. Ibm's tutorial on table extraction and understanding for scientific and enterprise applications, 2020. URL https://researcher.watson.ibm.com/researcher/view_group.php?id=10211.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019. ISSN 1939-3539. doi: 10.1109/tpami.2019.2956516. URL http://dx.doi.org/10.1109/tpami.2019.2956516.

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, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR 2021)*, 2021.

Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2022)*. IEEE, 2022.

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.

Yan LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

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.

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

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: A benchmark dataset for table detection and recognition. *arXiv preprint arXiv:1903.01949*, 2019.

Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Nandhinee PR, Harinath Krishnamoorthy, Koushik Srivatsan, Anil Goyal, and Sudarsun Santhiappan. Dexter: An end-to-end system to extract table contents from electronic medical health documents. *arXiv preprint arXiv:2207.06823*, 2022.

Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *CVPR 2020 Workshop on Text and Documents in the Deep Learning Era*, 2020.

Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate Information Bottleneck. *Neural Computation*, 18(8):1739–1789, 08 2006.

Shubhangi Vashisth, Anthony Mullen, Stephen Emmott, and Alejandra Lozada. Gartner market guide for intelligent document processing solutions, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *Technical Report*, 2008.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

A APPENDIX

EHRBank GUI Dataset: We collect and label GUI screenshots with twelve classes: button, dropdown, drowdown group, overlay, tab, tab group, table, table column, text box, text input group, horizontal scrollbar and vertical scrollbar.

- Button: Small, usually rectangle shape, clickable elements in the screen with or without text.
- Overlay: A separate piece of user interface that appears to be the front layer (e.g., a popup window).
- Dropdown and dropdown group: A clickable rectangle shape UI element or a group of such UI elements that can be clicked to expand to multiple choice selection interface.
- Tab and tab group: A clickable rectangle shape UI element or a group of such UI elements that can switch between different pages.
- Table and table column: Tabular element and column of such tabular element.
- Text box and text input group: Box that can type in text and group of such elements.
- Horizontal scrollbar and vertical: scrollbar: Horizontal and vertical draggable UI element that can move the displaying window accordingly.
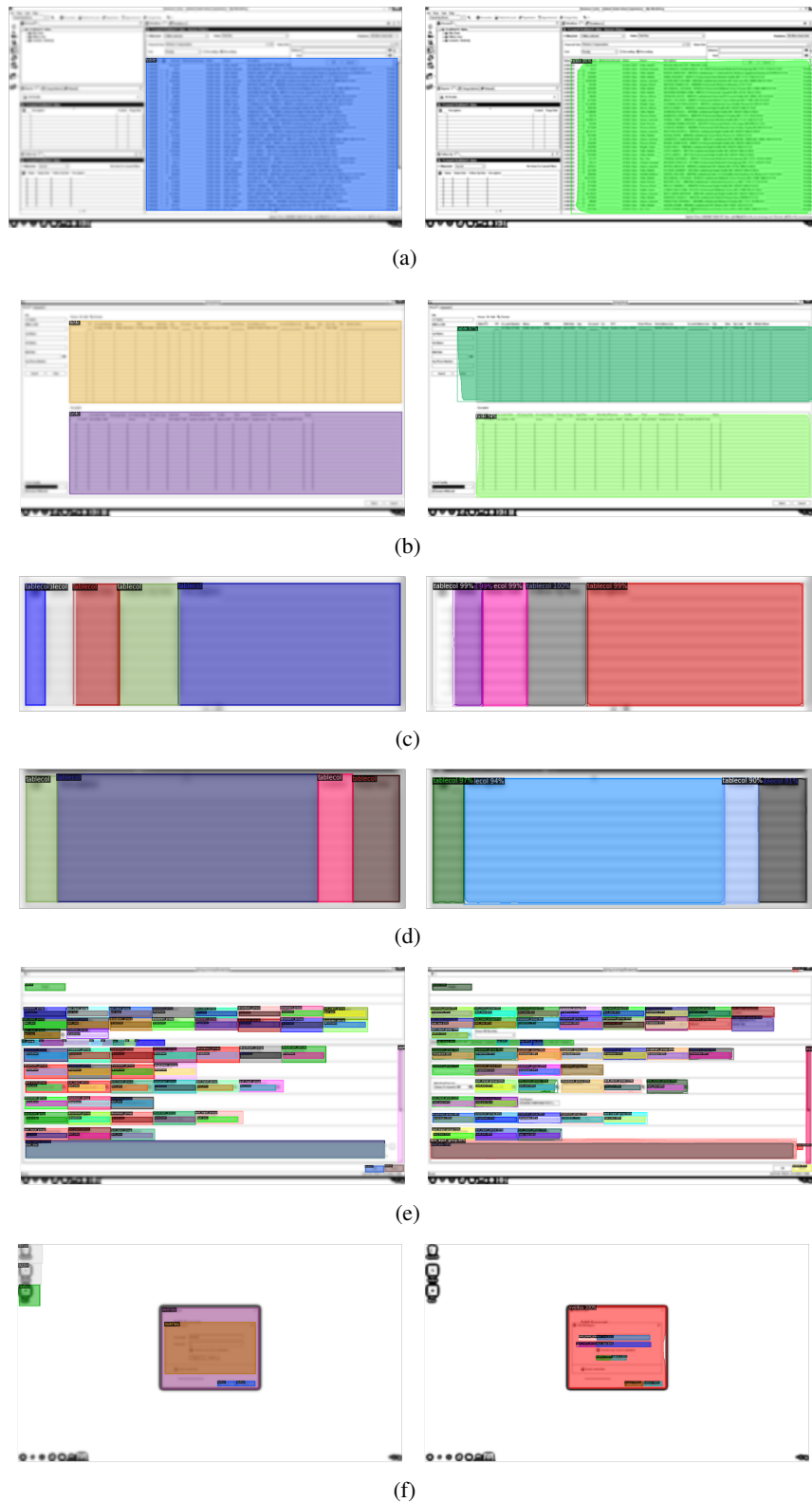
Figure 6: **EHRBank test set examples** (a-b) Table test set, (c-d) Column test set, and (e-f) GUI test set. The first column shows examples with human-annotated boundaries, while the second column presents the corresponding detection predictions by REGCLR. All images are purposefully scrubbed so that no PHI information about the patient is disclosed.