IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 March 2024, accepted 13 April 2024, date of publication 17 April 2024, date of current version 24 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390153

RESEARCH ARTICLE

Zero-Shot Pill-Prescription Matching With Graph Convolutional Network and Contrastive Learning

TRUNG THANH NGUYEN[1,3], **(Graduate Student Member, IEEE),**
PHI LE NGUYEN[2], **(Member, IEEE), YASUTOMO KAWANISHI**[1,3], **(Member, IEEE),**
TAKAHIRO KOMAMIZU[1,4], **(Member, IEEE), AND ICHIRO IDE**[1,4], **(Senior Member, IEEE)**

[1]Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan
[2]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 100000, Vietnam
[3]Guardian Robot Project, Information Research and Development and Strategy Headquarters, RIKEN, Seika, Kyoto 619-0288, Japan
[4]Mathematical and Data Science Center, Nagoya University, Nagoya, Aichi 464-8601, Japan

Corresponding author: Trung Thanh Nguyen (nguyent@cs.is.i.nagoya-u.ac.jp)

This work was supported by Vingroup Joint Stock Company (Vingroup JSC), Vingroup; in part by the Vingroup Innovation Foundation (VINIF) under Project VINIF.2021.DA00128; and in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP21H0355.

ABSTRACT Patients' safety is paramount in the healthcare industry, and reducing medication errors is essential for improvement. A promising solution to this problem involves the development of automated systems capable of assisting patients in verifying their pill intake mistakes. This paper investigates a Pill-Prescription matching task that seeks to associate pills in a multi-pill photo with their corresponding names in the prescription. We specifically aim to overcome the limitations of existing pill detection methods when faced with unseen pills, a situation characteristic of zero-shot learning. We propose a novel method named Zero-PIMA (Zero-shot Pill-Prescription Matching), designed to match pill images with prescription names effectively, even for pills not included in the training dataset. Zero-PIMA is an end-to-end model that includes an object localization module to determine and extract features of pill images and a graph convolutional network to capture the spatial relationship of the pills' text in the prescription. After that, we leverage the contrastive learning paradigm to increase the distance between mismatched pill images and pill name pairs while minimizing the distance between matched pairs. In addition, to deal with the zero-shot pill detection problem, we leverage pills' metadata retrieved from the DrugBank database to fine-tune a pre-trained text encoder, thereby incorporating visual information about pills (e.g., shape, color) into their names, making them more informative and ultimately enhancing the pill image-name matching accuracy. Extensive experiments are conducted on our collected real-world VAIPE$_{\mathrm{PP}}$ dataset of multi-pill photos and prescriptions. Through a series of comprehensive experiments, the proposed method outperforms other methods for both seen and unseen pills in terms of mean average precision. These results indicate that the proposed method could reduce medication errors and improve patients' safety.

INDEX TERMS Contrastive learning, graph convolutional network, object detection, pill-prescription matching, text-image matching, zero-shot learning.

I. INTRODUCTION

Medication is crucial in treating various diseases and improving patients' health. However, medication mistakes can lead to severe consequences, such as reducing the effectiveness of treatment, causing adverse effects, and even leading to death [1], [2], [3], [4]. According to a report by the United States National Coordinating Council for Medication Error Reporting and Prevention [5], drug abuse accounts for one-third of all deaths rather than the illness

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Lin.

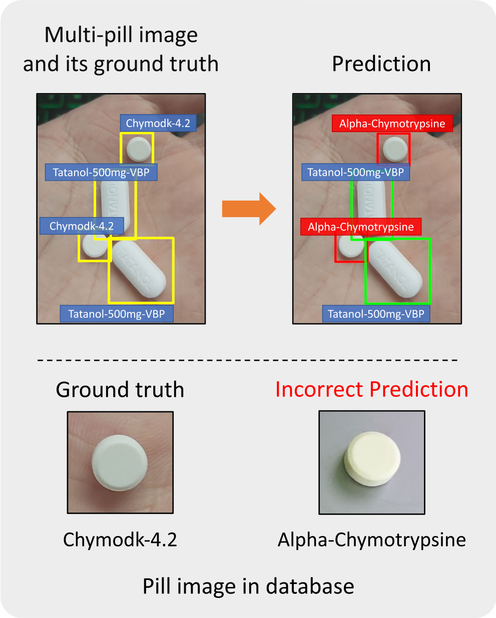


FIGURE 1. Illustration of incorrect pill detection using object detection models. In the task of pill detection, many pills are similar.

itself. Moreover, in the United States alone, approximately 7, 000 to 9, 000 people die yearly from medication mistakes. Several factors can lead to drug abuse, including taking the wrong amount or the wrong kind of medication and taking pills that are not prescribed. There are numerous causes of drug abuse, which may originate with the physician (prescribing error), the pharmacist (preparation error, dispensing error), or the patients themselves (wrong medication than prescribed) [6], [7]. In literature, considerable efforts have been dedicated to reducing the drug abuse caused by the first two groups [8], [9], [10]. However, only a small number of research has focused on drug abuse by patients themselves. In underdeveloped nations where regulations and processes concerning drug usage are not stringent and well-defined, medication errors caused by patients themselves occur frequently and become a critical issue, particularly among elderly patients and children [11], [12]. To this end, this study is one of the earliest attempts to reduce patient-caused drug abuse. In particular, we offer an approach for automatically matching information between pill names given in the prescription and the pill images presented in a photo of multi-pill intake, thereby aiding patients in detecting cases of taking unprescribed pills or mistaking the prescribed pills. We name our task as **Pill-Prescription matching**, which can be defined as follows: *Given a snapshot of a prescription and a photo of all the pills in a pill intake, match the names of the pills on the prescription and their corresponding regions in the multi-pill photo.*

A. EXISTING APPROACHES AND CHALLENGES

For the task of pill detection, most methods utilize object detection models [13], [14], [15], [16]. As illustrated in Figure 1, this approach encounters several challenges in this particular task due to numerous pills that look similar in shape and color. To address this issue, some methods have attempted to recognize characters printed on the surface of the pills [17]. However, not all pills have such characters, reducing the effectiveness of this approach.

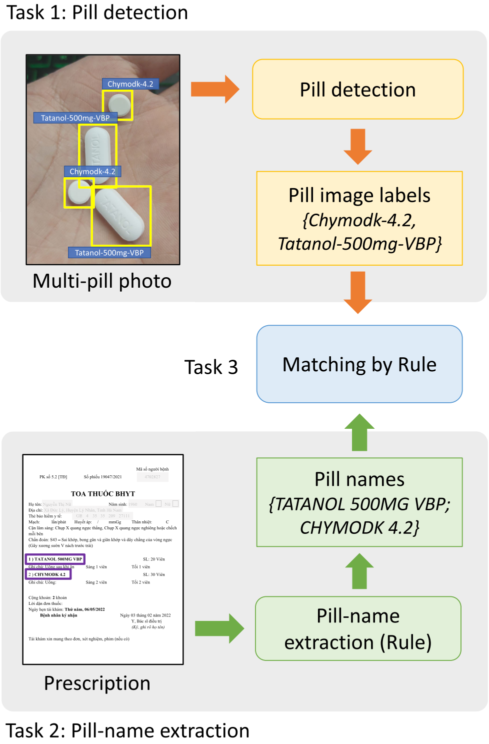


FIGURE 2. In conventional methods, the first step entails detecting pill labels (Task 1) and identifying pill names (Task 2) from the multi-pill photo and prescription text, respectively. Subsequently, the focus shifts to matching the pill names with their corresponding image labels (Task 3).

For the Pill-Prescription matching task, the conventional approach breaks into three sub-tasks: detecting the pill images presented in the provided multi-pill photo, extracting the pill names described in the prescription, and pill image-name matching. As shown in Figure 2, pill detection is typically achieved using object detection techniques. On the other hand, the extraction of pill names often relies on optical character recognition techniques for text localization and recognition and rule-based methods for pill-name information extraction. Once the pill image labels and pill names are extracted, they can be matched using rule-based matching techniques. Despite efforts in pill detection, this approach suffers from severely inaccurate classification, as stated previously. Moreover, the problem of pill name identification is challenging since the name of a pill might be expressed in numerous ways (e.g., a common pain reliever might be known as ‘‘Tylenol’’ in one country, ‘‘Paracetamol’’ in another, and ‘‘Acetaminophen’’ in yet others). When solving the Pill-Prescription matching by decomposing it into three sub-tasks, the inaccuracies of these single tasks will

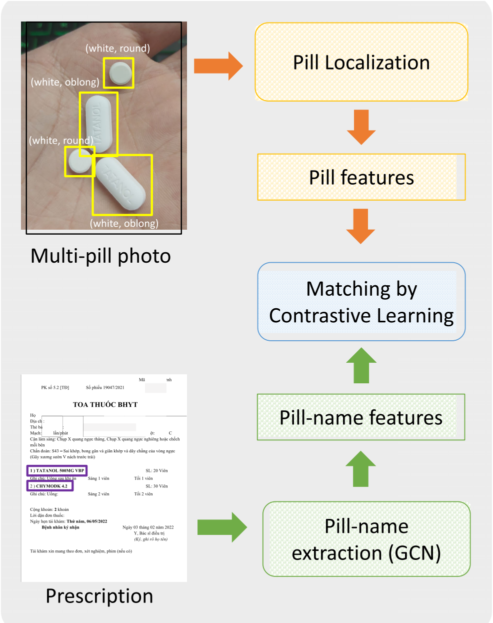


FIGURE 3. **Proposed method (1-stage approach) leverages the pill features obtained through object localization and employs a Graph Convolutional Network (GCN) to extract pill name features from prescriptions. The matching process is achieved through contrastive learning.**

accumulate, leading to a degradation in the pill identification accuracy.

Moreover, the conventional Pill Detection and Pill-Prescription matching approach fails to handle new pills that do not appear in the training dataset. It should be noted that new pills are frequently introduced (According to [18], an average of 43 new pharmaceuticals have been approved over a rolling 10-year period); therefore, the limitations of these methods make it challenging to put into practice. This necessitates a robust solution that can effectively identify seen pills (i.e., pills that appeared in the training dataset) and unseen pills (i.e., pills that have not appeared in the training dataset).

B. PROPOSED SOLUTION

In this work, we focus on *zero-shot learning for the Pill-Prescription matching problem* and propose ***Zero-PIMA*** *(**Zero**-shot **PI**ll-Prescription **MA**tching)*,[1] a novel approach that can accurately match pill images in a multi-pill intake photo and their corresponding names in the prescription even if the pills have not appeared in the training dataset. Our main idea lies in two points as follows:

1) To increase the overall accuracy of the Pill-Prescription matching problem, we propose an end-to-end deep learning model (Figure 3) that integrates the pill image localization, pill name extraction, and matching phases altogether. In this way, we are able to avoid the error accumulation issue associated with the conventional approach. To be more specific, we first employ an object localization module to determine and extract features of pill images. Meanwhile, a Graph Convolutional Network (GCN) is leveraged to capture the spatial relationship of text boxes in the prescription and highlight those containing pill names. After that, we leverage the contrastive learning paradigm to increase the distance between mismatched pill image and pill name pairs while minimizing the distance between matched pairs.
2) To identify unseen pills, we utilize the pills' metadata (i.e., shape, color), which is obtained from the DrugBank database [20]. The DrugBank database is a central repository storing data about almost all pills. Thus, leveraging this source allows us to retrieve information about unseen pills. This metadata is combined with the pill names to train a text embedding model. These text features are then aligned with the visual features of the pills. When an unseen pill appears, although its exact name may be unknown, its visual attributes like shape and color enable us to cross-reference and accurately identify it to the corresponding pill name in the prescription.

C. CONTRIBUTIONS

The main contributions of this paper are three-fold as follows:

- **Problem Definition and Solution Zero-PIMA:** We highlight the importance of the patient-caused drug abuse issue and define the Pill-Prescription matching problem. We then propose a novel end-to-end approach for handling the zero-shot learning for the Pill-Prescription matching problem. The proposed method leverages the GCN and Contrastive Learning to match pill images in a multi-pill intake photo and pill names in a prescription for both seen and unseen pills accurately.
- **Dataset Construction:** We provide a real-world dataset consisting of 2,156 multi-pill photos corresponding to 1,527 prescriptions. To the best of our knowledge, this dataset is the first one capturing the prescriptions and corresponding pill images.
- **Accurate Detection:** We perform extensive experiments to evaluate the performance of the proposed method and compare it with benchmark models. Experimental results demonstrate that the proposed method improves the accuracy in both seen and unseen pills on mean Average Precision (mAP) compared to the other methods.

The remainder of this paper is organized as follows: Firstly, we briefly summarize relevant works in Section II. We then present the details of the proposed method in Section III and evaluate its performance in Section IV. Section V concludes the paper and introduces our future direction.

[1] In our previous work [19], we proposed a pill-prescription matching approach, called PIMA. However, that solution did not address the pill localization problem (i.e., The pill images had to be cropped before being fed into the model). It also could not handle unseen pill images.

II. RELATED WORK

In this section, we introduce traditional Object Detection methods and Zero-shot Object Detection methods in Sections II-A and II-B, respectively, emphasizing this critical computer vision task's enhancement through deep learning. Following this, we explore related work on Pill Recognition and Pill Detection in Sections II-C and II-D, respectively.

A. TRADITIONAL OBJECT DETECTION

Object detection is a well-established research task in computer vision, and there have been many successful methods proposed in literature. One of the earliest and most widely used approaches is the Viola-Jones method [21], which uses Haar-like features and a cascade of classifiers to detect objects. More recently, the development of deep learning has led to significant progress in object detection, with methods such as Region-based Convolutional Neural Network (R-CNN) [22], Fast R-CNN [23], and Faster R-CNN [24] achieving excellent performance on popular benchmark datasets such as Microsoft Common Objects in COntext (MS-COCO) [25] and PASCAL Visual Object Classes (PASCAL VOC) [26]. Other notable approaches include You Only Look Once (YOLO) [27], Single Shot multibox Detector (SSD) [28], and RetinaNet [29], which are designed for real-time object detection and have achieved competitive results. Additionally, there have been efforts to improve object detection by incorporating attention mechanisms, such as in the recent work on DEtection TRansformer (DETR) [30], which uses a transformer-based architecture to directly output object detections without the need for anchor boxes.

B. ZERO-SHOT OBJECT DETECTION

Zero-Shot object Detection (ZSD) has emerged as a cutting-edge trend in modern object detection, aiming to detect objects beyond predefined categories. This field poses unique challenges, particularly in aligning visual features with semantic representations of objects. Bansal et al. [31] laid the groundwork for ZSD by adapting visual-semantic embeddings, highlighting the necessity of effectively distinguishing between background and unseen classes through models aware of the background context. Rahman et al. [32] further advanced ZSD with an enhanced visual-semantic alignment technique, employing a polarity loss function to improve discrimination between positive and negative predictions significantly. More recently, the advent of vision-language pre-training has led to ZSD being conceptualized as an image-text matching problem [33], [34], [35], leveraging large-scale image-text data to expand the number of training classes. Inspired by these methods, this research utilizes a pre-trained vision-language model for unseen pill detection. However, since the pre-trained model primarily focuses on standard image and text pairs, we fine-tune it on the proposed pill dataset.

C. PILL RECOGNITION

Since accurate identification of pills is important for patients' safety and healthcare delivery, advance in computer vision techniques and deep learning has led to an increasing interest in developing automated systems for pill identification. Wong et al. [36] proposed a deep learning model using a deep convolutional network [37] for automatic pill identification and verification that outperformed existing methods, using pill images captured with mobile phones under unconstrained environments. While this approach offers potential accuracy, its computational intensity and the need for extensive training data are notable drawbacks. Ling et al. [17] proposed a pill image recognition approach using a light-weight W^2-net for segmentation and a multi-stream deep network. Their two-stage training methodology with Batch All and Batch Hard strategies aimed to handle the hard samples taken under less controlled imaging conditions. However, this method uses pill images taken in laboratory settings, and the pills need to have imprinted pill codes marked on their surfaces. Besides that, we introduced a new approach named PIKA (which stands for Pill Identification with medical Knowledge grAph) [38] to enhance pill recognition accuracy under practical conditions. This approach leveraged external knowledge, specifically prescriptions, to model the implicit association between pills. By employing a walk-based graph embedding model, it extracted relational features from pills and merged them with image-based visual features to achieve the final classification.

D. PILL DETECTION

The current state-of-the-art on the pill detection task is still immature, with previous studies relying on traditional object detection models. Kwon et al. [13] proposed a regional deep learning algorithm to improve pill detection performance with limited training data. The method detects the location and type of individual pills in an image with multiple pills by limiting the training data to single-pill images. A two-step detection method based on Mask R-CNN [39] is used to improve local detection performance, where the first step detects only the number and area of pills in the image, and the second step detects the type of the corresponding pill. However, the reliance on single-pill images during training in this method limits the model's capability to generalize to situations where multiple pills are present in an image. In our previous work [19], we were the first to investigate the Pill-Prescription matching problem, called PIMA (PIll-Prescription MAtching). We exploited the contrastive learning paradigm to contrast pill images and names, i.e., minimizing the distance of a pill image and its corresponding name while maximizing those of mismatched pill image-name, thereby, enhance the matching accuracy. However, since this approach did not address the pill localization problem, the pill images had to be cropped before being fed into the model. In addition, it could not handle unseen pill images.

III. PROPOSED METHOD: ZERO-PIMA

We propose **Zero-PIMA** (**Zero**-shot **PI**ll-Prescription **MA**tching), a novel approach in dealing with the Zero-shot Pill-Prescription matching problem. We first provide a broad outline of Zero-PIMA in Section III-A. We then delve deeper into Zero-PIMA's modules in Sections III-B and III-C. Finally, we elaborate on the proposed learning objectives in Section III-D, which plays a crucial role in enhancing the overall performance of the proposed method.

A. OVERVIEW

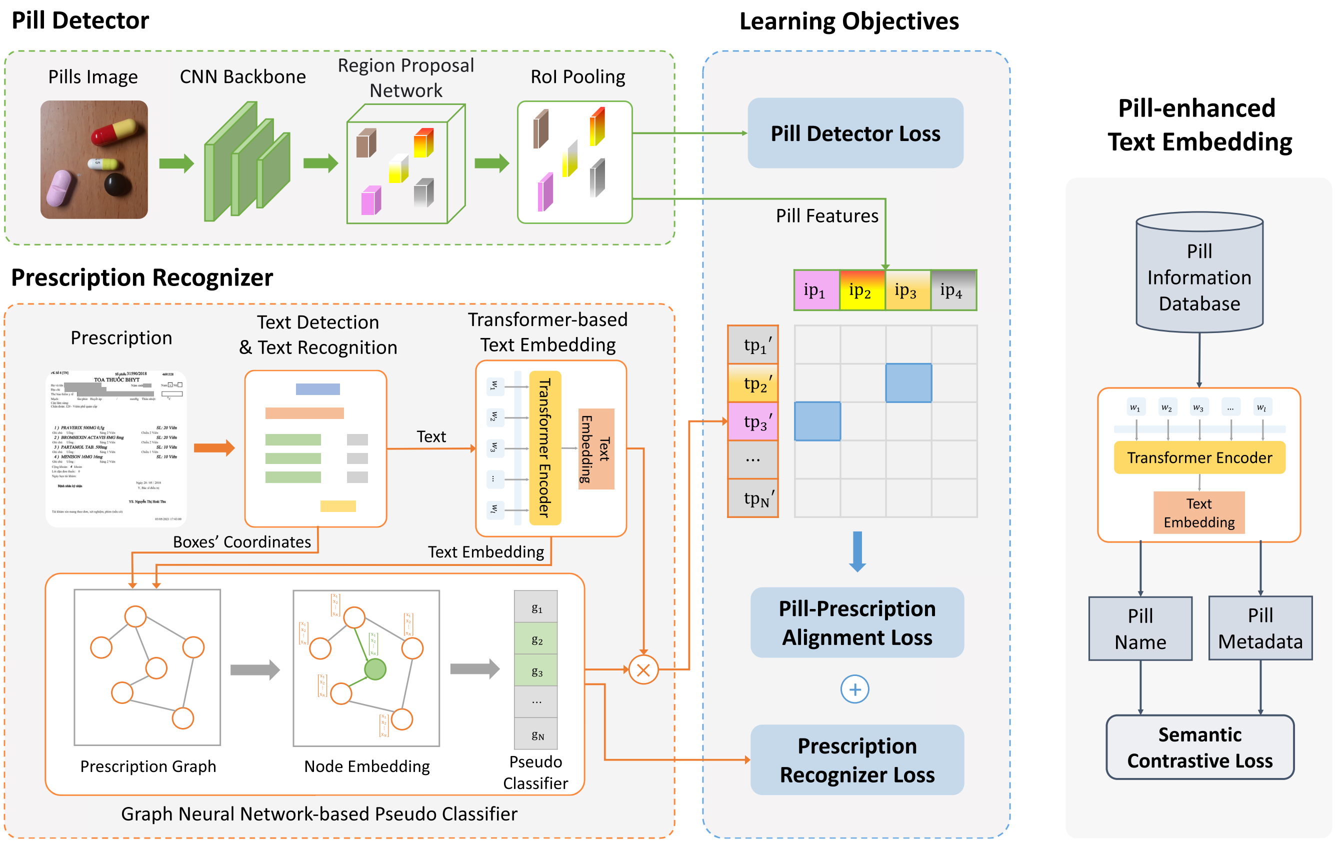
In this paper, we address the challenge of matching a set of pill images $P = \{p_1, p_2, \ldots, p_M\}$ (M is the number of pill objects) with their corresponding prescription names $S = \{s_1, s_2, \ldots, s_N\}$ (N is the number of texts) extracted from textual prescriptions. The proposed method formulates the matching task as an optimization problem, aiming to find a mapping function $f : P \rightarrow S$ that maximizes the accuracy and relevance of the pill-to-prescription name associations. We introduce constraints to ensure uniqueness and completeness in the matching process, where each pill is associated with no more than one prescription name. The proposed method leverages extracted features from both modalities to calculate similarity scores, facilitating the optimal pairing of pills with their respective prescription names.

We observe that the most challenging issues in handling the Pill-Prescription matching problem lie in the cases where (1) the pills to be identified have similar external appearances to others and (2) the pills were unseen in the training data. To tackle the first issue, we argue that it is insufficient to identify pills using only their visual information. Therefore, instead of decomposing the Pill-Prescription matching problem into three sub-tasks (i.e., pill detection, pill name extraction, and pill image-name matching), we combine all three into an end-to-end model and leverage information from extracted pill names to improve the pill image recognition accuracy, and vice versa. To address the second issue which entails coping with unseen pills, we utilize the metadata information (i.e., shape, color) retrieved from the DrugBank database [20]. This information is employed with the pill names to train the text embedding model. Thus, we can associate the name of each pill with its metadata, enhancing the deterministic nature of the pill names and, therefore, more effectively addressing the unseen pill issue.

Figure 4(a) illustrates the overview of the proposed model comprising three modules: *Pill Detector*, *Prescription Recognizer*, and *Learning Objectives*. The *Pill Detector* leverages the object detection technique to identify pills in the input image. Specifically, this module receives a multi-pill photo and produces the bounding boxes (enclosing pill objects) associated with their identities. We leverage a Convolutional Neural Network (CNN) as a backbone to extract features and a Region Proposal Network to suggest the objects' locations. The *Prescription Recognizer* is responsible for extracting textual information from a prescription. We employ a Transformer encoder to generate embeddings of the texts. Furthermore, we use a Graph Convolutional Network (GCN) to capture spatial relationships among the text boxes and highlight those representing pill names. To enrich the information of the pill names' text embeddings and better handle the unseen pill cases, we leverage pills' metadata (i.e., shape, color) retrieved from the DrugBank database to finetune the pre-trained text embeddings. The visual representations of the pill images (extracted by the Pill Detector) and textual features of pill names (generated by the Prescription Recognizer) are then projected onto a shared space and used as the inputs of the Pill-Prescription alignment in the *Learning Objectives* module. The Pill-Prescription alignment consists of a contrastive loss function aiming to establish associations between the pill name features and their corresponding pill image representation. The intuition behind the contrastive loss is to minimize the distance between features representing a pill image and its corresponding name while maximizing the distance between those depicting non-corresponding pill images and names. Furthermore, to enhance the deterministic of the features generated by the Pill Detector and Prescription Recognizer, we employ two losses in our learning objectives: Pill Detection loss and Prescription Recognizer loss. The former is responsible for detecting and localizing pill objects within the multi-pill photo, while the latter is to determine the text boxes containing pill names in the prescription.

B. PILL DETECTOR

The Pill Detector module is responsible for localizing pills and generating the representation for each pill in the multi-pill photo. For this purpose, in this work, we leverage Faster R-CNN [24] as the backbone model. However, it is worth noting that any other object detection technique can also be used as the Pill Detector. Figure 5 depicts the architecture of the Faster R-CNN with the main components of a CNN backbone for extracting a feature map of the input multi-pill photo and a Regions Proposal Network (RPN) for determining potential Region of Interest (RoI). The outputs of these two components are then used as inputs for the RoI pooling layer. The RoI pooling layer generates a representation for each RoI proposed by the RPN based on the feature map received from the CNN. To achieve at the same time two goals: (1) accurately determining the locations (i.e., bounding boxes' coordinates) of the objects, and (2) filtering out only those containing pill objects, we employ two loss functions: classification and bounding-box regression. Note that instead of using multi-class classification loss as in other object detection tasks, we use a binary classification loss to distinguish between the pill and non-pill objects. In this way, during the inference stage, we can filter out only bounding boxes that likely contain pills to perform the Pill-Prescription matching task. In addition, only pill-containing bounding



(a) Architecture of Zero-PIMA.

(b) Text embedding's finetuning flow.

FIGURE 4. **Overview of Zero-PIMA. (a) Illustration of the Zero-PIMA architecture consists of three modules: Pill Detector, Prescription Recognizer, and Learning Objectives. Pill Detector is responsible for localizing and extracting visual information from a multi-pill photo. Pill Prescription Recognizer utilizes a Graph Convolutional Network to highlight the text boxes likely to be pill names and a pill-enhanced text embedding to learn representations of pill names. Finally, textual and visual data are fed into the Pill-Prescription alignment in the Learning Objective module to produce a text-image retrieval result. (b) Semantic contrastive loss is applied to integrate pills' metadata into the pill names' embeddings.**

boxes are sent to the Pill-Prescription alignment during the training phase to enhance the training speed and matching precision.

Assuming that there are M pill objects in the input multi-pill photo denoted by $\{p_1, \ldots, p_M\}$, the Pill Detector will produce M feature vectors $\{\mathbf{i}^e_1, \ldots, \mathbf{i}^e_M\}$, where $\mathbf{i}^e_i$ represents the visual feature of pill p_i $(i = 1, \ldots, M)$. These feature vectors are then projected onto the same hyperplane with their counterparts in the prescription via a projection layer, resulting in the final representation of the pills as $I^p = \{\mathbf{i}^p_1, \ldots, \mathbf{i}^p_M\}$.

C. PRESCRIPTION RECOGNIZER

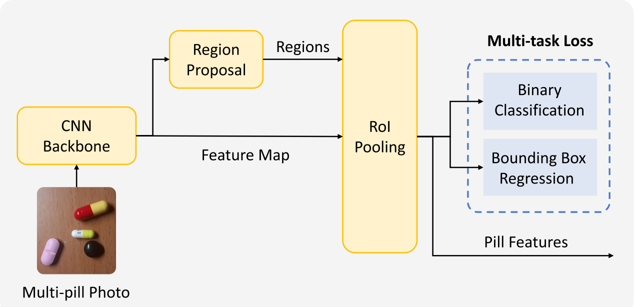


FIGURE 5. **Architecture of Faster R-CNN-based pill recognizer. The CNN backbone extracts the feature map from the input multi-pill photo. Region Proposal Network (RPN) identifies potential Regions of Interest (RoI) where the pills are located. Feature maps derived from RoI pooling are then used to detect and match pills to their respective prescriptions.**

The Prescription Recognizer aims to localize the text boxes containing pill names and generate their text embeddings. To accomplish this, we design a module consisting of three components: *Text Recognition*, *Transformer-based Text Embedding*, and *Graph Neural Network-based Pseudo Classifier*. Initially, the Text Recognition localizes the text boxes bounding texts in the prescription and extracts the texts. Let us denote by $\{b_1, \ldots, b_N\}$ the coordinates of the text boxes and $\{s_1, \ldots, s_N\}$ the corresponding texts (N is the number of texts in the prescription), then texts $\{s_1, \ldots, s_N\}$ are sent to the Transformer-based text embedding model [40] to produce text embedding vectors $\{\mathbf{t}^e_1, \ldots, \mathbf{t}^e_N\}$. The text

Algorithm 1 Algorithm for Learning Embedded Vectors of the Prescription Graph (Forward Propagation).

```
Input: Prescription graph $G = \{V, E\}$; input attribute $\{\mathbf{t}_v^e, \forall v \in V\}$, $\mathbf{t}^e$ is a text embedding; depth $K$; weight matrix $\mathbf{W}^k, \forall k \in \{1, \cdots, K\}$; non-linear function $\sigma$; differentiable aggregator functions $\text{AGGREGATE}_k$, $\forall k \in \{1, \cdots, K\}$, neighboring vertices $V_N : v \to 2^V$.
Output: Feature vectors $\mathbf{z}_v, \forall v \in V$.
$\mathbf{h}_v^0 \leftarrow \mathbf{t}_v^e, \forall v \in V$
for $k = 1 \cdots K$ do
    for $v \in V$ do
        $\mathbf{h}_{V_{N(v)}}^k \leftarrow \text{AGGREGATE}_k\left(\{\mathbf{h}_u^{k-1}, \forall u \in V_{N(v)}\}\right)$
        $\mathbf{h}_v^k = \sigma\left(W^k \cdot \text{MEAN}\left(\mathbf{h}_v^{k-1}, \mathbf{h}_{V_{N(v)}}^k\right)\right)$
    end for
    $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \left\|\mathbf{h}_v^k\right\|_2, \forall v \in V$
end for
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in V$
```

embedding vectors are then utilized in two ways: Firstly, in conjunction with the coordinates $\{b_1, \ldots, b_N\}$ of the text boxes, $\{\mathbf{t}_1^e, \ldots, \mathbf{t}_N^e\}$ are used to construct a graph representing the spatial relationship between them. The representation generated by a graph neural network is then fed into the pseudo-classifier to highlight text boxes containing pill names; Secondly, the text embedding vectors are sent to the Pill-Prescription alignment to perform the matching of pills' images and names.

1) TRANSFORMER-BASED TEXT EMBEDDING

We leverage the Transformer encoder to learn the text embeddings. Given a text $s_i = [w_1^{(i)}, \ldots, w_{l_i}^{(i)}]$ extracted from the prescription, where $w_t^{(i)}$ $(t = 1, \ldots, l_i)$ represents the t-th token of s_i, then the text embedding of s_i, denoted by $\mathbf{t}_i^e$, is obtained by feeding $[w_1^{(i)}, \ldots, w_{l_i}^{(i)}]$ into a transformer encoder. To enrich the information of the produced embeddings, we apply transfer learning to fine-tune the pre-trained text embeddings using the pills' metadata obtained from the DrugBank database [20]. The metadata consists of the color and shape of the pill. Specifically, we employ contrastive learning to contrast pill names and metadata, i.e., minimizing the distance between the name and metadata of the same pill and increasing it for distinct pills. The details of the contrastive loss are described in Section III-D4. Finally, we leverage a fully connected layer with skip-connection to project the text embeddings onto the same hyper-plane as their counterparts in the pill images. The final representations of the N text boxes are denoted by $T^p = \{\mathbf{t}_1^p, \cdots, \mathbf{t}_N^p\}$.

2) GRAPH NEURAL NETWORK-BASED PSEUDO CLASSIFIER

Prescriptions typically contain lots of noisy information, such as date, diagnosis, and note; identifying text boxes containing pill names is essential for the PIMA. Particularly for Zero-PIMA, where unseen pills have not been trained with their names, filtering out pill names from prescriptions helps reduce the misidentification of unseen pills with other texts (not pill names) in the prescription. To this end, our idea is to leverage the GCN to model the spatial relationship between text boxes in the prescription, differentiating between pill names and those that are not. We construct an unweighted graph $G = \{V, E\}$ with the vertices $V = \{v_1, \ldots, v_N\}$ representing the text boxes and the edges reflecting their relative positions in the prescription. To be more specific, each vertex v_i is associated with the attribue of its text embedding $\mathbf{t}_i^e$, and two vertices v_i and v_j are connected if one of them is the box with the shortest horizontal (or vertical) distance to the other. In this work, we leverage GraphSAGE [41], to convert from graph space to vector space. Any other GCN model can be used for this purpose, but investigating them is beyond the scope of this paper. The details of the forward propagation process in the prescription graph learning are presented in Algorithm 1. For each vertex v_i, we generate a graph embedding vector $\mathbf{z}_{v_i}$ that combines its textual information and relationship with neighbors within K-hops. This graph representation vector is then passed through the sigmoid activation function and the classifier to produce the classification result. The resulting vector $\mathbf{g} = (g_1, \ldots g_N)$ represents the probabilities for each text box to contain a pill name, i.e., g_i demonstrating the probability that the i-th text box contains a pill name. This pseudo-classifier is trained via a classification loss (see Section III-D for the details).

Finally, the pseudo classification result is multiplied by the text embeddings to obtain the weighted version, $T^{p'} = \{g_1\mathbf{t}_1^p, \ldots, g_N\mathbf{t}_N^p\}$, which emphasizes the most probable pill name while dimming the others.

D. LEARNING OBJECTIVES

We observe that the accuracy of the Zero-PIMA depends on four factors: (1) Localizing and extracting meaningful information about pill images, (2) Identifying and generating informative representations of pill names, (3)Matching pill images and names from the extracted ones, and (4) Capability in dealing with unseen pills whose images were unseen in the training process. To accomplish the first objective, rather than considering it as a multi-label classification, we instead utilize a binary classification loss to distinguish between the pill and non-pill bounding boxes (Section III-D1). For the second goal, we adopt GCN to model the spatial correlation between text boxes and design a cross-entropy loss to highlight boxes containing pill names (Section III-D2). The third objective is attained through a contrastive loss that compares the visual features of pill images and text embeddings extracted from the prescription (Section III-D3). Finally, we achieve the last goal using a semantic contrastive loss to finetune pre-trained text embeddings to capture better textual information from pill names (Section III-D4).

1) PILL DETECTOR LOSS

We adopt the multi-task loss following Faster-RCNN [24], which consists of a classification loss $\mathcal{L}_{\text{cls}}$ and regression loss $\mathcal{L}_{\text{reg}}$ to train the Pill Detector. Specifically, we define the Pill Detector's loss as follows:

$$\mathcal{L}_{\text{PD}} = \frac{1}{N_{\text{cls}}} \sum_{i} \mathcal{L}_{\text{cls}} \left(p_i, p_i^*\right) + \gamma \frac{1}{N_{\text{reg}}} \sum_{i} p_i^* \mathcal{L}_{\text{reg}} \left(t_i, t_i^*\right), \tag{1}$$

where i represents the index of an anchor in the mini-batch, and p_i the predicted probability that the anchor contains a pill object. The ground-truth label denoted as p_i^*, is set to 1 if the anchor represents a pill object and 0 otherwise. t_i is a vector that represents the coordinates of the predicted anchor, while t_i^* represents the ground-truth bounding box coordinates associated with the pill. The classification loss $\mathcal{L}_{\text{cls}}$ is computed over two classes (pill and non-pill objects), while the regression loss $\mathcal{L}_{\text{reg}}$ is calculated using the formula $\mathcal{L}_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*)$, where R represents the robust loss function (smooth L1) defined in [23]. Both losses are normalized by N_{cls} and N_{reg} and weighted by a balancing parameter γ, where N_{cls} is the mini-batch size and N_{reg} is the number of anchor locations.

2) PRESCRIPTION RECOGNIZER LOSS

We utilize the binary cross-entropy loss to identify whether a text box contains a pill name. We observe that the number of text boxes with pill names is significantly smaller than those without pill names. For this reason, we employ the following weighted cross-entropy loss to mitigate the bias:

$$\mathcal{L}_{\text{PR}} = -\frac{1}{N} \sum_{i=1}^{N} w_i \left[y_i \log\left(g_i\right) + \left(1 - y_i\right) \log\left(1 - g_i\right)\right], \tag{2}$$

where y_i and g_i represent the ground-truth label and the predicted result concerning a text box s_i, respectively, and w_i represents the ratio of text boxes with the label of $(1 - y_i)$. To be more specific, let N_{pill} be the number of text boxes with a pill name, and N the total number of text boxes, then w_i is determined as follows:

$$w_i = \begin{cases} 1 - \dfrac{N_{\text{pill}}}{N} & \text{, if text box } s_i \text{ contains a pill name,} \\ \dfrac{N_{\text{pill}}}{N} & \text{, otherwise.} \end{cases}$$

3) PILL-PRESCRIPTION CONTRASTIVE LOSS

This loss aims to model the cross-modal relationship between two modalities: pill image and pill name. The principle is to encourage the distance between representations of mismatched pill image and pill name pairs (a pill image and a name that do not correspond to the same medication), while minimizing those of the matched pairs (a pill image and its correct name, indicating they represent the same medication). Specifically, let $\mathbf{i}_i^p$ and $\mathbf{t}_j^p$ be the representations of a pill image p_i and a pill name s_j, respectively, then their similarity is defined by the cosine similarity as follows:

$$S\left(\mathbf{i}_i^p, \mathbf{t}_j^p\right) = \frac{\mathbf{i}_i^p \cdot \mathbf{t}_j^p}{\max\left(\left\|\mathbf{i}_i^p\right\|_2 \cdot \left\|\mathbf{t}_j^p\right\|_2, \varepsilon\right)},$$

where ε is a small offset responsible for avoiding the zero division problem. The learning objective $\mathcal{L}_{\text{PPC}}$ consists of two contrastive terms $\mathcal{L}_{I \to T}$ and $\mathcal{L}_{T \to I}$. The former is an image-to-text contrastive loss responsible for aligning the pill image corresponding to a given pill name, while the latter is a text-to-image contrastive loss responsible for matching the pill name with a given pill image. Details of the Pill-Prescription contrastive loss are as follows:

$$\begin{aligned} \mathcal{L}_{I \to T} &= -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\exp\left(S\left(\mathbf{i}_i^p, \mathbf{t}_i^p\right)^+ / \tau\right)}{\sum_{j=1}^{N_{\text{pill}}} \exp\left(S\left(\mathbf{i}_i^p, \mathbf{t}_j^p\right)^- / \tau\right)}, \\ \mathcal{L}_{T \to I} &= -\frac{1}{N_{\text{pill}}} \sum_{i=1}^{N_{\text{pill}}} \log \frac{\exp\left(S\left(\mathbf{t}_i^p, \mathbf{i}_i^p\right)^+ / \tau\right)}{\sum_{j=1}^{M} \exp\left(S\left(\mathbf{t}_i^p, \mathbf{i}_j^p\right)^- / \tau\right)}, \\ \mathcal{L}_{\text{PPC}} &= \mathcal{L}_{I \to T} + \mathcal{L}_{T \to I}, \end{aligned} \tag{3}$$

where M is the number of pill images, N_{pill} is the number of pill names, the symbol "+" represents a pair of samples that are similar, while the symbol "−" represents a pair of samples that are dissimilar, and τ is a temperature hyperparameter controlling the scaling of the distances between representations in the loss function. A lower temperature increases sensitivity, enhancing the distinction between positive and negative pairs, while a higher temperature reduces sensitivity, making the model less reactive to differences in similarity scores.

4) TEXT EMBEDDING'S FINETUNING LOSS

For realizing Zero-PIMA, matching unseen pills with their corresponding names in the prescription is one of the most crucial challenges. Obviously, in the cases of unseen pills, their pill images and names have not been included in the training data; relying on the visual appearance derived from the pill object image and the textual information learned from the prescription is insufficient for matching, as neither has been previously learned. To this end, we propose incorporating the metadata (i.e., color, shape) of pills retrieved from DrugBank [20] into their names. In this manner, the representations of pill names extracted from the prescription convey not only the textual information (from the name) but also the appearance indication (from the metadata) of pills, thereby improving the accuracy of pill image-names matching, particularly in the case of unseen pills. We employ a semantic contrastive loss to contrast pill names with the pills' metadata to minimize the distance between representations of a name ($\mathbf{e}^n$) and metadata ($\mathbf{e}^m$) belonging to the same pill while maximizing those of different pills.

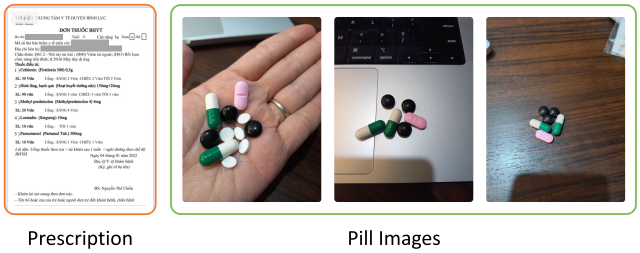


FIGURE 6. Representative examples from our VAIPE$_{\text{PP}}$ dataset. It was collected in real-world scenarios, where samples were taken in unconstrained environments.

TABLE 1. Comparison of the NIH and CURE datasets with our VAIPE$_{\text{PP}}$ dataset.

	NIH	CURE	VAIPE$_{\text{PP}}$
Number of pill objects	$7,000$	$8,973$	$6,366$
Number of pill classes	$1,000$	196	107
Instance's per category	7	$40-50$	>30
Illumination conditions	1	3	>50
Backgrounds	1	6	>50
Number of prescriptions	0	0	$1,527$

Details of the text embedding's finetuning loss are as follows:

$$\mathcal{L}_{\text{SC}} = -\frac{1}{P}\sum_{i=1}^{P}\log\frac{\exp\left(S\left(\mathbf{e}_i^n, \mathbf{e}_i^m\right)^+/\tau\right)}{\sum_{j=1}^{P}\exp\left(S\left(\mathbf{e}_i^n, \mathbf{e}_j^m\right)^-/\tau\right)}, \tag{4}$$

where P is the number of pill name-metadata pairs, S denotes the cosine similarity, the symbol "$+$" represents a pair of samples that are similar, while the symbol "$-$" represents a pair of samples that are dissimilar, and τ is a temperature hyperparameter.

The overall loss of the proposed model is determined by the weighted sum of all learning objectives (Eqn. 1, Eqn. 2, Eqn. 3, and Eqn. 4), which can be expressed as follows:

$$\mathcal{L}_{\text{Total}} = \lambda_1\mathcal{L}_{\text{PD}} + \lambda_2\mathcal{L}_{\text{PR}} + \lambda_3\mathcal{L}_{\text{PPC}} + \lambda_4\mathcal{L}_{\text{SC}}, \tag{5}$$

where λ_1, λ_2, λ_3, and λ_4 are balance coefficients.

IV. EXPERIMENTS

In this section, we introduce a thorough evaluation of the proposed method, Zero-PIMA, through comprehensive experiments. We compare Zero-PIMA with other object detection models and text-image retrieval methods under consistent experimental conditions. Additionally, we perform in-depth ablation studies to gain clearer insights into the key characteristics of Zero-PIMA.

TABLE 2. Example of illustration and pharmaceutical form details ("AMOXICILLIN 500mg" in this case).

Illustration	Pharmaceutical form	
	Color	Shape
	Pink and Blue	Capsule

To the best of our knowledge, previous studies addressing the Pill Detection problem have been restricted to datasets captured in laboratory environments with limited environmental conditions such as lighting, angle, and zoom level (e.g., NIH dataset [42]), typically containing only one pill per photo (e.g., CURE dataset [17]). As a result, these datasets do not accurately reflect reality, where patients may be taking multiple pills simultaneously to treat various symptoms. This limitation makes existing models less suitable for identifying pills in real-world medication photos taken by patients. In addition, it has been noted that there is a lack of publicly available datasets featuring pill images taken during actual patient consumption as well as corresponding prescription information. To fill in this gap, we devoted our efforts to building an open, large-scale dataset containing multi-pill photos and prescriptions called the VAIPE$_{\text{Pill}}$ and VAIPE$_{\text{Prescription}}$ datasets,[2] respectively. For Zero-PIMA, we selected a portion of these two datasets, referred to as VAIPE$_{\text{PP}}$. It consists of 2,156 multi-pill photos matching 1,527 prescriptions across 4 different templates. These were collected from anonymous patients at leading hospitals in Vietnam between 2021 and 2022. Following a thorough review for privacy concerns, the data were annotated by human annotators, with each prescription assigned relevant information. The pill intakes for each prescription were divided into morning, noon, and evening portions, with approximately five images taken for each portion. Figure 6 shows several representative examples from the dataset. Table 1 provides a summary of the meta-data details for the NIH [42], CURE [17], and the VAIPE$_{\text{PP}}$ datasets. The VAIPE$_{\text{PP}}$ dataset was constructed with a more flexible procedure, allowing fewer restrictions than the two conventional datasets, NIH and CURE. Due to this advantage, the VAIPE$_{\text{PP}}$ dataset demonstrates exceptional generalization capabilities, making it a trustworthy data source for training generic pill detection models.

A. CUSTOM DATASET

Furthermore, as a part of the data collection, we gathered and analyzed pill metadata from the DrugBank database [20]. This processed metadata includes various characteristics of the pills, including color and shape. However, due to resource constraints, we could only extract information for the pills collected in the VAIPE$_{\text{PP}}$ dataset. Table 2 gives an example of a pill and its corresponding metadata obtained from the DrugBank database.

[2] The dataset is made public from our project Web page at https://vaipe.org/##resource.

TABLE 3. Details of the data partition.

	Training set	Testing set		Total
		Seen	Unseen	
Number of classes	53	53	54	107
Number of pill images	1, 495 (69.34%)	350 (16.23%)	311 (14.43%)	2, 156 (100%)
Number of prescriptions	1, 057 (69.22%)	275 (18.00%)	195 (12.78%)	1, 527 (100%)

TABLE 4. Evaluation metrics.

Metrics	Description
mAP	Mean average AP at IoU threshold = $\overline{0.50, 0.95}$, step 0.05
AP_{50}	AP given IoU threshold = 0.50
AP_{75}	AP given IoU threshold = 0.75

B. EVALUATION METHODOLOGY

1) DATA SPLIT

We take a systematic approach to split the VAIPE_{PP} dataset into two distinct categories: seen classes ($D^s_{\text{test}} \subseteq D_{\text{train}}$) and unseen classes ($D^u_{\text{test}} \cap D_{\text{train}} = \emptyset$), where D_{train}, D^s_{test}, and D^u_{test} are training set, test set for seen classes, and test set for unseen classes, respectively. Our criteria for categorization are based on the frequency of each pill class in the collected prescriptions. Specifically, pills that are often prescribed are assigned to the seen classes, while those with a low occurrence frequency are placed in the unseen classes. Further details regarding this data split are demonstrated in Table 3. This data split allows us to evaluate the performance of the proposed model on both seen and unseen classes, providing a more comprehensive understanding of the model's generalization ability to new classes.

2) EVALUATION METRICS

We evaluate the performance of the proposed Zero-PIMA and other benchmarks using the Average Precision (AP) and mean Average Precision (mAP) metrics, which are commonly employed to assess the performance of object detection tasks. The AP metric measures the area under the Precision-Recall curve given an Intersection over Union (IoU) threshold. The IoU (defined as IoU $= \frac{\text{Intersection area}}{\text{Union area}}$) is the ratio of the overlapping region of a predicted bounding box and the corresponding ground truth to their intersecting union area. The choice of an IoU threshold determines whether a prediction is classified as a True Positive or a False Positive, thus impacting the AP results. To provide a comprehensive evaluation, we specifically use two settings of IoU thresholds: 0.50 and 0.75, denoted as AP_{50} and AP_{75}, respectively. These measurements provide valuable insights into the model's performance across different levels of bounding-box overlap. Additionally, the mAP is a comprehensive performance measure that takes into account the AP values within a specific range of IoU thresholds. In the following evaluation, we calculate mAP by averaging the AP values obtained for all classes, where the IoU thresholds range from 0.50 to 0.95 with an increment of 0.05. The evaluation metrics used for all experiments are summarized in Table 4.

3) BENCHMARK MODELS

In our evaluation, we compare the proposed Zero-PIMA with benchmark models on two distinct tasks: pill detection and pill-prescription matching.

For the first task, we evaluate the most popular object detection models.

- *Faster R-CNN model [24].* An object detection model that improves on Fast R-CNN [23] by utilizing RPN with the CNN model.
- *YOLOv8-S/M/L models [43].* The small/medium/large variants of the YOLO series is optimized for real-time object detection.
- *YOLOv8-L+RTDETR model [43].* Combines the large variant of YOLOv8 with Real-Time DEtection TRansformer (RTDETR) as a decoder, aiming to leverage the speed and accuracy of YOLO along with the efficient multiscale processing of DETR [30].
- *RTDETR-L model [44].* A variant of DETR optimized for real-time object detection, leveraging vision transformers to process multiscale features efficiently by decoupling intra-scale interaction and cross-scale fusion.

Concerning the second task, we compare Zero-PIMA with two baselines following the conventional approach. Specifically, we choose Faster R-CNN as the pill detection backbone and CLIP (Contrastive Language-Image Pre-Training) [45] as the multi-modal vision and language model and create two variants. The first one, denoted as *Faster-CLIP*, combines Faster R-CNN and vanilla CLIP, while the second, denoted as *Faster-CLIP (Text-finetuned)*, replaces vanilla CLIP by our text embedding model.

- *Faster-CLIP model.* We integrate the Faster R-CNN model with the CLIP model [45]. CLIP is a versatile model that handles a wide range of tasks by incorporating both visual and textual inputs. By combining the Faster R-CNN model with CLIP, we leverage the object localization results obtained from the Faster R-CNN as input for the CLIP model during vision-language tasks. The primary aim of the Faster-CLIP model is to explore the benefits of incorporating CLIP's vision-language capabilities for the Pill-Prescription matching task.
- *Faster-CLIP (Text-finetuned) model.* We explore the implementation of text fine-tuning to enhance the performance of the Faster-CLIP model further. The objective of the Faster-CLIP (Text-finetuned) model is to leverage textual information to improve the

TABLE 5. Experiment results of the proposed method compared to other methods. Best results in each task are highlighted in bold, while overall best results are underlined.

Task	Method	Backbone	Seen Accuracy			Unseen Accuracy		
			mAP	AP$_{50}$	AP$_{75}$	mAP	AP$_{50}$	AP$_{75}$
Pill Detection	Faster R-CNN [24]	MobileNetV3	54.74	78.90	66.89	—	—	—
		ResNet50	59.30	**83.99**	70.64	—	—	—
	YOLOv8-S [43]	CSPNet	56.92	77.60	68.56	—	—	—
	YOLOv8-M [43]	CSPNet	59.71	78.96	69.02	—	—	—
	YOLOv8-L [43]	CSPNet	**61.28**	82.04	**71.42**	—	—	—
	YOLOv8-L+RTDETR [43]	CSPNet	57.35	77.08	67.64	—	—	—
	RTDETR-L [44]	HGNetV2	59.28	78.44	67.17	—	—	—
Pill Prescription Matching	Faster-CLIP	MobileNetV3	59.08	82.78	70.15	21.33	28.99	25.41
		ResNet50	59.98	82.44	70.67	20.10	26.64	23.26
	Faster-CLIP (Text-finetuned)	MobileNetV3	62.83	90.33	75.68	45.91	61.99	54.02
		ResNet50	64.50	88.67	76.82	45.37	60.33	51.15
	Zero-PIMA (Proposed)	MobileNetV3	65.74	93.80	79.01	63.72	87.51	76.79
		ResNet50	**68.71**	**95.79**	**80.91**	**65.63**	**87.80**	**78.07**

object recognition performance and refine the joint vision-language representation of the model.

4) IMPLEMENTATION DETAILS

In our implementation, we ensure consistency using various pre-trained CNN models as the backbone network (i.e., MobileNetV3 [46], Residual Network (ResNet50) [47], Cross Stage Partial Network (CSPNet) [48], and High Performance GPU Network (HGNetV2) [44]). Within the pill-prescription matching task, we utilize pre-trained text embeddings, specifically MiniLM L12 multilingual [49], for both the proposed model and the comparison benchmarks. The input configurations for all models adhere to the requirements of the original architecture, while other parameters are fine-tuned for optimal performance. All implementations are performed using the PyTorch framework, and training is conducted for 100 epochs on a machine equipped with an NVIDIA V100 GPU (32 GB memory) and an Intel(R) Xeon(R) Gold 6248 CPU @ 2.50 GHz.

Proposed Model: We implement the proposed model based on the description provided in Section III. We use a projection layer consisting of two fully connected layers with Gaussian Error Linear Units (GELU) activation [50]. The input pill image size for the model is set to 224 × 224 × 3. The output dimension is set to 1 × 256, effectively capturing both visual and textual features. We compute contrastive loss in Eqn. 3 and Eqn. 4 by using all image and text pairs within a batch. The temperature hyperparameters in the contrastive loss are set to 1 for simplicity. The balance coefficients in Eqn. 5 are set equally to 1. For optimization, we employ AdamW [51] with an initial learning rate of 2.0×10^{-5}. During the training phase, we set the batch size to 8 to ensure efficient utilization of computational resources. To achieve optimal results, we carefully consider the model complexity and available training data size when selecting the hyperparameters.

C. BENCHMARK COMPARISON RESULTS

From now on, to ease the presentation, we use terms "seen accuracy" and "unseen accuracy" to indicate the accuracy of the models regarding the seen and unseen pill classes, respectively. Table 5 represents the accuracy of the proposed Zero-PIMA and the comparison benchmarks. We can see that Zero-PIMA outperformed the others in terms of all the evaluation metrics.

The experiment results, shown in Table 5, distinctly showcase the superiority of the proposed Zero-PIMA method over existing methods in pill detection and prescription matching tasks. Notably, the proposed method achieved the highest performance metrics, with its best results being underlined and highlighted in bold across both seen and unseen accuracy categories, specifically in terms of mAP, AP$_{50}$, and AP$_{75}$.

For the pill detection task, various methods have been used to evaluate the proposed dataset in the seen pill detection scenario. The comparison of Faster R-CNN, YOLO, and RTDETR approaches revealed notable differences in their performance metrics. Faster R-CNN showed the highest accuracy in AP$_{50}$ with 83.99%, while YOLOv8, particularly its large version (YOLOv8-L), outperformed others in mAP with 61.28%. Although RTDETR is one of the leading real-time object detection methods, the results only achieved 59.28% in mAP. In contrast to all models for pill detection tasks, the proposed method Zero-PIMA showed superiority, particularly when utilizing a ResNet50 backbone, marking a new benchmark with an mAP of 68.71%, AP$_{50}$ of 95.79%,

TABLE 6. Ablation study on different components of the proposed model on the unseen set. Best results are highlighted in bold.

Name	Graph Module	Text Finetune	mAP	AP_{50}	AP_{75}
Exp_1			24.49	35.45	30.69
Exp_2	✓		51.06	69.79	61.21
Exp_3		✓	63.35	86.80	75.23
Exp_4	✓	✓	**63.72**	**87.51**	**76.79**

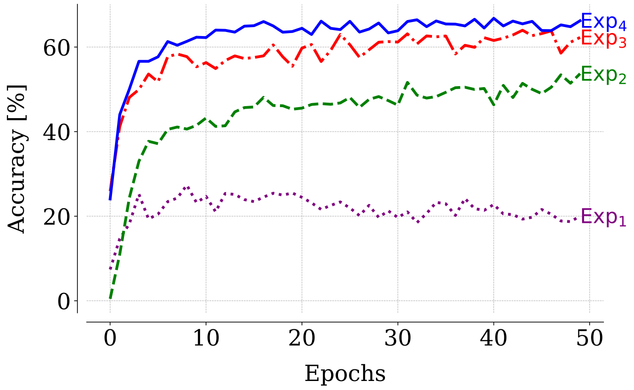


FIGURE 7. First 50 epochs of ablation study on different components of the proposed model on the unseen set (Table 6).

TABLE 7. Ablation study on different strategies involving the Graph module on the unseen set. Best results are highlighted in bold.

Strategy	mAP	AP_{50}	AP_{75}
Learnable	**63.72**	87.51	76.79
$\alpha = 0.7$	62.88	87.39	74.58
$\alpha = 0.8$	62.92	87.05	**76.89**
$\alpha = 0.9$	62.86	**87.57**	73.51

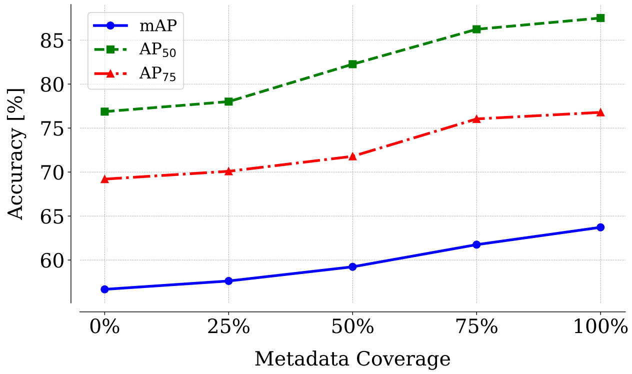


FIGURE 8. Ablation study on the impact of training pill metadata coverage on the unseen set.

and AP_{75} of 80.91% in seen pill detection. In the more challenging pill prescription matching task, the proposed method again proved its efficacy, significantly improving the unseen accuracy metrics to mAP of 65.63%, AP_{50} of 87.80%, and AP_{75} of 78.07%, utilizing the ResNet50 backbone.

These results validated the proposed method's capability to effectively bridge the gap between seen and unseen data performance, highlighting its potential for practical applications in the pharmaceutical field, where accurate detection and matching are crucial. Integrating advanced neural network architectures makes Zero-PIMA a promising approach for future research and application in pill identification and matching systems.

D. ABLATION STUDY

We conducte a series of excision experiments as part of an ablation study to evaluate the effectiveness of the proposed model in the unseen scenario ($D^u_{\text{test}} \cap D_{\text{train}} = \emptyset$). MobileNetV3 [46] is employed as the CNN backbone for all these experiments.

1) ASSESSING THE IMPACT OF COMPONENT REMOVAL ON MODEL EFFECTIVENESS

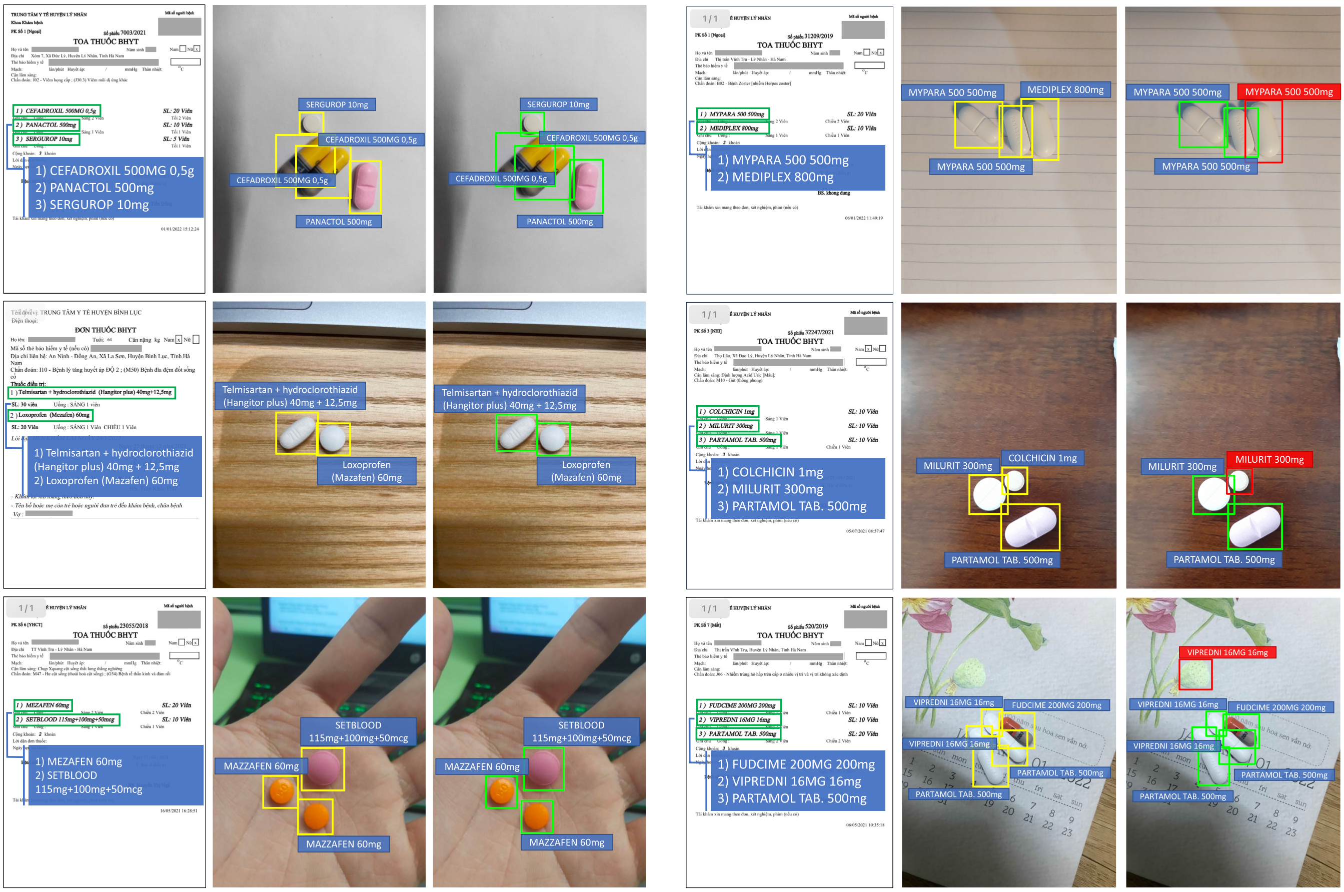
First, we aim to evaluate the impact of removing the Graph module and Text fine-tuning on the effectiveness of the proposed model. The results are presented in Table 6. In the absence of the Graph module and Text fine-tuning (Exp_1), the mAP achieved only 24.49%. However, when the Graph module (Exp_2) was integrated, there was a significant increase in the mAP accuracy, rising by 26.57%. This clearly illustrated the substantial impact of the Graph module. When Text fine-tuning was performed without the Graph module (Exp_3), the mAP showed a significant improvement of 38.83% compared to Exp_1. However, the highest accuracy was achieved when the Graph module was added in combination (Exp_4). Furthermore, Figure 7 illustrates that the model converged at a faster rate when the Graph module was present.

2) EVALUATING THE EFFECTIVENESS OF VARIOUS STRATEGIES FOR INCORPORATING THE GRAPH MODULE

We use various strategies to combine different Graph module approaches, as presented in Section III. Table 7 presents the results of using learnable or threshold-based strategies. The learnable approach involves directly multiplying the values from the Graph module with text embeddings, while a threshold-based approach selects text embeddings that meet the pre-set threshold α. The results demonstrated that both approaches yielded similar results. However, the learnable approach exhibited automatic efficiency on the data, whereas selecting the threshold relied on the distribution of each dataset, which may have led to differences or the loss of text containing pill names in prescriptions.

3) IMPACT OF TRAINING PILL METADATA COVERAGE ON THE UNSEEN SET

Next, we aim to investigate the impact of the coverage of pill metadata on the detection performance of the proposed model



(a) Illustration of some accurate predictions.

(b) Illustration of some incorrect predictions.

FIGURE 9. Visualization of some predictions for unseen pill detection. Each column presents the prescription, the ground-truth pill images, and the predictions.

on the unseen set. To achieve this, we evaluate the percentage of pill metadata covering the unseen set. The results shown in Figure 8 demonstrate that as the coverage of pill metadata decreased, the detection performance of the proposed model on the unseen set also decreased. However, we found that the Graph module, which we analyzed in the experiment of removing the Graph module in Section IV-D1, allowed the proposed model to detect pills even with limited information on the unseen set based on the prescription information. It is important to emphasize that the Graph module helped improve the accuracy of the proposed model in recognizing pills.

4) IMPACT OF DIFFERENT TRAINING PILL METADATA

Finally, we aim to evaluate the impact of different training pill metadata on the final result. Specifically, we examine the importance of shape and color information during the text fine-tuning process. The results presented in Table 8 demonstrate that both shape and color information significantly contributed to the final result. This ablation study holds significance as it emphasizes the need to consider different types of information when training models for pill recognition. We note that pills exhibit various shapes and colors, and this information is crucial for accurate identification. By including shape and color information in the training process, the proposed model learned to recognize unseen pills more accurately.

TABLE 8. Ablation study on the involvement of different training pill metadata on the unseen set. Best results are highlighted in bold.

Shape	Color	mAP	AP_{50}	AP_{75}
		51.06	69.79	61.21
✓		59.41	81.60	71.32
	✓	61.89	83.99	72.57
✓	✓	**63.72**	**87.51**	**76.79**

E. MODEL COMPLEXITY

Table 9 shows the complexity comparison between Zero-PIMA, Faster R-CNN, and Faster-CLIP. We observed that Faster R-CNN, which processes only pill images, had the

TABLE 9. Number of parameters and average inference time per sample. Faster R-CNN is designed to process only pill images, whereas Faster-CLIP and Zero-PIMA are intended to handle both pill images and their corresponding prescriptions.

Method	Backbone	Number of Parameters	Average Inference Time [s]
Faster R-CNN	MobileNetV3	18.87 M	0.00762
	ResNet50	41.07 M	0.02553
Faster-CLIP	MobileNetV3	21.90 M	0.03754
	ResNet50	44.10 M	0.04750
Zero-PIMA	MobileNetV3	20.02 M	0.03569
	ResNet50	42.22 M	0.04629

lowest number of parameters and the fastest inference times. In contrast, pill-prescription tasks, which require processing both the pill image and its corresponding prescription, lead to longer processing times for Zero-PIMA and Faster-CLIP. However, Zero-PIMA exhibited fewer parameters and achieved slightly faster inference times than Faster-CLIP. Comparing Zero-PIMA to Faster R-CNN, the parameter increase was just 1.15 M, indicating that integrating the GCN module into Zero-PIMA did not complicate the model.

F. QUALITATIVE VISUALIZATION

The visualizations in Figure 9 offer valuable insights into the proposed model's predictive capabilities and limitations. Figure 9(a) showcases instances where the model excels, accurately predicting the identification of various pills. These successes are notably attributed to the distinct variations in shape and color among the pills. Such results underscore the model's proficiency in recognizing and distinguishing pills based on these two critical features. Conversely, Figure 9(b) highlights scenarios where the model encounters difficulties. A recurring challenge arises when the model is presented with pills that have similar colors but differ in size—for example, between "MYPARA 500 500 mg" and "MEDIPLEX 800 mg" or "MILURIT 300 mg" and "COLCHICIN 1 mg". In these instances, the model struggles to differentiate between the pills, leading to misidentifications accurately. This limitation is further compounded by errors in detecting pills against complex backgrounds, where parts of the background can be mistaken for the pills themselves or obscure their visibility.

G. DISCUSSION

A notable limitation of this research is the reliance on four prescription templates, all formatted as lists. This structure allows the GCN module to excel in extracting the correct pill names from prescriptions. However, it may be a challenge to extract the pill name when prescriptions are presented in table formats. The accurate extraction of pill names is especially critical in cases of unseen pill detection, where the text and pills have not been previously matched during the training phase.

Moreover, in unseen scenarios, the proposed model's reliance on shape and color for pill identification may lead to inaccuracies when pills within the same prescription have similar appearances in terms of color and size. Such a narrow focus can lead to the failure of detection.

Furthermore, this research identifies a critical failure point in cases of incorrect pill localization. Mislocalization directly impacts the matching process, leading to incorrect identifications. This highlights a significant area for improvement in future model iterations, suggesting a need for enhanced localization techniques or incorporating additional distinguishing features beyond shape and color to improve accuracy and reliability in pill identification, particularly in challenging or unseen scenarios.

V. CONCLUSION

In this paper, we proposed a novel method for solving the zero-shot pill recognition and prescription matching task using GCN and Contrastive learning. The proposed method was evaluated on a real-world dataset that included actually prescribed pills. In addition to the proposed method, we also fine-tuned text embedding with pill metadata for the purpose of recognizing pills that were not included in the training data. The results showed that the proposed method outperformed other approaches in both seen and unseen accuracy in terms of mAP. We have made the source code for the proposed Zero-PIMA method available,[3] which can be accessed for further research and development in this domain.

For future work, we plan to explore the relationships among pills and consider additional attributes in the prescription, such as dosage quantities, to further enhance the accuracy of pill identification and detection. By incorporating these factors, we aim to improve the robustness and reliability of the proposed model in real-world scenarios. Furthermore, we envision deploying the proposed method in practical settings to assist healthcare professionals and patients in accurately identifying and matching the prescribed pills. This deployment can provide valuable support in healthcare services, ensuring the safe and effective use of medications.

ACKNOWLEDGMENT

The computation was carried out using the General Projects on the supercomputer "Flow" with the Information Technology Center, Nagoya University.

REFERENCES

[1] O. Mohamed Ibrahim, R. M. Ibrahim, A. Z. Al Meslamani, and N. Al Mazrouei, "Dispensing errors in community pharmacies in the united Arab emirates: Investigating incidence, types, severity, and causes," *Pharmacy Pract.*, vol. 18, no. 4, p. 2111, Oct. 2020.

[2] K. G. Zirpe, B. Seta, S. Gholap, K. Aurangabadi, S. K. Gurav, A. M. Deshmukh, P. Wankhede, P. Suryawanshi, S. Vasanth, M. Kurian, P. Elizabeth, J. Nirmala, and P. Esther, "Incidence of medication error in critical care unit of a tertiary care hospital: Where do we stand?" *Indian J. Crit. Care Med.*, vol. 24, no. 9, pp. 799–803, 2020.

[3] The code and pre-trained weights are publicly available at https://github.com/thanhhff/Zero-PIMA/.

[3] M. A. Sim, L. K. Ti, S. Mujumdar, S. T. H. Chew, D. J. B. Penanueva, B. M. Kumar, and S. B. L. Ang, "Sustaining the gains: A 7-Year follow-through of a hospital-wide patient safety improvement project on hospital-wide adverse event outcomes and patient safety culture," *J. Patient Saf.*, vol. 18, no. 1, pp. e189–e195, 2022.

[4] M. Aseeri, G. Banasser, O. Baduhduh, S. Baksh, and N. Ghalibi, "Evaluation of medication error incident reports at a tertiary care hospital," *Pharmacy*, vol. 8, no. 2, p. 69, Apr. 2020.

[5] R. A. Tariq, R. Vashisht, A. Sinha, and Y. Scherbak, "Medication dispensing errors and prevention," in *StatPearls [Internet]*. Treasure Island, FL, USA: StatPearls Publishing, 2023.

[6] S. Salmasi, T. M. Khan, Y. H. Hong, L. C. Ming, and T. W. Wong, "Medication errors in the Southeast Asian countries: A systematic review," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0136545.

[7] J. K. Aronson, "Medication errors: Definitions and classification," *Brit. J. Clin. Pharmacol.*, vol. 67, no. 6, pp. 599–604, Jun. 2009.

[8] L. P. Valdez, A. de Guzman, and R. Escolar-Chua, "A structural equation modeling of the factors affecting student nurses' medication errors," *Nurse Educ. Today*, vol. 33, no. 3, pp. 222–228, Mar. 2013.

[9] H.-T. Nguyen, T.-D. Nguyen, F. M. Haaijer-Ruskamp, and K. Taxis, "Errors in preparation and administration of insulin in two urban Vietnamese hospitals: An observational study," *Nursing Res.*, vol. 63, no. 1, pp. 68–72, 2014.

[10] S. S. Chua, H. M. Chua, and A. Omar, "Drug administration errors in paediatric wards: A direct observation approach," *Eur. J. Pediatrics*, vol. 169, no. 5, pp. 603–611, May 2010.

[11] B. Anh and T. Thang, "Vietnamese teen boy mistakenly consumes 21 birth control pills at a time," Tuoi Tre News, Vietnam, Oct. 2022.

[12] A. Nam, "Vietnamese woman loses baby after being prescribed abortion pills," VnExpress Int., Vietnam, Apr. 2018.

[13] H.-J. Kwon, H.-G. Kim, and S.-H. Lee, "Pill detection model for medicine inspection based on deep learning," *Chemosensors*, vol. 10, no. 1, p. 4, Dec. 2021.

[14] L. Tan, T. Huangfu, L. Wu, and W. Chen, "Comparison of RetinaNet, SSD, and YOLO V3 for real-time pill identification," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, pp. 1–11, Dec. 2021.

[15] Y. Y. Ou, A. C. Tsai, J. F. Wang, and J. Lin, "Automatic drug pills detection based on convolution neural network," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Oct. 2018, pp. 1–4.

[16] Y. Wang, J. Ribera, C. Liu, S. Yarlagadda, and F. Zhu, "Pill recognition using minimal labeled data," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2017, pp. 346–353.

[17] S. Ling, A. Pastor, J. Li, Z. Che, J. Wang, J. Kim, and P. Le Callet, "Few-shot pill recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9789–9798.

[18] United States Food & Drug Agency, *New Drug Therapy Approvals 2022*, FDA's Center Drug Eval. Res. (CDER), USA, 2022.

[19] T. T. Nguyen, H. D. Nguyen, T. H. Nguyen, H. H. Pham, I. Ide, and P. L. Nguyen, "A novel approach for pill-prescription matching with gnn assistance and contrastive learning," in *Proc. 19th Pacific Rim Int. Conf. Artif. Intell.* Shanghai, China: Springer, 2022, pp. 261–274.

[20] D. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res.*, vol. 34, pp. D668–D672, Jan. 2006.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Dec. 2001, pp. 511–518.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.

[26] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, Jun. 2015.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 213–229.

[31] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proc. 15th Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 384–400.

[32] S. Rahman, S. Khan, and N. Barnes, "Improved visual-semantic alignment for zero-shot object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11932–11939.

[33] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14084–14093.

[34] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "F-VLM: Open-vocabulary object detection upon frozen vision and language models," 2022, *arXiv:2209.15639*.

[35] S. Wu, W. Zhang, S. Jin, W. Liu, and C. Change Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15254–15264.

[36] Y. F. Wong, H. T. Ng, K. Y. Leung, K. Y. Chan, S. Y. Chan, and C. C. Loy, "Development of fine-grained pill identification algorithm using deep convolutional network," *J. Biomed. Inform.*, vol. 74, pp. 130–136, Oct. 2017.

[37] S. Mallat, "Understanding deep convolutional networks," *Philos. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, pp. 1–16, 2016.

[38] A. D. Nguyen, T. D. Nguyen, H. H. Pham, T. H. Nguyen, and P. L. Nguyen, "Image-based contextual pill recognition with medical knowledge graph assistance," in *Proc. 14th Asian Conf. Intell. Inf. Database Syst. (ACIIDS)*. Ho Chi Minh City, Vietnam: Springer, 2022, pp. 354–369.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Georgia, Oct. 2017, pp. 2961–2969.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[41] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1025–1035.

[42] Z. Yaniv, J. Faruque, S. Howe, K. Dunn, D. Sharlip, A. Bond, P. Perillan, O. Bodenreider, M. J. Ackerman, and T. S. Yoo, "The national library of medicine pill image recognition challenge: An initial report," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2016, pp. 1–9.

[43] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLO*. Accessed: Mar. 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[44] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[46] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," 2019, *arXiv:1905.02244*.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.

[49] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 3973–3983.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[51] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.

TRUNG THANH NGUYEN (Graduate Student Member, IEEE) received the B.Eng. degree in computer science from Hanoi University of Science and Technology, Vietnam, in 2022. He is currently pursuing the master's degree with the Graduate School of Informatics, Nagoya University, Japan. His research interests include computer vision, multimedia, and multimodal recognition.

PHI LE NGUYEN (Member, IEEE) received the B.E. and M.S. degrees from The University of Tokyo, Tokyo, Japan, in 2007 and 2010, respectively, and the Ph.D. degree in informatics from The Graduate University for Advanced Studies, National Institute of Informatics, Tokyo, in 2019. She is currently an Associate Professor with the School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam. Her research interests include network architectures and applied AI in various domains, such as smart healthcare, environment, and next-generation networks.

YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. In 2012, he became a Postdoctoral Fellow with Kyoto University. In 2014, he moved to Nagoya University, Japan, as a Designated Assistant Professor, where he became an Assistant Professor, in 2015, and a Lecturer, in 2020. Since 2021, he has been the Team Leader of the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and pattern recognition for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He is a member of IIEEJ and IEICE. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter.

TAKAHIRO KOMAMIZU (Member, IEEE) received the B.Eng. degree in computer science, the M.Eng. degree, and the Ph.D. degree in engineering from the University of Tsukuba, Japan, in 2009, 2011, and 2015, respectively. He became a Postdoctoral Researcher with the University of Tsukuba, in 2015, and an Assistant Professor with the Information Technology Center and a Designated Lecturer with the Institutes of Innovation for Future Society, Nagoya University, in 2018 and 2021, respectively. Since 2022, he has been an Associate Professor with the Mathematical and Data Science Center, Nagoya University. His research interests include databases, data analysis, linked open data, and multimedia data management. He is a member of ACM, IPSJ, IEICE, DBSJ, NLP, and JSAI.

ICHIRO IDE (Senior Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from The University of Tokyo, in 1994, 1996, and 2000, respectively. He became an Assistant Professor with the National Institute of Informatics, Japan, in 2000, and an Associate Professor with Nagoya University, Japan, in 2004, where he has been a Professor, since 2020. He was a Visiting Associate Professor with the National Institute of Informatics, from 2004 to 2010, an Invited Professor with Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), France, in 2005, 2006, and 2007, and a Senior Visiting Researcher with ISLA, Instituut voor Informatica, Universiteit van Amsterdam, The Netherlands, from 2010 to 2011. His research interests include the analysis and indexing to authoring and generation of multimedia content, especially in large-scale broadcast video archives and social media, mostly on news, cooking, and sports content. He is a Senior Member of IEICE and IPS Japan and a member of ACM, JSAI, and ITE.