Visual Information Extraction in the Wild: Practical Dataset and End-to-end Solution

Jianfeng Kuang[1*], Wei Hua[1*], Dingkang Liang[1], Mingkun Yang[1], Deqiang Jiang[2], Bo Ren[2], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology
[2] Tencent YouTu Lab
{kuangjf, whua_hust, dkliang,
yangmingkun, xbai}@hust.edu.cn
{dqiangjiang,timren}@tencent.com

Abstract. Visual information extraction (VIE), which aims to simultaneously perform OCR and information extraction in a unified framework, has drawn increasing attention due to its essential role in various applications like understanding receipts, goods, and traffic signs. However, as existing benchmark datasets for VIE mainly consist of document images without the adequate diversity of layout structures, background disturbs, and entity categories, they cannot fully reveal the challenges of real-world applications. In this paper, we propose a large-scale dataset consisting of camera images for VIE, which contains not only the larger variance of layout, backgrounds, and fonts but also much more types of entities. Besides, we propose a novel framework for end-to-end VIE that combines the stages of OCR and information extraction in an end-to-end learning fashion. Different from the previous end-to-end approaches that directly adopt OCR features as the input of an information extraction module, we propose to use contrastive learning to narrow the semantic gap caused by the difference between the tasks of OCR and information extraction. We evaluate the existing end-to-end methods for VIE on the proposed dataset and observe that the performance of these methods has a distinguishable drop from SROIE (a widely used English dataset) to our proposed dataset due to the larger variance of layout and entities. These results demonstrate our dataset is more practical for promoting advanced VIE algorithms. In addition, experiments demonstrate that the proposed VIE method consistently achieves the obvious performance gains on the proposed and SROIE datasets. The code and dataset will be available at https://github.com/jfkuang/CFAM.

Keywords: Visual Information Extraction · Document Understanding · Document Semantics Extraction.

---

[*] Equal contribution. [†]Corresponding author.

Work done when Wei Hua was an intern at Tencent.

(a) Existing framework

(b) Our framework

Fig. 1. The overview of the existing framework and our framework. The existing framework directly adopts OCR features as the input of an information module, while our framework additionally proposes a contrast-guided feature adjustment module to the existing framework.

Table 1. Statistics of the representative VIE datasets and our proposed dataset.

Dataset	Year	Number of images			Number of entities	Number of instance	Language	Scene Type
		Train	Test	Total				
SROIE [7]	ICDAR2019	626	347	973	4	52,316	EN	Receipt
EPHOIE [30]	AAAI2021	1,183	311	1,494	10	15,771	CN	Examination Paper
POIE (**Ours**)	-	**2,250**	**750**	**3,000**	**21**	**111,155**	EN	Product

1   Introduction

Visual Information Extraction (VIE), automatically extracting structured information from visually-rich document images, is an essential step towards document intelligence [42,31,9]. It involves extracting values of entities from images and reasoning about their relations, which is a challenging cross-domain problem that requires both visual and textual understanding, such as table comprehension and document analysis [10,32,39,29]. This task becomes more difficult in real-world scenarios where diverse layouts, noisy backgrounds, and large variances of entities exhibit in the wild.

Datasets play an essential role in data-driven problems. In the document VIE area, a number of datasets [7,21,30,2], have been proposed. Specifically, SROIE [7] is the most widely used dataset, in which the images are scanned receipts printed in English. EPHOIE [30] is collected for Chinese document VIE, which is thus unsuitable for assessing previous methods for English VIE. These

Entities	Forms of entities
CE-PS	Calories per serving, Energy per serving,…
CAR-PS	Total Carb. per serving, Total Carbohydrate per serving,…
PRO-PS	Protein per serving, Protein per bottle,…
SO-D	Sodium % Daily Value*, Sodium %DV*,…
...	...
...	...
TF-D	Total Fat %Daily Value*, Total Fat %DV*,…
CE-P1	Calories per 100g, Calories per 100ml,…
CAR-P1	Total Carb. per 100g, Total Carbohydrate per 100ml,…
PRO-P1	Protein per 100g, Protein per 100ml,…

Our dataset   (a)   Other datasets   (b)

Fig. 2. (a) exhibits a few typical examples from our dataset and others. (b) shows some typical entities and their different forms in our dataset. Note that the actual entity name is very long, so we use the abbreviation of the entity name. More details of our POIE dataset can be found at https://github.com/jfkuang/CFAM

datasets are composed of only document images with moderate scale. For more general scenarios, like VIE in the wild, the captured images have larger variances in layouts, backgrounds, and fonts, as well as various entities. Consequently, the available datasets cannot fully reveal the challenges of real-world applications.

To better study VIE in the wild, we collect a large-scale challenging dataset called **P**roducts for **O**CR and **I**nformation **E**xtraction (POIE), which consists of camera images from Nutrition Facts label of products in English. Compared with existing VIE datasets [7,30], POIE has several distinct characteristics. First, it is the largest VIE dataset with high-quality annotations, where 3,000 images with 111,155 text instances are collected. As indicated in Table 1, it is over 2× larger than the previous largest public VIE dataset. Second, POIE is particularly challenging, where the images are in diverse layouts and distorted with various folds, bends, deformations as well as perspectives. We show some typical examples in Figure 2 (a). Third, POIE has up to 21 kinds of entities, some of which come in numerous forms. For example, in Figure 2 (b), several forms of some entities are given. It is quite common in the real-world applications that each entity presents in various forms, demonstrating VIE in the wild is more difficult.

We observe that previous end-to-end approaches [40,30] achieve lower performance on POIE, especially on the information extraction task, due to complex layouts and variable entities. Besides, these methods directly adopt OCR features as the input of the following information extraction module. We argue that the reason is there is a severe semantic gap between the tasks of OCR and information extraction while directly feeding OCR features into the following information extraction module. Recently, we have witnessed the rise of contrastive

learning in computer vision, which has received extensive attention in various visual tasks [12,24,1,6,37] while has rarely been noticed in VIE tasks. In this paper, we propose a novel end-to-end framework for VIE. We adopt contrastive learning to effectively establish the connections between the tasks of OCR and information extraction. Specifically, the key component of our framework is a plug-and-play Contrast-guided Feature Adjustment Module (CFAM), as shown in Figure 1. In CFAM, we design the feature representation for OCR and information extraction (i.e., instance features and entity features for OCR and information extraction). As a result, CFAM constructs a similarity matrix reflecting the relations between entity features and instance features to adjust the instance features more appropriately for the following information extraction task.

In summary, the main contributions of this paper are two-fold: 1) we propose a large-scale dataset consisting of camera images with variable layouts, backgrounds, fonts, and much more types of entities for VIE in the wild. 2) we design a novel end-to-end framework with a plug-and-play CFAM for VIE, which adopts contrastive learning to narrow the semantic gap caused by the difference between the tasks of OCR and information extraction.

2   Related Works

2.1   VIE Datasets

Currently, a few datasets [7,26,21,30,2] are proposed for VIE on document images. Specifically, SROIE [7] is the most widely used dataset, which significantly promotes the development of this field. The images of SROIE are scanned receipts in printed English. Each image is associated with complete OCR annotations and the values of four key text fields. Besides, EPHOIE [30] is collected for Chinese document VIE, where the images have complex backgrounds and diverse text styles. The details of these datasets mentioned above are shown in Table 1. We can observe that these existing datasets are composed of only document images with moderate scale. To better explore VIE in the wild, a large-scale dataset with more entities and larger variances in layouts, backgrounds, and fronts is urgently required.

2.2   VIE Methods

According to the pipeline of VIE, existing approaches [34,38,2,27,16] can be divided into two categories: methods in two-stage and end-to-end paradigms. Most works with a two-stage pipeline focus on the second stage for information extraction. In these methods, the OCR results are first obtained via an individual OCR extractor. In Post-OCR parsing [8], the coordinates of text bounding boxes are applied during the second stage. To better model the layout structure and visual cues of documents, LayoutLM [34] employed a pre-training strategy inspired by BERT. GraphIE [23] and PICK [38] constructed a graph according

to the OCR results and applied Graph Neural Networks (GNNs) to extract the global representation for further improvement. In CharGrid [11], CNN is used to integrate semantic clues and the layout information. MatchVIE [27] noticed the importance of modeling the relationships between entities and text instances. However, it required additional annotations of all key-value pairs. All of these methods concentrated on the context modeling between OCR results in the second stage but ignored the accumulative errors from the preceding OCR module.

Recently, an increasing number of VIE methods have been proposed in an end-to-end fashion. EATEN [3] first generated feature maps from input images and attached several entity-aware decoders to predict all entities. However, it can only handle documents with a fixed layout. TRIE [40] was an end-to-end trainable framework to solve the VIE task, which focused more on the performance of entity extraction. VIES [30] improved each part of VIE, like text detection, recognition, and information extraction, but incurred obvious costs. Donut [13] can directly extract information from the input images without text spotting, while it needed massive data for pre-training. All of the above methods directly took OCR features as the input of the following information extraction module. Different from the previous end-to-end approaches, we propose a novel framework in an end-to-end manner to establish the connections between the tasks of OCR and information extraction.

2.3   Contrastive Learning

Contrastive learning [12], a typical way for visual representation learning, has attracted lots of attention in many fields and obtained great progress [24,36,1]. Contrastive learning allows samples of positive pairs to lie close together in the latent space, while samples belonging to negative pairs are repelled in the latent space. CLIP [24] is a representative work for vision-language pre-training via contrastive learning on image-text pairs. Since then, plenty of multi-modal contrastive learning methods have been proposed [22,25,41,15,28,35]. However, to the best of our knowledge, contrastive learning has not been thoroughly studied in the VIE field. In this paper, we propose to use contrastive learning to supervise the construction of relations between the tasks of OCR and information extraction. Different from the above approaches, where the image features and text features are taken as inputs, only the features of text instances are fed into the information extraction module in the VIE task. Therefore, how to represent OCR features as well as entity features, and construct the relations between them to narrow the semantic gap should be of great importance for end-to-end VIE.

3   POIE Dataset

3.1   Data Collection

Products for **O**CR and **I**nformation **E**xtraction (POIE) dataset derives from camera images of various products in the real world. The images are carefully

selected and manually annotated. Our labeling team consists of 8 experienced labelers. We first crop the nutrition tables from product images and adopt multiple commercial OCR engines (Azure and Baidu OCR) for pre-labeling. Then we use LabelMe *** to manually check the annotation of the location as well as transcription of every text box, and the values of entities for all the text in the images and repaired the OCR errors found. After discarding low-quality and blurred images, we obtain 3,000 images with 111,155 text instances.

3.2   Data Characteristics

To the best of our knowledge, POIE is the largest dataset with both OCR and VIE annotations for the end-to-end VIE in the wild. The images in POIE contain Nutrition Facts labels from various commodities in the real world, which have larger variances in layout, severe distortion, noisy backgrounds, and more types of entities than existing datasets. The comparison of dataset statistics is shown in Table 1. Existing datasets mainly consist of document images with insufficient diversities of layout, background disturbs, and entity categories. Therefore, they cannot fully illustrate the challenges of some practical applications, like VIE on the Nutrition Facts label. Compared with these datasets, POIE contains images with variable appearances and styles (such as structured, semi-structured, and unstructured styles), complex layouts, and noisy backgrounds distorted by folds, bends, deformations, and perspectives (typical examples are shown in Figure 2 (a)). Moreover, the types of entities in POIE reach 21, and a few entities have different forms (some typical entities with various forms are shown in Figure 2 (b)), which is very common and pretty challenging for VIE in the wild. Besides there are often multiple words in each entity, which appears zero or once in every image. These properties mentioned above can help enhance the robustness and generalization of VIE models to better cope with more challenging applications.

3.3   Data Split and Evaluation Protocol

POIE is divided into training and testing sets, with 2,250 and 750 images, respectively. We use the performance of detection (DET), recognition (REC), and information extraction (IE) in the end-to-end pipeline to evaluate all methods on the proposed POIE. Following the settings of SROIE [7] and EPHOIE [30], F1-Score is applied as the evaluation metric for the three tasks.

4   Our Method

The overall pipeline of our method is illustrated in Figure 3. Different from other methods, we introduce contrastive learning into the end-to-end trainable framework for effectively bridging the modules of OCR and information extraction. Specifically, it is composed of a text detector, a Contrast-guided Feature Adjustment Module (CFAM), a recognizer module, and an information extraction module. Given an input image,

---

*** https://github.com/wkentaro/labelme

Fig. 3. The overall framework of our method. It consists of a text detector, a contrast-guided feature adjustment module, a recognition module, and an information extraction module.

- The CNN-based text detector not only localizes all text instances but also yields visual feature maps F for subsequent modules. The text detector detects at the line level.
- With RoIAlign, the visual features of each instance I are obtained from F. Then, I is simultaneously fed into CFAM and a recognition module to generate a similarity matrix S and the recognition features T. In the recognition module which is two-layers transformer decoder, we use the intermediate features in transformer as recognition features T and obtain the recognition results of all text. In the CFAM, I is first encoded to E. Meanwhile, we use a serial of learnable vectors as the entity features L'. Subsequently, the pair-wise similarity between L' and E is calculated.
- Finally, the recognition features T, the encoded features E, and similarity matrix S are added, and the results are fed into the following information extraction module (LSTM-CRF) to predict the structural outputs.

4.1 Contrast-guided Feature Adjustment Module

Contrastive learning is first proposed to bridge the tasks of OCR and information extraction. The major difficulty is how to design proper representations on behalf of the two tasks. Therefore, we propose CFAM, the key component of our framework, as shown in Figure 4. Given visual features of each instance $I \in \mathbb{R}^{N \times C}$, the encoded features $E \in \mathbb{R}^{N \times C}$ are obtained via context modeling between I, where N and C indicate the number of instances and the channel of features, respectively. Then E is used to guide the generation of entity features $L' \in \mathbb{R}^{M \times C}$, where M indicates the number of entity categories. Next, we take E and L' as inputs and calculate a similarity matrix $S \in \mathbb{R}^{N \times M}$ between them. To supervise the generation of S, we use the ground truth $\tilde{S}$ of correspondences between instances and entity categories, which is obtained from the ground truth of the entity. Finally, the outputs of CFAM are S and E.

Generation of encoded features. The visual features of each instance I usually are discrete among instances. Thus, we use three Self-Attention (SA) layers to encode $I \in \mathbb{R}^{N \times C}$ and generate the encoded features $E \in \mathbb{R}^{N \times C}$, which can effectively establish connections among instances and transfer rich visual features

I from the text detection to the information extraction. The SA consists of three inputs, including queries (Q), keys (K), and values (V), defined as follows:

$$SA(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{C/m}}) \cdot V, \tag{1}$$

where Q, K, and V are obtained from the same input I (e.g., $Q=IW_Q$). Particularly, we use the multi-head self-attention (MSA) to construct the complex feature relations, $MSA = [SA_1; SA_2; ..; SA_m]W_O$, where W_O is a projection matrix and m is the number of attention heads, set as 8.

Generation of entity features. It is crucial to design the entity features $L' \in \mathbb{R}^{M \times C}$, which represents the information extraction task. The natural way is directly adopting the learnable entity embeddings $L \in \mathbb{R}^{M \times C}$ as entity features. However, the entity features are almost the same among all samples for the same initialization of learnable entity embeddings and cannot fully use the information from OCR. Additionally, simply using encoded features E as entity features causes lower generalization for information extraction. Thus, based on this consideration, we adopt E and L together to generate the entity features L'. Given the encoded features E, which are first transposed as E^T , then the E^T are fed into FC layer, which results in $E' \in \mathbb{R}^{C \times M}$:

$$E' = FC(E^T) \tag{2}$$

Next, we transform the E', then use E' to guide the generation of the entity features $L' \in \mathbb{R}^{M \times C}$ by adding the E' to learnable entity embeddings $L \in \mathbb{R}^{M \times C}$, defined as follows:

$$L' = L + E'^T \tag{3}$$

Generation of similarity matrix. The correspondences between encoded features E and entity features L' are significant for establishing the relations between the tasks of OCR and information extraction. Hence, we take E and L' jointly as inputs to construct the relations between them via calculating the product of the two matrices, which results in a similarity matrix $S \in \mathbb{R}^{N \times M}$ reflecting the relations between encoded features E and entity features L':

$$S = E \cdot L'^T \tag{4}$$

Moreover, we use the ground truth $\tilde{S}$ that is obtained from the annotated labels of the entity to supervise the generation of S.

4.2   Information Extraction Module

Information extraction requires both visual and textual understanding to automatically assign visual elements to various entities. Therefore, the recognition features T, encoded features E, and similarity matrix S, indicating the visual, textual and correlative information are used for information extraction. Following previous representative methods [30,40], we adopt LSTM-CRF[14] as an information extraction module to predict entity categories for all characters at the character level and output the final structure results.

Fig. 4. The structure of the proposed contrast-guided feature adjustment module (CFAM).

4.3   Loss Function

In the training phase, our proposed framework can be trained in an end-to-end manner with the weighted sum of the losses generated from four parts of text detection, recognition, information extraction, and contrastive learning, which are defined as follows:

$$L = \alpha L_{det} + \beta L_{rec} + \gamma L_{ie} + \lambda L_c, \tag{5}$$

where α, β, γ, and λ are hyper-parameters that control the tradeoff between losses.

L_{det} is the same loss as Mask R-CNN [4] for text detection. L_{rec} is the cross entropy loss for character recognition as follows:

$$L_{rec} = CE(R, \tilde{R}), \tag{6}$$

where CE indicates the cross entropy loss, $\tilde{R}$ denotes the ground truth of each character.

L_{ie} is also the cross entropy loss for entity classification as follows:

$$L_{ie} = CE(A, \tilde{A}), \tag{7}$$

where CE indicates the cross entropy loss, $\tilde{A}$ denotes the ground truth of entity category.

L_c indicates the loss for contrastive learning. Due to the imbalanced distribution of the number of entities, we adopt the focal loss [19] for L_c as follows:

$$L_c = FL(S, \tilde{S}), \tag{8}$$

where FL indicates the focal loss, $\tilde{S}$ denotes the actual entity category of each character.

Table 2. Performance comparison of the SOTA algorithms on the SROIE and POIE datasets. For text spotting, the results contain two parts (the left and right parts indicate the results of text detection and recognition, respectively). The Δ means the drop of results from SROIE to POIE dataset. * indicates our reproduced results. Note that all results are exhibited in F1-Score.

Setting	Method	SROIE	POIE	Δ
Text Spotting	Mask Textspotter [17]*	97.38/91.23	97.68/91.03	0.30/-0.20
Pure IE	VIES [30]	96.12	84.08*	-12.04
	PICK [38]	96.12	83.23*	-12.89
	TRIE [40]	96.18	82.45*	-13.73
	LayoutLMv2 [33]	96.25	84.18	-12.07
End-to-End IE	TRIE [40]	82.06	76.37*	-5.69
	VIES [30]*	83.44	77.19	-6.25

5   Experiments

5.1   Implementation Details

Our proposed method is implemented in Pytorch. We use four TITAN Xp with 12GB RAM to train our model with batch size four and the Adadelta optimizer. The learning rate starts from 2e-4 and decays to 2e-6 following the linear decay schedule. For the SROIE dataset, the total training epoch is set to 600 for a fair comparison. The total training epoch for the POIE is set to 200 for a fair comparison. We adopt Mask R-CNN [4] as our text detector with ResNet-50 [5] followed by FPN [18]. The hyper-parameters α and β are set to 1.0 while γ and λ are set to 10.0. We choose the best results of all methods in total epoch as the final results. Besides, we use bounding boxes and transcripts for all the text in the images.

5.2   Analysis of Our Proposed Dataset

In this part, to verify the practical utility of our POIE dataset, we make comprehensive comparisons between the POIE dataset and the SROIE [7] dataset. Although SROIE is the most widely used English dataset in the field of VIE, there exists a large number of errors in the annotations of SROIE. Additionally, the SROIE mainly focuses on document images with a single scene layout and few entities, which cannot fully reflect the challenges of information extraction in the wild. The comparisons are from three aspects: 1) We first explore the performance of popular text spotting method [17] on these two datasets. The aim of this setting is to verify the differences between the OCR task and VIE task. 2) Then, we evaluate the performance of a few typical methods for the pure IE task (i.e., directly using the ground truth of bounding boxes and texts as the inputs for information extraction). This setting can effectively reveal the difficulty of variable layouts and entities for information extraction. 3) Finally,

Table 3. Performance comparison of the state-of-the-art algorithms on SROIE and POIE datasets. * indicates our reproduced results. Backbone of all algorithms is ResNet50. Note that all methods are evaluated in F1-Score without post-processing.

Method	SROIE	POIE
GRAPHIE [23]	76.51	-
Chargrid [11]	78.24	-
GCN [20]	80.76	-
TRIE [40]	82.06	-
TRIE [40]*	82.82	76.37
VIES [30]*	83.44	77.19
Ours	**85.87**	**79.54**

to further evaluate the challenges of our dataset, we evaluate some methods in the end-to-end setting (i.e., using OCR results from text spotting as inputs for information extraction.) The details comparisons of these three settings are listed as follows:

Text spotting setting. As shown in the first part of Table 2, we observe that the performances of the pioneering text spotting algorithm Mask Textspotter [17] are slightly different on POIE and SROIE, which demonstrates that owing to the rapid development of text spotting, document and scene texts can be effectively detected and recognized by the advanced algorithms.

Pure IE setting. As listed in the second part of Table 2, we find that the performance of representative methods [40,30] have a significant drop from SROIE to our dataset, e.g., the performance of VIES and TRIE from 96.12 to 82.91 and from 96.18 to 81.79 respectively. The reason is that the variable layouts and entities are much more difficult for information extraction, demonstrating that our dataset is more practical for promoting advanced VIE algorithms.

End-to-end IE setting. As shown in the third part of Table 2, the performance of the TRIE [40] and VIE [30] has distinct differences from SROIE to ours, e.g., the performance of TRIE and VIES from 82.06 to 76.37 and from 83.44 to 77.19, respectively, which further proves our dataset is also more practical for end-to-end methods.

Qualitative analysis. In this part, we show the qualitative analysis for the difficulty of our POIE dataset. To extract the values of entities from images, the method requires adequate semantic understanding, which is very challenging for algorithms. A typical challenge of our POIE dataset can be seen in Figure 5. We find that without any clues for the "serving size" (an entity in our dataset), the algorithm cannot accurately predict the value of "serving size".

Ground-truth	Ours

SS: 100g
TF-P1: 1.1g
…
CE-P1: 106
TF-D: 1.6%

SS: 100g
TF-P1: 1.1g
…
CE-P1: 106
TF-D: 1.6%

Fig. 5. Qualitative analysis of the proposed POIE dataset. Red boxes indicate the predicted detection results. Texts show the results of information extraction. Note that green texts indicate the false negative results.

Table 4. Ablation study of CFAM on the proposed POIE dataset. SCH indicates a simple classification head.

SCH	CFAM	POIE dataset		
		F1-Score (DET)	F1-Score (REC)	F1-Score (IE)
-	-	97.50	89.40	75.05
✓	-	97.10	89.01	77.59
-	✓	**97.89**	**91.68**	**79.54**

5.3   The Comparisons of Our Method and SOTA

In this part, we compare our method with previous state-of-the-art (SOTA) methods [23,11,20,40,30] on POIE and SROIE datasets. Our method is in an end-to-end manner, and the core of our approach is to narrow the semantic gap between the tasks of OCR and information extraction. Consequently, we mainly compare our proposed method with the end-to-end methods and focus on the performance of IE (i.e., F1-score of information extraction).

As shown in Table 3, our method achieves SOTA results on POIE and SROIE datasets, and outperforms other methods by an obvious margin. For the POIE dataset, our method exhibits superior performance. For the SROIE dataset, most previous methods do not use post-progress, thus we mainly focus on the results produced without post-progress. Additionally, our method outperforms other methods by obvious gains. Moreover, Figure 6 shows some qualitative results of ours and other methods. We can observe that our method is more precisely to predict the values of entities.

Ground truth	TRIE	VIES	Ours
SS: 1 bottle (8 fl oz) **TF-PS**: 6g … **SO-PS**: 200mg **PRO-D**: 18%	**SS**: 1 bottle (8 fl oz) **TF-PS**: 6g … **SO-PS**: 200mg **PRO-D**: 180	**SS**: 1 bottle (8 fl oz) **TF-PS**: 6g … **SO-PS**: 200mg **PRO-D**: 18%	**SS**: 1 bottle (8 fl oz) **TF-PS**: 6g … **SO-PS**: 200mg **PRO-D**: 18%
TF-P1: 29.7 g **TF-PS**: 8.9g … **CAR-P1**: 37.8 g **PRO-P1**: 7 g	**TF-P1**: 29.7 g **TF-PS**: 8.9 g … **CAR-P1**: 37.8 g **PRO-P1**: 7 g	**TF-P1**: 29.7 g **TF-PS**: 8.9 g … **CAR-P1**: 37.8 g **PRO-P1**: 7 g	**TF-P1**: 29.7 g **TF-PS**: 8.9 g … **CAR-P1**: 37.8 g **PRO-P1**: 7 g
SS: 100g **TF-PS**: 1.1g … **CE-PS**: 106 **TF-D**: 1.6%	**SS**: 100g **TF-PS**: 1.1g … **CE-PS**: 106 **TF-D**: 1.6%	**SS**: 100g **TF-PS**: 1.1g … **CE-PS**: Sugar **TF-D**: 1.6%	**SS**: 100g **TF-PS**: 1.1g … **CE-PS**: 106 **TF-D**: 1.6%
SS: (25.5 g) **CAR-P1**: 74 g … **PRO-PS**: 0.6 g **TF-P1**: 15 g	**SS**: (25.5 g) **CAR-P1**: 74 g … **PRO-PS**: 0.6 g **TF-P1**: 15 g	**SS**: (25.5 g) **CAR-P1**: 74 g … **PRO-PS**: 0.6 g **TF-P1**: 15 g	**SS**: (25.5 g) **CAR-P1**: 74 g … **PRO-PS**: 0.6 g **TF-P1**: 15 g

Fig. 6. Some visualization results of our method and others on the proposed POIE dataset. There are results of TRIE, VIES, and ours, from left to right, respectively. Red boxes indicate the predicted detection results. Texts show the information extraction results. Note that green and blue indicate the false negative and false positive results, respectively.

5.4   Ablation Studies

The ablation studies are carried out on our POIE dataset. For a complete comparison, we show the experimental results of DET, REC, and IE to evaluate the effectiveness of our proposed modules. Specifically, as described below:

Effectiveness of CFAM. To verify the effectiveness of CFAM, we conduct experiments in the settings: classification and CFAM. Classification indicates directly classifying visual features to the corresponding entity. As listed in Table 4, we observe that classification can promote the performance of information extraction but have a negative impact on text detection and recognition tasks. We argue that it is because our method is an end-to-end method, where the OCR and information extraction tasks are joint optimizations. Moreover, compared with classification, we find that with CFAM, our method achieves superior results (97.89% in DET, 91.68% in REC, 79.54% in IE) for the reason that CFAM serves as an auxiliary module to bridge the relations between the tasks of OCR and information extraction. Additionally, to verify the necessity of using supervision for contrastive learning in CFAM, we conduct experiments in the setting: with L_c and without L_c. As shown in Table 5, we found that the contrastive loss L_c can promote the performance.

Analysis of generated entity features. In our proposed CFAM module, it is important to design the entity features representing the information extraction task. There are three strategies for generating the entity features (i.e., directly using learnable entity embeddings, adopting encoded features, and combining the

Table 5. The effectiveness of L_c in CFAM.

L_c	POIE dataset		
	F1-Score (DET)	F1-Score (REC)	F1-Score (IE)
-	97.48	91.54	78.69
✓	**97.89**	**91.68**	**79.54**

Table 6. Analysis of generated entity features. Note that all results are shown in F1-Score.

Learnable entity embeddings	Encoded features	POIE dataset		
		DET	REC	IE
✓	-	97.48	90.51	78.86
-	✓	97.81	91.48	76.25
✓	✓	**97.89**	**91.68**	**79.54**

learnable entity embeddings and encoded features). The performance of these strategies is shown in Table 6. We observe that using learnable entity embeddings and encoded features together achieves the best results. However, either using learnable entity embeddings or encoded features will get lower results. We argue that the reasons are two-fold: 1) without encoded features, the generated entity features are almost the same among all entity features and cannot fully use the information from instance. 2) without learnable entity embeddings, the generated entity features completely rely on the encoded feature and cannot fully represent the feature of information extraction.

Generalization of CFAM. To further demonstrate the generalization of our proposed CFAM, we evaluate CFAM by applying it to other SOTA methods [40,30]. As listed in Table 7, the proposed CFAM improves the performance of TRIE and VIES by a distinguish margin even though they are strong methods. For our POIE dataset, CFAM improves the SOTA method TRIE and VIES by 2.45% and 1.92%, respectively. For the SROIE dataset, CFAM enhances the performance of TRIE and VIES by an obvious margin. The results demonstrate the generalization of our proposed CFAM, which can be effectively applied to other methods.

5.5   Limitation

Our method has achieved superior results on SROIE and ours. However, we find that the challenges, e.g., different forms for the same entity (Figure 2 (b)), have not been solved well and still influence the performance of information extraction. In the future, we will further explore the commonality between different forms of the same entity and design a new method to effectively solve this problem.

Table 7. The effectiveness of CFAM in other methods on the SROIE and POIE datasets. We reproduce the TRIE and VIES. Note that all results are exhibited in F1-Score.

Method	SROIE	POIE
TRIE [40]	82.82	76.37
TRIE [40] + CFAM (**ours**)	**84.25**	**79.17**
VIES [30]	83.44	77.19
VIES [30] + CFAM (**ours**)	**84.32**	**79.11**

6 Conclusion

In this paper, we have proposed a large-scale English dataset (called POIE) consisting of camera images, which can reflect the challenges of the real world. We have designed a novel end-to-end framework with a plug-and-play CFAM for VIE tasks, which adopts contrastive learning and properly designs the representation of VIE tasks for contrastive learning. The experimental results prove that our dataset is more practical for promoting advanced VIE algorithms. Additionally, the experiments demonstrate that our proposed framework consistently achieves the obvious performance gains on SROIE and ours. In the future, we will consider to extend the dataset with more images, entities, and diverse natural scene disturbances. We hope our proposed dataset and framework can promote further investigation in VIE.

References

1. Aberdam, A., Litman, R., Tsiper, S., Anschel, O., Slossberg, R., Mazor, S., Manmatha, R., Perona, P.: Sequence-to-sequence contrastive learning for text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15302–15312 (2021) 4, 5
2. Cao, H., Li, X., Ma, J., Jiang, D., Guo, A., Hu, Y., Liu, H., Liu, Y., Ren, B.: Query-driven generative network for document information extraction in the wild. In: ACM MM. pp. 4261–4271 (2022) 2, 4
3. Guo, H., Qin, X., Liu, J., Han, J., Liu, J., Ding, E.: Eaten: Entity-aware attention for single shot visual text extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 254–259. IEEE (2019) 5
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017) 9, 10
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 10
6. Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. Advances in Neural Information Processing Systems **34**, 3635–3649 (2021) 4

7. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019) 2, 3, 4, 6, 10

8. Hwang, W., Kim, S., Seo, M., Yim, J., Park, S., Park, S., Lee, J., Lee, B., Lee, H.: Post-ocr parsing: building simple and robust parser via bio tagging (2019) 4

9. Jiang, X., Long, R., Xue, N., Yang, Z., Yao, C., Xia, G.S.: Revisiting document image dewarping by grid regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4543–4552 (2022) 2

10. Kahraman, H.T., Sagiroglu, S., Colak, I.: Development of adaptive and intelligent web-based educational systems. In: 2010 4th international conference on application of information and communication technologies. pp. 1–5. IEEE (2010) 2

11. Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., Faddoul, J.B.: Chargrid: Towards understanding 2d documents. In: EMNLP (2018) 5, 11, 12

12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning **33**, 18661–18673 (2020) 4, 5

13. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: ECCV. pp. 498–517. Springer (2022) 5

14. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT. pp. 260–270 (2016) 8

15. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. 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). pp. 2592–2607 (2021) 5

16. Li, X.H., Yin, F., Dai, H.S., Liu, C.L.: Table structure recognition and form parsing by end-to-end object detection and relation parsing. Pattern Recognition **132**, 108946 (2022) 4

17. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: ECCV. pp. 706–722. Springer (2020) 10, 11

18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 10

19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017) 9

20. Liu, X., Gao, F., Zhang, Q., Zhao, H.: Graph convolution for multimodal information extraction from visually rich documents. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). pp. 32–39 (2019) 11, 12

21. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: Cord: a consolidated receipt dataset for post-ocr parsing (2019) 2, 4

22. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021) 5

23. Qian, Y., Santus, E., Jin, Z., Guo, J., Barzilay, R.: Graphie: A graph-based frame-work for information extraction. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 751–761 (2019) 4, 11, 12

24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 4, 5

25. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Dense-clip: Language-guided dense prediction with context-aware prompting. In: Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022) 5

26. Sun, H., Kuang, Z., Yue, X., Lin, C., Zhang, W.: Spatial dual-modality graph reasoning for key information extraction. arXiv preprint arXiv:2103.14470 (2021) 4

27. Tang, G., Xie, L., Jin, L., Wang, J., Chen, J., Xu, Z., Wang, Q., Wu, Y., Li, H.: Matchvie: Exploiting match relevancy between entities for visual information extraction 4, 5

28. Wang, H., Bai, X., Yang, M., Zhu, S., Wang, J., Liu, W.: Scene text retrieval via joint text detection and similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4558–4567 (2021) 5

29. Wang, H., Liao, J., Cheng, T., Gao, Z., Liu, H., Ren, B., Bai, X., Liu, W.: Knowl-edge mining with scene text for fine-grained recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4624–4633 (2022) 2

30. Wang, J., Liu, C., Jin, L., Tang, G., Zhang, J., Zhang, S., Wang, Q., Wu, Y., Cai, M.: Towards robust visual information extraction in real world: new dataset and novel solution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2738–2745 (2021) 2, 3, 4, 5, 6, 8, 10, 11, 12, 14, 15

31. Wang, W., Huang, Z., Luo, B., Chen, Q., Peng, Q., Pan, Y., Yin, W., Feng, S., Sun, Y., Yu, D., et al.: mmlayout: Multi-grained multimodal transformer for doc-ument understanding. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4877–4886 (2022) 2

32. Wong, K.Y., Casey, R.G., Wahl, F.M.: Document analysis system. IBM journal of research and development **26**(6), 647–656 (1982) 2

33. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., 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). pp. 2579–2591 (2021) 10

34. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: ACMSIGKDD. pp. 1192–1200 (2020) 4

35. Yang, M., Liao, M., Lu, P., Wang, J., Zhu, S., Luo, H., Tian, Q., Bai, X.: Reading and writing: Discriminative and generative modeling for self-supervised text recog-nition. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4214–4223 (2022) 5

36. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y.: Graph contrastive learn-ing with augmentations **33**, 5812–5823 (2020) 5

37. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. arXiv preprint arXiv:2302.14338 (2023) 4
38. Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In: ICPR. pp. 4363–4370. IEEE (2021) 4, 10
39. Zeng, G., Zhang, Y., Zhou, Y., Yang, X., Jiang, N., Zhao, G., Wang, W., Yin, X.C.: Beyond ocr+ vqa: Towards end-to-end reading and reasoning for robust and accurate textvqa. Pattern Recognition p. 109337 (2023) 2
40. Zhang, P., Xu, Y., Cheng, Z., Pu, S., Lu, J., Qiao, L., Niu, Y., Wu, F.: Trie: end-to-end text reading and information extraction for document understanding. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1413–1422 (2020) 3, 5, 8, 10, 11, 12, 14, 15
41. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022) 5
42. Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., Chua, T.S.: Towards complex document understanding by discrete reasoning. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4857–4866 (2022) 2