Few-Shot Pill Recognition

Suiyi Ling\textsuperscript{1,*}, Andrés Pastor\textsuperscript{1,*}, Jing Li\textsuperscript{2}, Zhaohui Che\textsuperscript{3}, Junle Wang\textsuperscript{4}, Jieun Kim\textsuperscript{5}, Patrick Le Callet\textsuperscript{1}

\textsuperscript{1}University of Nantes \textsuperscript{2}Alibaba Group \textsuperscript{3}Shanghai Jiao Tong University \textsuperscript{4}Tencent \textsuperscript{5}Hanyang University

\{suiyi.ling, andreas.pastor, patrick.lecallet\}@univ-nantes.fr jing.li.univ@gmail.com chezhaohui@sjtu.edu.cn wangjunle@gmail.com jkim2@hanyang.ac.kr

Abstract

Pill image recognition is vital for many personal/public health-care applications and should be robust to diverse unconstrained real-world conditions. Most existing pill recognition models are limited in tackling this challenging few-shot learning problem due to the insufficient instances per category. With limited training data, neural network based models have limitations in discovering most discriminating features, or going deeper. Especially, existing models fail to handle the hard samples taken under less controlled imaging conditions. In this study, a new pill image database, namely CURE, is first developed with more varied imaging conditions and instances for each pill category. Secondly, a light-weight W\textsuperscript{2}-net is proposed for better pill segmentation. Thirdly, a Multi-Stream (MS) deep network that captures task-related features along with a novel two-stage training methodology are proposed. Within the proposed framework, a Batch All strategy that considers all the samples is first employed for the sub-streams, and then a Batch Hard strategy that considers only the hard samples mined in the first stage is utilized for the fusion network. By doing so, complex samples that could not be represented by one type of feature could be focused and the model could be forced to exploit other domain-related information more effectively.

Experiment results show that the proposed model outperforms state-of-the-art models on both the National Institute of Health (NIH) and our CURE database.

1. Introduction

Accurately recognizing prescription pill images according to their visual appearance helps to ensure patients’ safety and facilitate contemporary healthcare system for patients/old people. Furthermore, it can be useful in avoiding errors across the pharmacological chain; it can also improve the care provided by experts on poison control [28], increase medication persistence [4], minimize the loss of medications and prescriptions in evacuation scenarios [22], and promote the development of remote/self-diagnosis technology and smart health-care applications [34].

However, accurate pill recognition in daily life is usually hindered by the few-shot learning problem, where the number of samples per category is small; for instance, the NIH dataset [35] contains only 7 samples per category. Moreover, although there are various commercial products and web-based services for pill image identification, no complete solution has been found to make the system sufficiently robust to different noisy imaging conditions in both professional and general public healthcare services. In academic literature, most existing pill recognition models fail in the few-shot regime. This failure is more likely under less-controlled noisy imaging conditions, especially regarding the hard samples. There are mainly two types of hard samples: 1) hard negatives: different pill categories with similar visual appearance; 2) hard positives: pills under the same category but with significantly different visual characteristics due to the noisy imaging conditions. For instance,
in Figure 2 (a), the same pill under different lighting conditions has different colors, while in Figure 2 (b), the three different pills tend to be classified under the same pill category by existing pill recognizers because of the similarity in shape and color. The main characteristic that distinguishes the three different pills in Figure 2 (b) is the imprinted texts, which are however difficult to identify even for human eyes. As noted in [17], imprinted text plays a key role in facilitating accurate pill recognition; thus approaches make better use of the domain-related information, such as text information, could hold the key to more effective pill recognition with limited data. According to our observations, more imprinted information can be gleaned from the pill’s texture. For example, Figure 2 (c) shows the texture map of (b). A comparison of the texture maps and the original RGB images shows that the imprinted texts on the pills are more visible in the texture maps.

Therefore, we propose a Multi-Stream (MS) deep learning model based on a novel two-stage training strategy, where individual streams are first trained using the Batch All (BA) strategy that considers all the samples, in addition to a late fusion process using the Batch Hard (BH) strategy that solely focuses on the hard samples that could not be processed by the individual streams in the preceding training stage. It is worthy to note that the BH proposed in this study is different from the one in [11], which selects only the hardest positive/negative samples from each batch using min/max function. The overall framework is depicted in Figure 1. More specifically, a $W^2$-net is first proposed to extract the pills regions from the background. Using the segmented pill regions, we trained three streams that process the RGB image, contour, and texture maps separately using the triplet loss with the BA strategy. Furthermore, we retrained the Deep Text Spotter (DTS) [1] that detects and recognizes imprinted texts on the texture maps of pill image as the fourth stream. Finally, we trained a fusion network to combine the four streams using triplet loss considering only the hard samples that violate the triplet constraint in the first stage, along with the imprinted text information provided by the retrained DTS. Specifically, this scheme facilitates the compensation of different features with the auxiliary information of the high-level imprinted text.

2. Related Work

Pill dataset: Recently, the U.S. National Library of Medicine (NLM) of the NIH released a pill image dataset and called for submission of prescription pill images recognition models [35]. However, the images in the NIH dataset have limitations on lighting, background conditions, and equipment, among others. The summary of the NIH database is shown in Table 1.

| Pill recognition model: In addition to designing invariant descriptors for identifying pills, Caban et al. [2] proposed a modified shape distribution technique for examining the shape, color, and imprinted text of pills. However, the imprint descriptors within the model are limited, and the images considered are not representative of the variability of practical situations. In [13], the structure-related features of pills were exploited by first localizing the pills within query images according to non-zero gradient magnitude. Nevertheless, this model may not be applicable under a less-controlled imaging condition. Similar features were considered in [3, 5] to estimate the size of pills, and recognize them. Unfortunately, these methods disregard the fact that the sizes of pills can easily change under different zooming effects. Yu et al. [36] suggested exploiting the shape and other features of pills to represent the imprinted symbols on pills; however, this method fails in cases, where the imprints of the captured pill images are obscure or invisible to humans.

| Table 1. Comparison of the CURE and the NIH dataset. |
|-----------------|-----------------|-----------------|
| Number of pill images | NIH NLM | CURE |
| Number of pill categories | 1000 | 196 |
| Instance per category | 7 | 40-50 |
| Illumination conditions | 1 | 3 |
| Backgrounds | 1 | 6 |
| Imprinted text labels | No | yes |
| Segmentation labels | No | partially labeled |

Apart from the approaches that exploit the traditional handcrafted features of pills, recently, with the breakthrough of deep learning in computer vision and image processing, four deep-features-based methods [35] have been proposed, and have yielded reasonable results on the NLM NIH pill image recognition challenge. Among them, Mobile Deep Pill (MDP) [37], one of the state-of-the-art proposals, won the first prize in the challenge. First, three convolutional neural networks (CNN) that take RGB image, grayscale image, and edge maps as input correspondingly are trained; then, the dissimilarity values calculated using each single CNN model for pill recognition are linearly summed.

Although some of the aforementioned models consider the imprinted text on pills, they solely use structural de-
scripts and do not attempt to recognize the symbols on the pills. Furthermore, most of them fail to cater for the hard samples described in the previous section under noisy conditions as they simply extract different features, and proceed straightforward to train the classifiers without considering the complementary relations between different features using a well-designed learning strategy.

**Few-shot learning** algorithms have been developed and proven to be a promising tool in small data scenarios. They could be categorized as 1) the metric learning based approach [31, 27, 30, 33] whereby a similarity metric/space is learned; 2) the memory network approach [19, 25, 23, 9] whereby the model is trained to store ‘experience’; 3) The gradient-descent-based approach [7, 9], where a meta-learner is trained to adapt a base-learner through different tasks. Since most of these models use shallow networks to avoid the over-fitting problem with limited samples, their performances are limited. To tackle this limitation, MTL [29] was proposed to utilize a deep network for few-shot cases based on the hard task meta-batch strategy. In [14], CTM was proposed to handle few-shot problem by select the most relevant feature dimensions after traversing both across and within categories. Nevertheless, none of these models are designed for pill recognition. Therefore, they do not sufficiently exploit the domain-related information in small data scenarios to deal with the hard samples.

### 3. Proposed CURE Pill database

In this section, the novel CURE pill dataset is introduced. This dataset summarized in Table 1, contains 8973 images of 196 categories, and approximately 45 samples were obtained for each pill category.

The illumination conditions include 1) indoor light, 2) weak outdoor light, and 3) strong outdoor light. Examples of the reference images in the dataset are shown in the first column of Figure 4. We believe that generating reference images using consumer images is more practical, because of the following reasons: 1) In cases where reference images are uploaded by the pill manufacturers, high-quality cameras could be too expensive and collecting professional images under professionally controlled conditions is more time consuming; 2) In cases where reference images are uploaded by consumers, the developed pill recognition models should achieve acceptable performances even with lower quality reference images. The reference images are labeled with pixel-wise pill location and imprinted text/symbols.

As summarized in Table 1, this dataset considers more challenging real-world conditions (i.e., with more diverse backgrounds, light, and zooming conditions); thus, it reflects practical cases better, compared to the NIH dataset [35]. Examples of images in the dataset are shown in Figure 4. as observed, 1) images in the last row were taken under different light condition, which could result in sig-
significant changes to the pill color (especially for (h), where the color of the images taken under different lighting conditions with the MPI equipment vary significantly); 2) (c) and (d) are taken under different zooming conditions; 3) the backgrounds considered in this dataset are diverse.

4. The Proposed Model

4.1. Pill Segmentation and Localization

Backgrounds of pill images provide few useful information and could even deteriorate the training process of pill recognizer as a source of noises. Dealing with different noisy backgrounds, challenging lighting, and zoom in/out conditions necessitates a model that yields more precise segmentation results. Thus, we propose a $W^2$-net to delineate the pill regions from the backgrounds so that pill recognizer can be trained on localized pill images, and ignore the perturbations from noisy and superfluous backgrounds.

It has been demonstrated in [21] that repeated bottom-up, top-down processing used in conjunction with intermediate supervision is critical for improving the performance of a network. Inspired by this idea and the concept of Knowledge Distilling [12], the proposed $W^2$-net was constructed using four simplified $U$-Net [24]. It is worth noting that, $W^2$ is 17.5 times smaller than $U$-Net, i.e., 2M vs. 35M. This was achieved through 1) Using 1.4% of the parameters of the original $U$-net for each simplified $U$-net; 2) Feeding the intermediate output from the previous simplified $U$-Net into the next one. The detailed network architecture is shown in the supplemental material. As there are only two categories in our study, i.e., background and pill regions, we employ the pixel-wise binary cross-entropy loss for the $i_{th}$ simplified $U$-net:

$$L_{U_i} = \sum_{p \in P_i} -l(p) \cdot \log(s(p)) + (1 - l(p)) \cdot \log(1 - s(p)), \tag{1}$$

where $l(\cdot)$ is the true label of each pixel $p \in P_i$ and $s(\cdot)$ is the score predicted by the $i_{th}$ $U$-net with sigmoid function as activation function. The loss function of the proposed $W^2$-net is then defined as:

$$L_{W^2} = \sum_{i=1}^{4} \lambda_{U_i} L_{U_i}, \tag{2}$$

where $\lambda_{U_i}$ are the parameters balancing the losses of the corresponding simplified $U$-nets and are set equally to 1/4 in this study. In the following sections, only the segmented/located pill images are considered.

4.2. Multi-Stream CNN for Pill Recognition

4.2.1 Metric Embedding Learning using Triplet Loss

Pill recognition is a typical few-shot learning problem, where insufficient data is available for each pill class. Recently, effective few-shot methodologies adapted a metric-learning scheme to learn an similarity metric to compare the difference between a test/query example and the few ones used in training [14, 31]. In this study, to better handle the positive/negative hard samples described in Section 1, the triplet loss was employed to optimize the similarity metric (embedding space) such that images of the same pill are closer to each other and inversely for the different ones.

Theoretically speaking, given a set of triplets $(I_a, I_p, I_n)$ ($I_a$ is considered as an anchor image, $I_p$ is the positive sample that is of the same category with $I_a$, while $I_n$ is the negative sample that is of different category), the goal of metric embedding learning is to learn a function $f_\theta(I) : \mathbb{R}^F \rightarrow \mathbb{R}^E$ parametrized by $\theta$ to map similar or different pill images, i.e., $(I_a, I_p)$ for same pills or $(I_a, I_n)$ for different pills, from the feature manifold $\mathbb{R}^F$ onto metrically close/far points in an embedding space $\mathbb{R}^E$ with the objective function defined as [26]:

$$L_{tri}(\theta) = \sum_{a,p,n \; y_a=y_p \neq y_n} [m + D(f_\theta(I_a), f_\theta(I_p)) - D(f_\theta(I_a), f_\theta(I_n))]_+, \tag{3}$$

where $m$ is a margin that is enforced between positive and negative pairs [26], $[x]_+ = \max\{0, x\}$. $y_i$ is the pill category label for the $i_{th}$ sample, $D(f_\theta(I_a), f_\theta(I_n)) : \mathbb{R}^E \times \mathbb{R}^E \rightarrow \mathbb{R}$ denotes the metric function that measures the distances between two images $I_a, I_n$ in the embedding space. Throughout this study, Euclidean distance was used as the distance measure, i.e., $D(\cdot, \cdot)$, as in [11]. Nevertheless, pill recognition model that processes different features separately, e.g., MDP [37], cannot effectively handle the hard samples. Thus, it is important to devise a proper strategy for training considering the fact that 1) If using all the possible triplets, the number of triplets will increase cubically with the growth of data numbers, rendering the training inefficient; 2) If only the hardest triplets are considered, the model would select the outliers in the dataset, resulting in the failure of $f_\theta$ in learning the 'normal' associations [11]; 3) Training separate networks using different features as done in [37] neglects the complementary relationships between different features.

To this end, we propose a MS CNN that is composed of four individual streams (stream RGB, Texture, Contour and Imprinted Text) that are consecutively combined with a late fusion network. The proposed MS CNN was trained in a stage-wise manner, similarly to [6]. In the first training stage, we used the BA strategy to train the stream RGB, Texture, and Contour individually, where all the samples are considered. In this stage, the hard samples that could not be tackled using each stream alone were selected for the second stage. For example, for stream RGB, the hard samples could be the same pill but under different illumination conditions as shown in Figure 2 (a); for stream Texture, hard samples could be different pills with the same texture,
shape, imprinted text but different color; for stream Contour, hard samples could be different pills with the same shape but different texture or imprinted text (when imprinted texts are visually unclear). In the second training stage, we propose a new BH strategy to train the fusion net to combine the three individual streams along with a retrained imprinted text stream (based on text-regions detection and recognition). During the second training stage, the four streams were fixed and the fusion network was trained using only the hard samples mined in the first stage. The rationale behind this two-stage training strategy is to build a bridge between different feature spaces so that they could compensate each other by concentrating on the hard samples excavated in the first training stage. Details are given in the following subsections aligned with the two-stage learning procedure, which is summarized in Algorithm 1.

### 4.2.2 RGB, Texture, and Contour Streams

Among the hand-crafted features based pill recognition models [2, 13, 3, 5, 36], color, contour and texture related descriptors were commonly considered and proven to be effective in the task. Therefore, we selected color, contour and texture channel (task-related channels) as the input of the sub-streams \( f_{rgb}(\cdot) \), \( f_{texture}(\cdot) \), and \( f_{contour}(\cdot) \) empirically. The first training stage is summarized in Algorithm 1 (line: 3-16). Details are described below.

Firstly, a low-pass-based model [15] is adapted in this paper to obtain clearer contour and texture maps. More specifically, as depicted in Figure 1, given the gray-level image \( I_{gray} \) of a segmented pill image \( I_{RGB} \), the contour map \( I_c \) is obtained by employing Canny edge detector on the response \( I_{res} \) of using Gaussian filter on \( I_{gray} \). Then, the residual that maintains the high-frequency components, where texture information is emphasized, is obtained by subtracting the \( I_{res} \) from \( I_{gray} \). Here, it is named as the texture map and denoted as \( I_t \). Afterwards, the three individual streams taking RGB image \( I_{RGB} \), contour map \( I_c \), and texture map \( I_t \) as input separately using the BA strategy [11]:

\[
\mathcal{L}_{BA}(\theta, X_b) = \sum_{i=1}^{P} \sum_{a=1}^{K} \sum_{j=1}^{K} \sum_{n=1}^{m} \left[ m + d_{j,a,n}^{a,p} + \right],
\]

\[
d_{j,a,n}^{a,p} = D(f_\theta(I_a^i), f_\theta(I_p^j)) - D(f_\theta(I_a^i), f_\theta(I_h^n)),
\]

where \( I_j \) denotes the data points of the \( j \)th instance for the \( i \)th pill category in the current mini-batch \( X_b \). \( P \) and \( K \) are the numbers of randomly sampled pills categories and the corresponding pill images in each batch. For each batch, all possible \( PK(K-K)(K-1) \) combination of triplets were considered. During the first training stage, the hard samples that violated the constraint \( d_{j,a,n}^{a,p} < m \) were forwarded to the second training stage.

The usage of a pretrained network could lead to a design lock-in [11]. Therefore, in this study, we designed the three individual streams from scratch (details of the network architectures are summarized in the supplementary material).

### Algorithm 1 Two stage BA-BH learning strategy.

1. **Input:** Data set \( X \). Triplet generator \( \tau(g) \).
2. **Output:** Multi-Stream pill recognizer \( f_{MS}(\cdot) \).
3. **Stage 1** Task-related streams training (Section 4.2.2):
   4. Randomly initialize three individual streams \( f_{rgb}(\cdot), f_{texture}(\cdot), \) and \( f_{contour}(\cdot) \).
   5. Initialize Hard triplet set: \( H_{str} \leftarrow \emptyset \), where \( str = \{rgb, contour, texture\} \).
   6. for batch \( X_b \) in \( X \) do
      7. \( T_a = \tau(g)(X_b), T_a \) is the set contains all the possible triplets from batch \( X_b \).
      8. Evaluate \( \mathcal{L}_{BA}(f_{rgb}(\cdot), T_a), \mathcal{L}_{BA}(f_{texture}(\cdot), T_a), \) and \( \mathcal{L}_{BA}(f_{contour}(\cdot), T_a) \) by Eq.(4).
      9. Optimize \( f_{rgb}(\cdot), f_{texture}(\cdot), \) and \( f_{contour}(\cdot) \) by Adam optimizer.
   10. end for

11. Freeze \( f_{rgb}(\cdot), f_{texture}(\cdot), \) and \( f_{contour}(\cdot) \).
12. for any triplet \( (I_a^i, I_p^j, I_h^n) \in T_a(X) \) do
      13. if \( D(f_{str}(I_a^i), f_{str}(I_p^j)) - D(f_{str}(I_a^i), f_{str}(I_h^n)) < m \), where \( str = \{rgb, contour, texture\} \) then
         14. \( H_{str} \leftarrow (I_a^i, I_p^j, I_h^n) \)
      15. end if
   16. end for
17. Retrain DTS, and obtain \( f_{avg}(\cdot) \) (Section 4.2.3).
18. **Stage 2** Fusion Network training, focuses on hard triplets (Section 4.2.4):
   19. Randomly initialize fusion network \( f_{fusion}(\cdot) \).
20. for batch \( X_b \) in \( \{H_{rgb} \cup H_{contour} \cup H_{texture}\} \) do
      21. \( T_h = \tau(g)(X_b), T_h \) is the triple set contains hard triplets within batch \( X_b \) mined in the first stage.
      22. Obtain \( f_{MS}(\cdot) \) by plugging \( f_{avg}(\cdot), f_{rgb}(\cdot), f_{texture}(\cdot), \) and \( f_{contour}(\cdot) \) with \( f_{fusion}(\cdot) \) as shown in Fig. 1.
      23. Evaluate \( \mathcal{L}_{BH}(f_{MS}(\cdot), T_h) \) by Eq.(8).
      24. Optimize \( f_{fusion}(\cdot) \) by Adam optimizer.
   25. end for

### 4.2.3 Imprinted Text Stream

In the proposed model, imprinted text information on pills was captured by first detecting possible text regions and then recognizing the texts/symbols within them by retraining the DTS model [11] as depicted in the lower part of Figure 1. To reshape detected text regions into canonical tensor with consistent dimension, adapted bilinear sampling proposed in [11] was first employed. For a detected text region \( r \in \mathbb{R}^{w \times h \times C} \), it is normalized into a tensor with a fixed-height \( r_n \in \mathbb{R}^{w \times h' \times C} \) using:

9793
\[
    r_n = \sum_{x=1}^{w} \sum_{y=1}^{h} \max(0, 1 - |x - \tau_x(x'')|) \\
    \cdot \max(0, 1 - |y - \tau_y(y'')|),
\]

where \( \tau \) is a point-wise coordinate transformation and \( H' \) is the fixed-height that was set as 32 in [1].

We adapted the text recognizer proposed in [1] using the texture image \( I_t \) obtained after low-pass filtering as input. For each normalized \( r_n \), it was converted into a conditional probability distribution using Connectionist Temporal Classification [10], so that the most probable series of symbols could be chosen for the text regions. More specifically, the text recognizer in DTS was trained with an alphabet \( A \) to return a matrix \( M_t \) of size \( \frac{W}{H} \times |A| \) for an input \( r_n \) of size \( \frac{W}{H} \times H' \), where \( W = \frac{wH}{h} \) and \( |A| \) is the length of the alphabet. Here, each column at a position \( i \) of the matrix is a vector \( v^i = (v^i_1, \ldots, v^i_j, \ldots, v^i_{|A|}) \), where each \( v^i_j \) indicates the likelihood of the \( j \)-th label within the alphabet, e.g. letter ‘a’, exists at the \( k \)-th position, and \( \sum_{j=1}^{v^i_j} = 1 \). Then, the probability of a sequence of labels \( s \) within a detected region \( r_n \) is defined as

\[
p(s|\mathbf{v}) = \prod_{i=1}^{W/H} v^i_j, \quad s \in A^{W/H}.
\]

To remove the blanks or repeated labels, the many-to-one mapping \( M_A: A^{W/H} \rightarrow A^{W/H} \), was employed to get the conditional probability of the final sequence \( s_f \). The objective function used for training the text recognition network could be then defined as in [1, 10]:

\[
s_f \in A^{W/H} \quad \sum_{s: M_A(s) = s_f} p(s|\mathbf{v}).
\]

After retraining the network \( f_{text}(\cdot) \), for an input \( I_{RGB} \), the texture map \( I_t \) was first generated. Then \( I_t \) was fed to the text detection network to generate all the possible text regions, the two highest ranked text proposals were normalized, tallied as one \( r_n \) and then fed to the text recognition network to obtain the matrix \( M_t = f_{text}(r_n) \) in size of \( \frac{W}{H} \times |A| \). Afterwards, it was averaged over the position dimension, \( i.e. \), the width of the tallied text region, to obtain a final text probability vector \( \mathbf{v}_t \) with a size of \( (1, |A|) \), where each dimension of the vector \( \mathbf{v}_t_j = \frac{\sum_{i=1}^{|A|} v^i_j}{W/H} \) corresponds to one item in the alphabet \( A \) showing the average probability value of the likelihood of this corresponding symbol appears in the pill image over the entire tailed text regions. The procedure of taking one \( I_t \) as input to obtain the average text vector \( \mathbf{v}_t \) is denoted as \( f_{text}(M_t) = \mathbf{v}_t \). The retrained DTS is fixed and will be then combined by the later fusion net with other streams in the fusion stage. Intuitively, \( \mathbf{v}_t \) provides information of the existence of imprinted symbols on the pills, and was subsequently utilized to compensate the other three streams.

### 4.2.4 Multi-Stream Fusion Network

The second training stage is summarized in Algorithm 1 (line: 11-24). After training the \( RGB \), \( Texture \), and \( Contour \) streams separately using the BA strategy in the first training stage, where the hard samples were selected, they were then fixed. Afterwards, in the second training stage, they were combined with the Imprinted Text stream (it was fixed after retraining with texture maps, and was used as extra information) to train the fusion network \( f_{fusion}(\cdot) \) using BH strategy that focuses only on the hard samples. More specifically, they were concatenated with two fully connected layers (white color layers in Figure 1). The remaining fusion network was trained with emphasis on the hard samples collected during the first training stage (\( i.e. \), pretraining of individual streams) with the following objective function:

\[
    L_{BH}(\theta, X_b) = \sum_{str=1}^{3} N_{str} \sum_{\text{hard neg.}} \underset{\text{str}}{\text{max}} H_{str} \left( f_{str}(I^a_a), f_{str}(I^b_p) \right),
\]

where \( H_{str} \) is the set of hard samples obtained during the first training stage, \( str = \{\text{rgb}, \text{texture}, \text{contour}\} \) corresponds to the \( RGB \), \( Texture \) and \( Contour \) streams respectively. \( N_{str} \) is the number of hard samples from the stream of \( str \) in the batch. It must be emphasized that only the \( RGB \), \( Texture \) and \( Contour \) streams were trained firstly using the triplet loss with the BA strategy to mine the hard samples that could not be settled by the corresponding feature. The Imprinted Text stream was combined directly with the other streams as auxiliary information, where the fusion net was trained using the BH strategy in the second stage for the following reasons: 1) the bottom-side of a great number of pills do not contain any imprinted text; 2) imprinted texts could be occluded; 3) pills with different characteristics but from the same manufacturer could have the same imprinted text. Thus, most of the samples that could be easily represented by other information, \( e.g. \), shape, would become hard samples for stream \( Imprinted Text \), and hence weaken the advantage of using the proposed two-stage strategy.

### 4.2.5 Pill Retrieval/Recognition

With the embedded metric learned using the proposed multi-stream model \( f_{MS} \), the category of any query consumer image \( I_{con} \) could then be predicted by measuring the similarity between \( I_{con} \) and all the reference images \( I_{ref} \) in the learned embedded space. In this study, the similarity score was computed directly using \( D(f_{MS}(I_{con}), f_{MS}(I_{ref})) \) instead of firstly summing the
similarity scores computed by different networks based on different features as done in [37].

5. Experimental Results

5.1. Pill Segmentation

The $W^2$-net was trained using our CURE dataset. Reference images were utilized via data augmentation (details in supplementary material) to train the network and the performance was tested on 20% of the consumer images with pixel-wise labels. The $W^2$-net was trained for 5 epochs using Adam optimizer with a learning rate started from $10^{-4}$ (divided by 10 every 2 epochs).

Intersection Over Union (IOU) [24] was applied for performance evaluations. To check the superiority of $W^2$-net, it was compared to the original $U$-net proposed in [24], and the state-of-the-art light-weight Espnetv2 [18]. The performances are summarized in Table 2. As observed, the proposed network achieves superior performance even compared with $U$-net (35M). Examples of segmentation results on both the CURE dataset are shown in Figure 5. As shown, 1) the first row that $W^2$-net is more robust to complicated background; 2) the second row that $W^2$-net is better in dealing with the regions of pills’ shadows.

![Figure 5.](image)

Figure 5. Columns: from left to right are the pill image, the segmented results using (1) $U$-net; (2) ESPNetv2; (3) $W^2$-net.

5.2. Imprinted Text Detection & Recognition

The DTS [1] that detects and recognizes imprinted text in pill images was retrained on our novel CURE dataset (details of data augmentations are summarized in the supplementary material). We directly retrained both the text detection and recognition network of the DTS simultaneously for 6 epochs using mini-batch Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a learning rate of $10^{-3}$, divided by 10 every two epochs.

![Table 2. Performances of pill segmentation.](image)

<table>
<thead>
<tr>
<th></th>
<th>$U$-net (35M)</th>
<th>Espnetv2 (0.3M)</th>
<th>$W^2$-net (2M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOU</td>
<td>0.90</td>
<td>0.78</td>
<td>0.94</td>
</tr>
</tbody>
</table>

The performance of the imprinted text recognition model was evaluated by the f-measure [1]. To confirm our hypothesis that imprinted texts on pills could be more easily recognized using the texture map compared to the RGB image, we conducted our experiment using the DTS model on both texture maps and RGB images. Results shown in Table 3 demonstrate that the model trained with texture maps yields superior performance. Except for DTS, we have also tested the state-of-the-art end-to-end Mask TextSpotter (MTS) [16] using RGB images. However, its performance is lower compared to using DTS on RGB images.

5.3. Pill Recognition

Our MS model was tested on both the NIH and CURE datasets. In the first training stage, for the three individual streams, dropout (with $p_d = 0.9, 0.8, 0.8$ for the RGB, Texture, and Contour streams respectively, where $p_d$ is the probability of retaining the hidden unit) and $l_2$ norm regularization ($\lambda_{l_2}$ was set as 0.1, 0.06, 0.06 for the three streams respectively, where $\lambda_{l_2}$ is the regularization parameter) were used on each of the fully connected (dense) layers before the output. The mini-batch Adam was selected as the optimizer, with learning rate $l_r$ equal to $3 \times 10^{-4}$ and divided by 3 after every 5 iterations until 20 iterations are complete. During the second training stage, $p_d = 0.5$, $\lambda_{l_2} = 10^{-3}$, and the mini-batch Adam is employed with $l_r = 10^{-4}$, which is divided by 3 for every 2 epochs. For both training stages, margin $m$ was set to 0.5 and the mini-batch size was 64.

![Table 3. Performances of imprinted text recognition.](image)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f-measure</td>
<td>0.446</td>
<td>0.47</td>
<td>0.56</td>
</tr>
</tbody>
</table>

The performances of the pill recognition models were evaluated based on the Mean Average Precision (MAP), and Top-K accuracy [37, 35]. The evaluation scheme of one-side pill recognition with 5-fold cross validation [37] was used. The performances of the proposed model compared to MDP [37], the state-of-the-art pill recognition model, on the NIH and CURE datasets are listed in Table 4. As shown, the proposed MS model outperforms MDP on both of the datasets in terms of Top-1 and MAP values. It is noteworthy that, our network (including segmentation and recognition) is much lighter than the one of MDP (15.6 M vs. 39 M.), and the inference time is shorter (87ms vs. 89.8ms).

Solved hard samples: To showcase the advantage of using our proposed two-stage learning strategies, examples of hard triplets from the NIH/Cure dataset, which are mined in
the first stage (i.e., violate the triplet constraint $d_{i,a,p}^n < m$) and solved (accurately recognized) in the second stage, are shown in Figure 6. Images on the right, middle, and right of each triple are the positive samples, the anchor images, and the negative samples respectively. Samples (negative/positive) within the hard triplets are hard samples.

![Figure 6](image)

Table 5. Performance comparison with few-shot learning models (5-way 1-shot Accuracy (%)).

<table>
<thead>
<tr>
<th>Database</th>
<th>NIH</th>
<th>CURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM [14]</td>
<td>61.2</td>
<td>50.4</td>
</tr>
<tr>
<td>MTL [29]</td>
<td>58.7</td>
<td>47.7</td>
</tr>
<tr>
<td>MS (ours)</td>
<td>64.2</td>
<td>53.7</td>
</tr>
</tbody>
</table>

### Few-shot/meta learning regime

We also compared our model with state-of-the-art meta/few-shot learning models, i.e., MTL [29] and CTM [14]. Similar to the experimental setup on MiniImagenet dataset in [7, 32], we followed the protocol proposed in [32] and divided the NIH and our CURE dataset into 16% , 64% and 20% as meta-validation (NIH:160, CURE:31 classes), meta-train (NIH:640, CURE:125 classes), and meta-test (NIH:200, CURE:40 classes) sets according to the pills’ categories. Categories in test set are unseen during training/validation process. During the meta-training phase, as done in [14], the entire meta-train set was employed to train the similarity metric/embedder with the proposed multi-stream network boosted by the two-stage learning strategy. During the meta-testing phase, to set up an $N$-way $K$-shots recognition evaluation scheme, $N$ unseen classes were selected, provide the model with K different instances of each of the $N$ classes, and evaluate the model’s ability to classify new instances within the $N$ classes [31, 7]. For fair comparison, only the segmented images obtained using $W^2$-net are considered for the compared few-shot models. For MTL model, the ResNet-12 architecture was adopted as in their paper and in [8, 20]. For CTM, as their model achieved better performance by employing deeper backbone for the feature extractors [14], we tested CTM with ResNet-18 in the experiment. Please refer to the supplementary material for more details on the experimental set up. The results are presented in Table 5. As shown, our model achieves the best few-shot classification performance.

### Ablation Study

Extensive comparisons with ablative models were performed, and the results are presented in Table 6. By comparing the performances of the ablative models to the one of the MS model shown in Table 4, it could be seen that: 1) Impact of each stream (row 3-8): the proposed MS model outperforms the individual models. By removing a certain stream, the performance drops. Domain-related information, e.g., imprinted text, helps to improve the recognition performance; 2) Impact of segmentation models (row 9-10): by removing/replacing the proposed $W^2$-net, the performances drop; 3) Impact of learning batch strategy (row 10-12): the proposed two-stage BA-BH learning strategy is superior to the traditional BH and BA strategy.

Table 6. Recognition results for ablative models (one-side).

<table>
<thead>
<tr>
<th>Database</th>
<th>Abitative models</th>
<th>MAP</th>
<th>TOP-1</th>
<th>MAP</th>
<th>TOP-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH</td>
<td>MTL</td>
<td>0.612</td>
<td>54.6</td>
<td>0.562</td>
<td>50.9</td>
</tr>
<tr>
<td>CURE</td>
<td>Stream RGB</td>
<td>0.259</td>
<td>20.6</td>
<td>0.507</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>Stream Contour</td>
<td>0.179</td>
<td>12.6</td>
<td>0.534</td>
<td>25.98</td>
</tr>
<tr>
<td></td>
<td>No segmentation</td>
<td>0.406</td>
<td>45.6</td>
<td>0.447</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>With Bh</td>
<td>0.664</td>
<td>60.2</td>
<td>0.682</td>
<td>65.1</td>
</tr>
<tr>
<td></td>
<td>With Bh</td>
<td>0.651</td>
<td>58.7</td>
<td>0.677</td>
<td>64.5</td>
</tr>
</tbody>
</table>

### 6. Conclusion

In this study, we present a new pill images dataset CURE, which provides more instances per class. For better tackling the few-shot pill recognition problem, a $W^2$-net is first proposed for pill segmentation. Then, a multi-stream deep architecture along with a two-stage learning strategy is proposed to better exploit the domain-related information in small data scenarios. It deploys first the BA strategy for the RGB, Texture, Contour streams to mine the hard samples, and second a novel BH strategy to train a fusion-net that combines the three individual streams with a stream of imprinted text as auxiliary information. Experimental results show that 1) $W^2$-net is superior to both $U$-net and ESPNev2; 2) using high-frequency components with emphasized texture helps to solve the formidable problem of recognizing imprinted text on pills; 3) The proposed model achieves top performance by its more accurate recognition of the hard samples that cannot be handled by individual features.

### Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT Future Planning (NRF-018X1A3A1070163). Jieun Kim is the corresponding author. The authors would like to thank Xiao Zeng for evaluating their MDP model on the CURE dataset.
References


