Supplementary Material

1. Introduction

In this supplementary material, we first present the network architectures of the proposed segmentation W^2-Net, and the *RGB*, *Texture* and *Contour* streams of the proposed multi-streams network. In addition, more examples of the pill images in our proposed *CURE* dataset are shown. Afterwards, data augmentation strategies used in our study are summarized. Moreover, more details of the experimental setup for few-shot regime is provided. Furthermore, we provide the information of the performance evaluation metrics used in our study. More importantly, more experimental results are presented including 1) more pill segmentation results; 2) the two-side pill recognition results; 3) more ablation studies. Finally, notations of all the variables and functions are summarized in the last section.

2. Network architecture of the proposed W^2-Net, and the RGB, Contour and Texture streams

The architecture of the proposed W^2-Net is shown in Fig. 1. It is consisted of four simplified U-nets: the input of each intermediate U-net is the concatenation of (1) the segmentation output from previous U-net, (2) the output of the second last layer from the previous U-net and (3) the input image.

The architectures of RGB stream is presented in Table 1, when the ones of Contour and Texture streams are summarized in Table 2. The *Texture* and *Contour* streams share the same architecture. The only difference between their architecture and the one of the RGB stream is the number of channels (*i.e.*, the network of the RGB stream has twice the numbers of channels) at each layer. The number of parameters of *RGB*, *Contour*, *Texture* streams are 9 M, 2.2 M and 2.2 M respectively. More complex architectures were tested for the three streams, however, performances drop. Please refer to Section 'Extra ablation studies' in this file for more details.

Table 1. The CNN architecture of the*RGB* stream. **nb C.** denotes the number of channels.

Type	**nb C.**	**output size**	**stride**	**kernel**
Input	3	128×128	-	-
conv1	64	64×64	2	7×7
conv2	64	64×64	1	3×3
conv3	64	64×64	1	3×3
pool1	64	32×32	-	4×4
conv4	96	32×32	1	3×3
conv5	96	32×32	1	3×3
conv6	96	32×32	1	3×3
pool2	96	16×16	-	4×4
conv7	128	16×16	1	3×3
conv8	128	16×16	1	3×3
conv9	128	16×16	1	3×3
pool3	128	8×8	-	4×4
conv10	256	8×8	1	3×3
conv11	256	8×8	1	3×3
conv12	256	8×8	1	3×3
pool4	$256/128$	4×4	-	4×4
conv13	384	4×4	1	3×3
conv14	384	4×4	1	3×3
conv15	384	4×4	1	3×3
fc	-	512	-	-
reg	-	256	-	-

Table 2. The CNN architecture of the *Texture* and *Contour* streams. **nb C.** denotes the number of channels.

Type	**nb C.**	**output size**	**stride**	**kernel**
Input	1	128×128	-	-
conv1	32	64×64	2	7×7
conv2	32	64×64	1	3×3
conv3	32	64×64	1	3×3
pool1	32	32×32	-	4×4
conv4	48	32×32	1	3×3
conv5	48	32×32	1	3×3
conv6	48	32×32	1	3×3
pool2	48	16×16	-	4×4
conv7	64	16×16	1	3×3
conv8	64	16×16	1	3×3
conv9	64	16×16	1	3×3
pool3	64	8×8	-	4×4
conv10	128	8×8	1	3×3
conv11	128	8×8	1	3×3
conv12	128	8×8	1	3×3
pool4	128	4×4	-	4×4
conv13	192	4×4	1	3×3
conv14	192	4×4	1	3×3
conv15	192	4×4	1	3×3
fc	-	256	-	-
reg	-	128	-	-

3. More examples from the proposed CURE pill image dataset

More examples of pill images in our CURE dataset are presented in Figure 2. It could be observed that diverse lighting, zooming, backgrounds conditions are considered.

4. Details of data augmentations strategies used in the experiments

4.1. Pill segmentation

The W^2-net was trained using our *CURE* dataset. Reference images were utilized to train the network with data

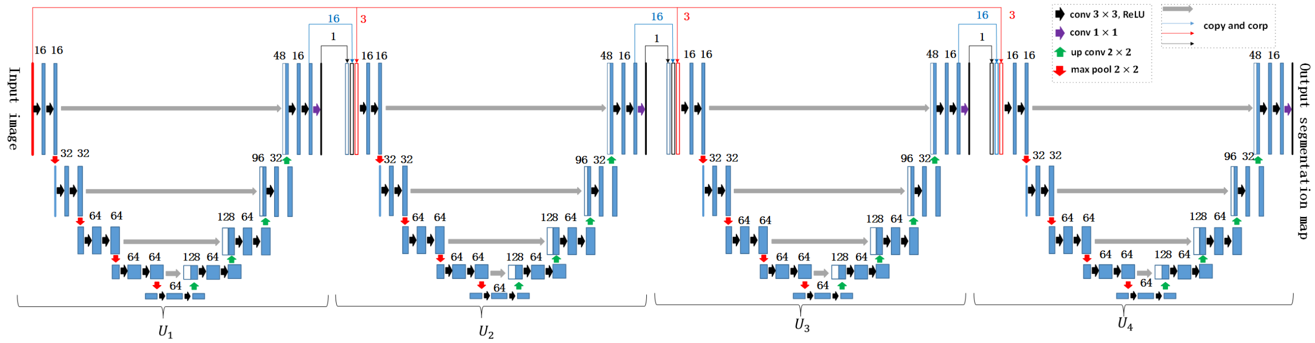


Figure 1. The W^2-net architecture.

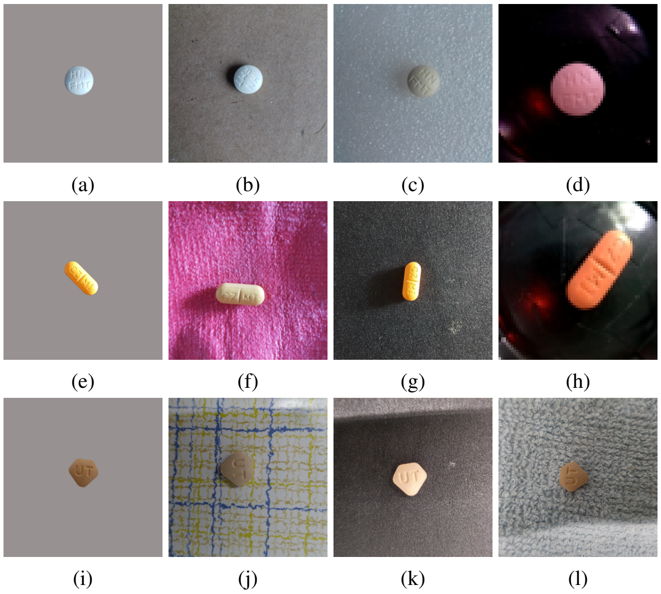


Figure 2. Examples of images in CURE. Row: each row corresponds to one category of the pill. Column: (1) 1st column: reference images; (2) other columns: consumer images.

augmentation and the performance was tested on 20 % of the consumer images with pixel-wise labels. To augment the training set, we synthesized around 10^5 customer images by 1) replacing the background using the texture images from the *Describable Textures Dataset* [2] or background patches manually extracted from the *CURE* dataset; 2) rotating the labeled pill region in the range of $(-180°, 180°)$; 3) changing foreground-background contrast calculated in terms of the ratio of foreground-background illuminance; 4) randomly changing the ratio of the height/width of the maximum circumscribed rectangle of the foreground pill area versus the height/width of the image to mimic different zoom in/out conditions. As thus, the network is robust to the variance of pill sizes; 5) randomly switching the pill location.

4.2. Stream Imprinted Text Pre-training

We first used the text regions proposal model of the Deep TextSpotter (DTS) to generate possible text regions and manually selected the correct text regions as ground-truth text regions. Since imprinted texts are challenging to be detected and recognized by DTS, for images where no text region candidate was returned, we manually labeled the bounding boxes for all the text regions. For those pill images that contain no texts, the largest inscribed rectangular of the pill mask obtained after pill segmentation was taken as the ground-truth regions, and labeled with a blank symbol '-'. As the imprinted text/symbol regions were labeled, we augmented the data by (1) blurring the text regions using *Gaussian* filtering; (2) rotating the text boxes in the range of $(-180°, 180°)$ with a constraint that the rotated bounding boxes should still locate within the pill.

4.3. Pill Recognition

To augment the data, we simply rotated the images in the range of $(-180°, 180°)$ for both the NIH and CURE datasets, *i.e.*, we employed the same data augmentation strategy on the two datasets separately.

5. Experimental setup for few-shot regime

To compare with the state-of-the-art few-shot learning models, we followed the experimental protocol proposed by [6]. Similar to the experimental setup on MiniImagenet dataset in [3], we divided the NIH and our CURE dataset into 16% , 64% and 20% as validation, training and testing set according to the pills' categories. Categories in test set are unseen during training/validation process. More specifically, to set up an $N-$way $K-$shots classification/recognition problem, 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.

For some pill categories in the NIH dataset, there are less than 5 instances for top/bottom side of the pill. Thus, in the paper, we only tested the models under the 1-shot setting.

6. Details of performance evaluation metrics used in the experiments

6.1. Pill segmentation

- **Intersection Over Union (IOU):**

$$IOU = \frac{p_{ii}}{\sum_{j=1}^{2} p_{ij} + \sum_{j=0}^{2} (p_{ji} - p_{ii})}, \qquad (1)$$

where p_{ii} indicates the true positives, while p_{ij} and p_{ji} are the false positives and false negatives relatively.

6.2. Imprinted text recognition

As done in [1], when the IOU of a predicted text region is higher than 0.5 and the transcription is identical (using case-insensitive comparison [5]), it is considered as correctly recognized. Performance is then evaluated with f-measure.

- **f-measure:**

$$f - measure = \frac{2 \times precision \times recall}{precision + recall}, \qquad (2)$$

where $precision = \frac{t_p}{t_p + f_p}$, and $recall = \frac{t_p}{t_p + f_n}$. t_p, f_p, and f_n indicates the true positives, the false positives and the false negatives correspondingly.

6.3. Pill recognition

- **Mean Average Precision (MAP):**

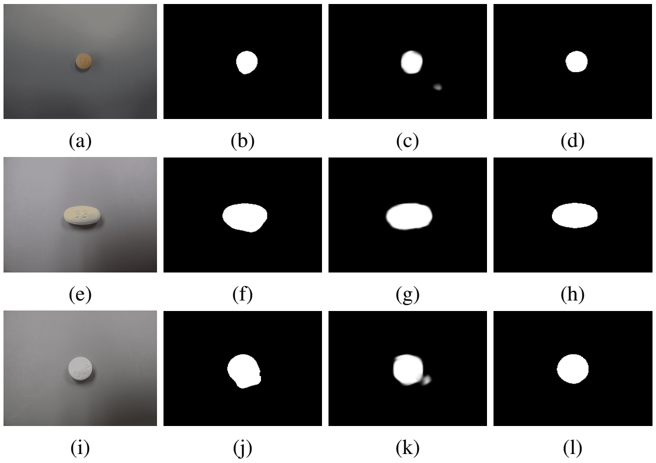
$$MAP = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \frac{j}{MT(i,j)} \right), \qquad (3)$$

  where N denotes the number of consumer images, N_i represents the number of the reference images, j is the number of correctly matched images (*i.e.*, 1/2), and $MT(i,j)$ indicates the correct ranking of the reference images.

7. More experimental results

7.1. More pill segmentation results

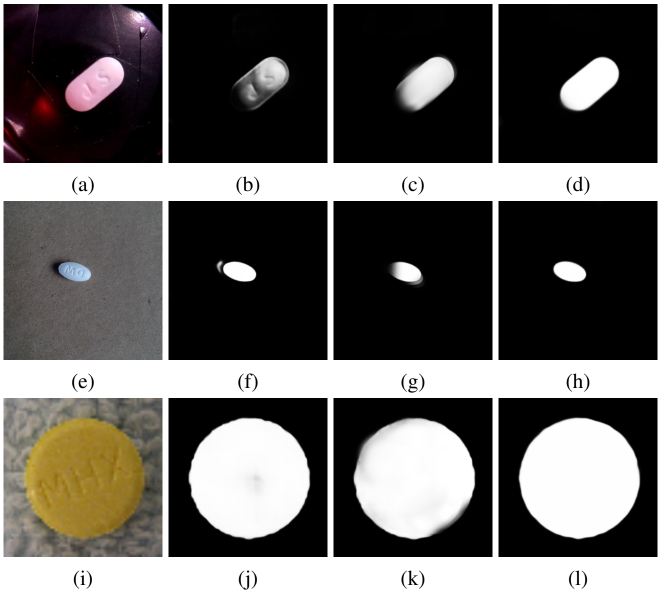
More pill segmentation results on *NIH* and *CURE* dataset are shown in Figure 3 and 4 correspondingly. Columns in the figures from left to right are the pill image, the segmented results using U-net, ESPNetV2, and W^2-net respectively. It could be observed from Figure 3 (d-l) that W^2-net is better in preserving the shape of the pills (smoother borders). It is verified that even the proposed model are trained on our *CURE*, it could also be employed on other pill image dataset.


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

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

(i) (j) (k) (l)

Figure 3. Examples of segmentation results on *NIH*. U-Net, ESPNetV2, W2


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

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

(i) (j) (k) (l)

Figure 4. Examples of segmentation results on *CURE*. U-Net, ESPNetV2, W2

7.2. Two-side performance

The performance comparison results in terms of two-side MAP, Top-1 are summarized in Table 3. As shown, in the two-side case, the proposed model is also superior to the state-of-the-art model MDP. Similar to [7], under two-side setting, if one pill recognizer is able to indicate the category of the query pill image no matter which side it is (top/bottom), then the image is considered as accurately recognized. However, under one-side setting, top and bottom side of one pill category are considered as different categories [8]. Therefore, the two-side setting is commonly easier than the one-side setting, and the pill recognizers could achieve higher performance under the two-side setting.

The two-side performances of the ablative models (corresponds to Table 6 in the paper) are shown in Table 4. Consistent conclusions could be made: 1) Impact of individual stream (row 2-7): the proposed MS model outperforms the individual models. By removing a certain stream, the performance drops. Domain knowledge, *e.g.*, imprinted text, helps to improve the recognition performance; 2) Impact of segmentation models(row 7-9): by removing/replacing the proposed W^2-net, the performances drop; 3) Impact of learning batch strategy (row 8-10): the proposed two-stage BA-BH learning strategy is superior to the traditional BH and BA strategy.

Table 3. Performance of pill recognition models (**two-side**).

Database	NIH		CURE	
	MAP	**TOP-1**	**MAP**	**TOP-1**
MDP [8]	0.837	74.1	0.853	79.8
MS (ours)	**0.852**	**77.3**	**0.876**	**84.6**

Table 4. Recognition results for ablative models (two-side).

Database		NIH		CURE	
	Ablative models	**MAP**	**TOP-1**	**MAP**	**TOP-1**
Individual stream	Stream RGB	0.738	62.6	0.786	66.6
	Stream Texture	0.475	34.2	0.641	49.5
	Stream Contour	0.469	32.4	0.620	47.9
Impact of domain-related features	Without Text	0.727	64.7	0.784	70.1
	Without Contour	0.801	71.5	0.842	78.3
	Without Texture	0.788	69.9	0.830	76.7
Impact of segmentation	No segmentation	0.478	43.7	0.513	47.1
	With U-net	0.828	74.3	0.857	78.1
Impact of strategy	With BA	0.793	71.0	0.827	74.2
	With BH	0.789	70.4	0.809	73.8

7.3. Extra ablation studies:

With enough samples and network parameters (network with deep/complicated enough architecture), important domain-related information, such as texture, contour characteristics or imprinted text could be extracted and applicable for the task. But for this few-shot learning task, we found it much more efficient to take advantage of handcrafted channels as domain knowledge at the first stage with simpler networks, and then boost at the second stage.

For comparisons, we have added extra experiments on the CURE dataset: 1) single complicated stream RGB 50% more parameters (see row 1 in Table 5); 2) same architecture (as the proposed stream RGB) for stream *Contour* and *Texture* (9 M) are checked (see row 2&3 in Table 5). It could be observed that using more complicated network in a few-shot regime does not guarantee performance gains.

7.4. Impact of margin m

Similar to [4], we tested m (in equation (4) in the paper) in the range of $[0, 1]$. The one-side MAP values of the proposed MS model with varying m on our CURE and the NIH dataset are depecicted in Figure 5. As shown, the performance of the proposed model on the CURE dataset increases steadily until it achieves highest MAP with $m = 0.5$, and then drops significantly after. Similar trend could be observed for the performances of our model on the NIH dataset. Therefore, we selected $m = 0.5$ in our study.

Table 5. Extra ablation study on the *CURE* and NIH database.

Database	NIH		CURE	
Model	**MAP**	**TOP-1**	**MAP**	**TOP-1**
Stream RGB (18.1 M)	0.573	42.0	0.482	48.0
Stream Contour (9 M)	0.172	8.8	0.281	14.9
Stream Texture (9 M)	0.577	43.8	0.241	19.5

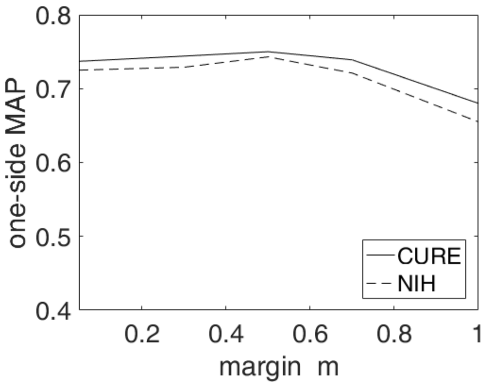


Figure 5. The impact of margin m on the performance.

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] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[3] 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.

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

[5] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[6] 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.

[7] 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.

[8] 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.