Few-Shot Pill Recognition

Suiyi Ling[1]*, Andréas Pastor[1]*, Jing Li[2], Zhaohui Che[3], Junle Wang[4], Jieun Kim[5], Patrick Le Callet[1]

[1]University of Nantes [2]Alibaba Group [3]Shanghai Jiao Tong University [4]Tencent [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^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],

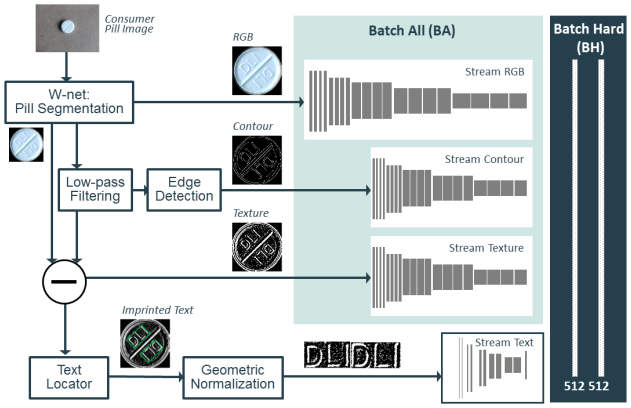


Figure 1. The framework of the proposed pill recognition model that 1) takes color, texture, contour, and imprinted text information as input; 2) uses first the *Batch ALL* (BA) and then the *Batch Hard* (BH) strategy. $\ominus$ denotes the subtraction operation.

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,

*Equal contribution.

(a) (b) (c)

Figure 2. Examples of *hard samples* in pill recognition: (a) hard positive pill samples from *CURE* dataset; (b) hard negative pill samples from *NIH* dataset; (c) texture maps obtained using low-pass filtering of the hard negative samples in (b).

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 TextSpotter* (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.

	NIH NLM	**CURE**
Number of pill images	7000	8973
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, *MobileDeepPill* (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, gray-level 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.

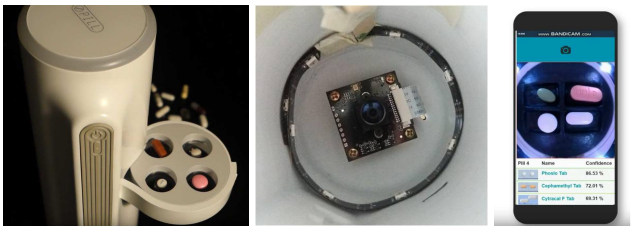
Although some of the aforementioned models consider the imprinted text on pills, they solely use structural de-

scriptors 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 straightaway 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 [1]

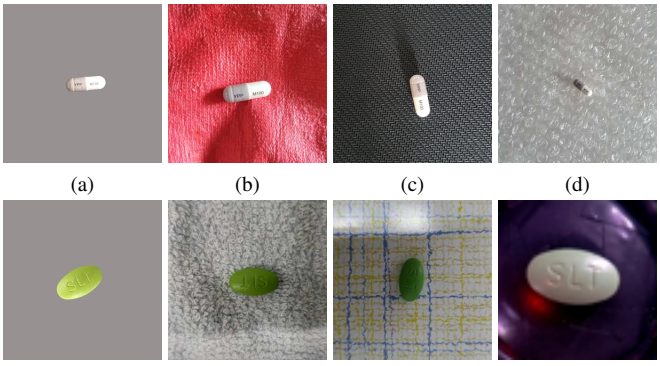
In this section, the novel *CURE* pill dataset is introduced. This dataset summarized in Table 1, contains 8973 images of 196 categories, and approximately 45 samples were obtained for each pill category.



(a) (b) (c)

Figure 3. (a) *MPI* device; (b) embedded camera in the device; (c) smartphone software connected to the device.

Equipment: The pill images within this database were taken using three phones (*i.e.*, *Samsung SM-J320FN*, *SM-N920S* and *LG F500L*) and one *Multi-Pill Identifier* (MPI) device. The MPI device provides healthcare personnel and the general public with descriptions for unknown pills. Information obtained using this device could be used for purposes such as checking the compatibility between different pills and detecting expired medicines. The device is shown in Fig 3 (a), where a camera was mounted on a *Raspberry Pi 3*, as shown in Fig 3 (b), and was set above the pill holder.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4. Examples of images in CURE. Row: each row corresponds to one category of the pill. Column: (1) 1st column: reference images; (2) other columns: consumer images.

Consumer Images: are pictures not taken under professionally controlled conditions [35]. In real cases, consumer images are likely to be taken with varying backgrounds, illumination, focus, and orientations. To make the database more diverse, when collecting the consumer images, backgrounds of different levels of texture granularity, illumination, and dynamic zooming in/out conditions are considered. The illumination conditions include 1) indoor light, 2) weak outdoor light, and 3) strong outdoor light.

Reference Images: For each pill category, the reference image was generated using the best-quality consumer images. More specifically, the pixel-level pill regions in the selected pill images with the better-controlled conditions are first manually labeled. Then, the backgrounds of the selected images are replaced with clean gray backgrounds. 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-

[1] Our CURE dataset is available at `https://github.com/suiyiling/Few-shot-pill-recognition`.

nificant 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:

$$\mathcal{L}_{U_i} = \sum_{p \in P_I} -l(p) \cdot log(s(p)) + (1 - l(p)) \cdot log(1 - s(p)), \tag{1}$$

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:

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

where λ_{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 θ 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]:

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

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, $D(f_\theta(I_i), f_\theta(I_j)) : \mathbb{R}^E \times \mathbb{R}^E \rightarrow \mathbb{R}$ denotes the metric function that measures the distances between two images I_i, I_j in the embedding space. Throughout this study, Euclidean distance was used as the distance measure, *i.e.,*$D(\cdot)$, 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_θ 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 *Coutour*, 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]:

$$\mathcal{L}_{BA}(\theta, X_b) = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} \overbrace{\sum_{\substack{i=1\\p\neq a}}^{K}}^{all\ pos.} \overbrace{\sum_{j=1}^{P}\sum_{\substack{n=1\\j\neq i}}^{K}}^{all\ neg.} \left[m + d^{i,a,p}_{j,a,n}\right]_+ ,$$
$$d^{i,a,p}_{j,a,n} = D(f_\theta(I^i_a), f_\theta(I^i_p)) - D(f_\theta(I^i_a), f_\theta(I^j_n)), \tag{4}$$

where I^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 $PK(PK-K)(K-1)$ combination of triplets were considered. During the first training stage, the hard samples that violated the constraint $d^{i,a,p}_{j,a,n} < m$ 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 $\tau_g(\cdot)$.
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 $(I^i_a, I^i_p, I^j_n) \in T_g(X)$ **do**
13: **if** $D(f_{str}(I^i_a), f_{str}(I^i_p)) - D(f_{str}(I^i_a) f_{str}(I^j_n)) < m$, where $str = \{rgb, contour, texture\}$ **then**
14: $H_{str} \leftarrow (I^i_a, I^i_p, I^j_n)$
15: **end if**
16: **end for**
17: Retrain DTS, and obtain $f^{avg}_{text}(\cdot)$ (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 triplet set contains hard triplets within batch X_b minded in the first stage.
22: Obtain $f_{MS}(\cdot)$ by plugging $f^{avg}_{text}(\cdot)$, $f_{rgb}(\cdot)$, $f_{texture}(\cdot)$, 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 [1] 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 [1] 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 $r_n \in \mathbb{R}^{\frac{wH'}{h} \times H' \times C}$ using:

$$r_n = \sum_{x=1}^{w}\sum_{y=1}^{h} max(0, 1-|x-\tau_x(x')|) \cdot max(0, 1-|y-\tau_y(y')|)), \tag{5}$$

where τ 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 $\mathcal{A}$ to return a matrix $\mathbf{M_t}$ of size $\frac{\overline{W}}{4} \times |\mathcal{A}|$ for an input r_n of size $\overline{W} \times H'$, where $\overline{W} = \frac{wH'}{h}$ and $|\mathcal{A}|$ is the length of the alphabet. Here, each column at a position i of the matrix is a vector $\mathbf{v}^i = (v_1^i, \ldots, v_j^i, \ldots, v_{|\mathcal{A}|}^i)$, where each v_j^i indicates the likelihood of the j_{th} label within the alphabet, *e.g.* letter 'a', exists at the i_{th} position, and $\sum_{j=1}^{\mathcal{A}} v_j^i = 1$. Then, the probability of a sequence of labels s within a detected region r_n is defined as

$$p(s|\mathbf{v}) = \prod_{i=1}^{\overline{W}/4} v_j^i, \ s \in \mathcal{A}^{\overline{W}/4}. \tag{6}$$

To remove the blanks or repeated labels, the many-to-one mapping $\mathcal{M}_{\mathcal{A}} : \mathcal{A}^{\overline{W}/4} \mapsto \mathcal{A}^{\leq \overline{W}/4}$, 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]:

$$\sum_{s_f \in \mathcal{A}^{\leq \overline{W}/4}} \ \sum_{s:\mathcal{M}_{\mathcal{A}}(\mathbf{v})=s_f} p(s|\mathbf{v}). \tag{7}$$

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, tailed as one r_n and then fed to the text recognition network to obtain the matrix $\mathbf{M_t} = f_{text}(r_n)$ in size of $\frac{\overline{W}}{4} \times |\mathcal{A}|$. Afterwards, it was averaged over the position dimension, *i.e.,* the width of the tailed text region, to obtain a final text probability vector $\mathbf{v_t}$ with a size of $(1, |\mathcal{A}|)$, where each dimension of the vector $\mathbf{v_{t_j}} = \frac{\sum_{i=1}^{\overline{W}/4} v_j^i}{\overline{W}/4}$ corresponds to one item in the alphabet $\mathcal{A}$ 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 $f_{text}^{avg}(\mathbf{M_t}) = \mathbf{v_t}$. 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:

$$\mathcal{L}_{BH}(\theta, X_b) = \sum_{str=1}^{3} \sum_{\substack{(I_a^i, I_p^i, I_n^j) \in H_{str} \\ (I_a^i, I_p^i, I_n^j) \in X_b}}^{N_{str}} [m + \overbrace{D(f_\theta(I_a^i), f_\theta(I_p^i))}^{str_{th}\ hard\ pos.} - \underbrace{D(f_\theta(I_a^i), f_\theta(I_n^j))}_{str_{th}\ hard\ neg.}]_+, \tag{8}$$

where H_{str} is the set of hard samples obtained during the first training stage, $str = \{rgb, texture, contour\}$ 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

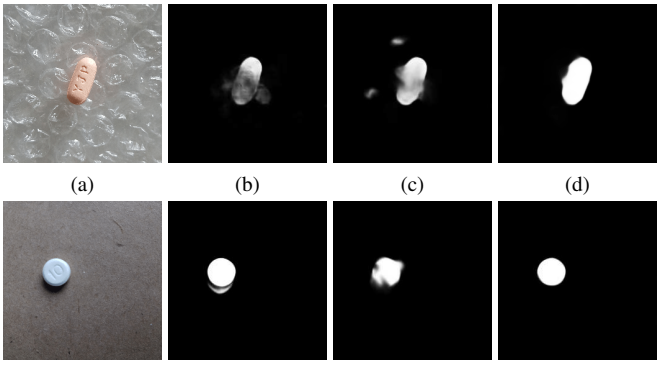
5.1. Pill Segmentation

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).

Table 2. Performances of pill segmentation.

	U-net (35M)	**Espnetv2** (0.3M)	$\mathbf{W^2}$-net (2M)
IOU	0.90	0.78	**0.94**

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.



(a) (b) (c) (d)

(e) (f) (g) (h)

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 3. Performances of imprinted text recognition.

	MTS-RGB [16]	***DTS-RGB*** [1]	***DTS-Texture***
f-measure	0.446	0.47	**0.56**

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 (λ_{l_2} was set as 0.1, 0.06, 0.06 for the three streams respectively, where λ_{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 3×10^{-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 4. Performance of pill recognition models (**one-side**).

Database	**NIH**		**CURE**	
	MAP	**TOP-1**	**MAP**	**TOP-1**
MDP [37]	0.582	53.1	0.704	63.7
MS (ours)	0.722	65.9	0.749	68.3

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 (57ms 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 $d^{i,a,p}_{j,a,n} < m$) 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.

(a) (b) (c) (d)

Figure 6. Examples of hard triplets that are minded in the first stage but are solved (accurately recognized) in the second stage.

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

Database	NIH	CURE
CTM [14]	61.2	50.4
MTL [29]	58.7	47.7
MS (ours)	**64.2**	**53.7**

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).

Database		NIH		CURE	
	Ablative models	**MAP**	**TOP-1**	**MAP**	**TOP-1**
Individual stream	Stream RGB	0.612	54.6	0.562	50.9
	Stream Texture	0.259	20.6	0.507	49.2
	Stream Contour	0.179	12.6	0.348	25.98
Impact of domain-related features	Without Text	0.612	54.4	0.594	52.2
	Without Contour	0.653	60.9	0.677	66.9
	Without Texture	0.633	56.9	0.604	54.5
Impact of segmentation	No segmentation	0.406	45.6	0.447	48.7
	With U-net (35M)	0.577	54.1	0.641	60.4
Impact of strategy	With BA	0.664	60.2	0.682	65.1
	With BH	0.651	58.7	0.677	64.5

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 *ESPNetv2*; 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

[1] Michal Busta, Lukas Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.

[2] Jesus J Caban, Adrian Rosebrock, and Terry S Yoo. Automatic identification of prescription drugs using shape distribution models. In *2012 19th IEEE International Conference on Image Processing*, pages 1005–1008. IEEE, 2012.

[3] Rung-Ching Chen, Yung-Kuan Chan, Ying-Hao Chen, and Cho-Tsan Bau. An automatic drug image identification system based on multiple image features and dynamic weights. *International Journal of Innovative Computing, Information and Control*, 8(5):2995–3013, 2012.

[4] Joyce A Cramer, Anuja Roy, Anita Burrell, Carol J Fairchild, Mahesh J Fuldeore, Daniel A Ollendorf, and Peter K Wong. Medication compliance and persistence: terminology and definitions. *Value in health*, 11(1):44–47, 2008.

[5] António Cunha, Telmo Adão, and Paula Trigueiros. Helpmepills: A mobile pill recognition tool for elderly persons. *Procedia Technology*, 16:1523–1532, 2014.

[6] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[8] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *arXiv preprint arXiv:1806.04910*, 2018.

[9] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Young-Beom Lee, Unsang Park, Anil K Jain, and Seong-Whan Lee. Pill-id: Matching and retrieval of drug pill images. *Pattern Recognition Letters*, 33(7):904–910, 2012.

[14] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2019.

[15] Suiyi Ling, Patrick Le Callet, and Zitong Yu. The role of structure and textural information in image utility and quality assessment tasks. *Electronic Imaging*, 2018(14):1–13, 2018.

[16] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

[17] Ingo Lütkebohle. NIH NLM Pill Image Recognition Challemge. `https://pir.nlm.nih.gov/challenge/`, 2016.

[18] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9190–9200, 2019.

[19] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.

[20] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. *arXiv preprint arXiv:1712.09926*, 2017.

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[22] Sae Ochi, Susan Hodgson, Owen Landeg, Lidia Mayner, and Virginia Murray. Disaster-driven evacuation and medication loss: a systematic literature review. *PLoS currents*, 6, 2014.

[23] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[25] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[28] Henry A Spiller and Jill RK Griffith. Increasing burden of pill identification requests to us poison centers. *Clinical Toxicology*, 47(3):253–255, 2009.

[29] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[30] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[33] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[34] Darrell West et al. How mobile devices are transforming healthcare. *Issues in technology innovation*, 18(1):1–11, 2012.

[35] Ziv Yaniv, Jessica Faruque, Sally Howe, Kathel Dunn, David Sharlip, Andrew Bond, Pablo Perillan, Olivier Bodenreider, Michael J Ackerman, and Terry S Yoo. The national library of medicine pill image recognition challenge: An initial report. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2016.

[36] Jiye Yu, Zhiyuan Chen, Sei-ichiro Kamata, and Jie Yang. Accurate system for automatic pill recognition using imprint information. *IET Image Processing*, 9(12):1039–1047, 2015.

[37] Xiao Zeng, Kai Cao, and Mi Zhang. Mobiledeeppill: A small-footprint mobile deep learning system for recognizing unconstrained pill images. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 56–67. ACM, 2017.