Word and Image Embeddings in Pill Recognition

Richárd Rádli[a], Zsolt Vörösházi[b] and László Czúni[c]
University of Pannonia, 8200 Veszprém, Egyetem u. 10., Hungary

Keywords: Metrics Learning, Pill Recognition, Multi-Modal Learning, Multihead Attention, Multi-Stream Network, Dynamic Margin Triplet Loss.

Abstract: Pill recognition is a key task in healthcare and has a wide range of applications. In this study, we are addressing the challenge to improve the accuracy of pill recognition in a metrics learning framework. A multi-stream visual feature extraction and processing architecture, with multi-head attention layers, is used to estimate the similarity of pills. We are introducing an essential enhancement to the triplet loss function to leverage word embeddings for the injection of textual pill similarity into the visual model. This improvement refines the visual embedding on a finer scale than conventional triplet loss models resulting in higher accuracy of the visual model. Experiments and evaluations are made on a new pill dataset, freely available.

1 INTRODUCTION

Accurate recognition of prescription pill images, based on their visual characteristics, can play a crucial role in ensuring patient safety and optimizing modern healthcare systems, particularly for elderly patients. Since medication errors are the most common mistakes in healthcare (Cronenwett et al., 2007) recognition technology has the potential to prevent errors throughout the pharmaceutical supply chain, enhance the expertise of poison control professionals, improve medication adherence, mitigate losses of medications and prescriptions during evacuation scenarios, and drive advancements in remote and self-diagnosis technologies as well as smart healthcare applications. Pill recognition systems can significantly enhance the quality of medication dispensing, either for home usage or in large-scale automated pill dispensing systems.

Strictly theoretically, medication pills are designed with distinctive features, including size, colour, shape, engravings, and imprints. However, various factors contribute to the challenges of accurate recognition, such as:

- Pill photographs are captured under diverse conditions, such as varying illumination, viewing angles, distances, and camera settings.
- Due to their small size, local features on pills are often not clearly visible or may be distorted.
- Number of possible classes can be very large (tens of thousands of possible pills), while few-shot learning is a requirement for many applications.

The two main approaches for optical recognition or verification systems are object classification and metric learning (Yang and Jin, 2006). In both cases characteristics such as imprinted or carved signs, size, color, and shape are considered as observed by the camera sensors during training and inference processes. Both of these methods apply 'yes' or 'no' evaluations (meaning that a pill belongs to the same or another class) in a sense that no scaled similarity is considered between the different classes. Metric learning aims to measure the similarity among objects while using a distance metric function. In the procedure we have elaborated, we also take into account textual information from pill information leaflets that was previously ignored

In our models we deploy multi-stream metrics learning since this way we can control the usage of different image features in the model and also few-shot learning is easy to carry out.

The main contributions of our paper are the followings:

- We showed that high-level textual information, gained from free-text, can be utilized in optical pill recognition.
- We introduced a new triplet loss (dynamic margin

[a] https://orcid.org/0009-0009-3160-1275
[b] https://orcid.org/0009-0004-3032-8784
[c] https://orcid.org/0000-0001-7667-9513

Rádli, R., Vörösházi, Z. and Czúni, L.
Word and Image Embeddings in Pill Recognition.
DOI: 10.5220/0012460800003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 729-736
ISBN: 978-989-758-679-8; ISSN: 2184-4321

triplet loss - DMTL), where the margin is dynamically controlled with distances of word embeddings of class textual descriptors;

- We created a new dataset (OGYEIv2) of 112 classes of pill with 4480 images.[1]
- We evaluated our model on our dataset with 5-fold validation.

2 RELATED WORKS

Generic deep neural network (DNN) object detectors have been applied for pill recognition in several recent articles, such as (Tan et al., 2021), (Nguyen et al., 2022), and (Heo et al., 2023).

In (Tan et al., 2021) three well-known DNN object detectors (YOLOv3, RetinaNet, and SSD) were compared on a custom dataset, resulting only in small differences in mAP ($\sim 2\%$, all above 0.80). More specific approaches are described in (Nguyen et al., 2022) and (Heo et al., 2023). In (Nguyen et al., 2022) the proposed solution used a prescription-based knowledge graph, representing the relationship between pills. A graph embedding network extracted the relational features of pills and a framework was applied to fuse the graph-based relational information with the image-based visual features for the final classification. The drawback of this method is that it requires medical prescriptions, or equivalently it can be applied when there are multiple pills on the image. In (Heo et al., 2023) the authors trained not only RGB images of the pills but also imprinted characters. In the pill recognition step, the different modules separately recognize both the features of pills and their imprints, meanwhile correcting the recognized imprint to fit the actual data of other features. A trained language model was also applied to the imprint correction. It was shown through an ablation study that the language model could significantly improve the pill identification ability of the system. The drawback of this solution is that a specific language model (including an OCR - optical character recognition module) is required for the application.

In contrast to these approaches, our solution avoids the use of specific language models and uses only generic models to process the information leaflet of pills. In the above models the training would require the processing of textual printed information and/or language-specific OCR modules. They face problems when texts are not visible (see Fig.1 for illustration) or when new classes are to be added to the model, also these texts should be added manually. Our primary purpose is to elaborate a more general and easily extensible framework.

For the above reasons, we followed the approaches (Zeng et al., 2017) and (Ling et al., 2020) where the utilization of metrics learning was demonstrated in order to embed the pill images.

The winner (Zeng et al., 2017) of an algorithm challenge on pill recognition in 2016, announced by the United States National Library of Medicine (NLM-NIH) (Yaniv et al., 2016), used a multi-stream technique. In (Zeng et al., 2017) the visual information (e.g. colour, gray-scale, and gradient images of already localized pills) are processed by so called 'training CNNs'. A knowledge distillation model compression framework then condensed the training CNNs into smaller footprint CNNs ('student CNNs'), employed during inference. The CNNs were designed to embed features in a metric space, where cosine distance was utilized as a metric to determine how similar the features, produced by the CNNs. During the training of the streams, siamese neural networks (SNNs) were used with three inputs: the anchor image, a positive, and a negative sample, while the applied triplet loss was responsible to minimize the distance between the anchor image and positive samples, and to increase the distance between the anchor image and negative samples.

The winner model was improved in (Ling et al., 2020) with better accuracy tested on the CURE dataset. The teacher-student compression approach was left and a separate OCR stream, and a stream fusion network was introduced. The OCR stream was responsible for text localization, geometric normalization, and feature embedding with the Deep TextSpotter (Busta et al., 2017). In addition to the OCR stream, RGB, texture, and contour streams were used; segmentation was performed using an improved U-Net model to generate the stream inputs.

Our approach has similar structure to (Ling et al., 2020) but with a few modifications: we replaced the OCR method with an LBP (local binary pattern) (Ojala et al., 1994) stream, we utilize a more refined backbone in streams, we use state-of-the-art YOLO network for object detection, and we added attention mechanisms to the models. The performance of our multi-stream framework was compared to the architecture of (Ling et al., 2020) in (Rádli et al., 23b), using the CURE dataset, showing a few percentage advantage in all test settings. The main contribution of this article is the improvement of our previous model by the introduction of a new triplet loss, which utilizes textual information about medicines. Details of our custom model are given in Section 4.

[1]The dataset is available at: https://www.kaggle.com/datasets/richardradli/ogyeiv2

3 DATASETS

3.1 Pill Images

Our fundamental use-case model is the operation of a dispensing verification device in order to capture images of various pharmaceutical pills, mostly in a controlled environment. Our custom dataset (OGYEIv2) was created under the following conditions:

- pills have uniform mid-gray background,
- some images were taken under an upper LED lamp equipped with a diffuser, which gave the tablets a good overall appearance, with no significant shadows but clear imprints,
- other images were captured using a side mounted LED strip lamp: the engraved surface patterns become clearly visible.

All images in our dataset went through a lens distortion correction (undistortion) and have been also pixel-wise annotated. The main parameters of OGYEIv2 are given in Table 1. For comparison we included those of CURE.

Table 1: Comparison of the CURE and our novel OGYEIv2 dataset.

	CURE	OGYEIv2
Number of pill classes	196	112
Number of images	8973	4480
Raw image resolution	800×800 2448×2448	3840×2160
Image resolution after undistortion	-	3746×2019
Instance per class	40-50	40
Backgrounds	6	1
Segmentation labels	No	Yes
Free-text description	No	Yes

3.2 Text Obtained from Pill Information Leaflets

To complement the visual data, we have collected the official information leaflets about the pills provided by the manufacturers. These leaflets include a general description of the visual appearance of medicines, which should be localized first. It was performed using NLP (natural language processing) methods since these information (typically 2 or 3 sentences in length) was in a specific section of the documents as free text. These sentences form the textual part of our online available OGYEIv2 dataset. At this moment, according to our knowledge, this is the only freely available free-text+image pill dataset. The extracted sentences were then tokenized, with particular attention to features such as pharmaceutical form, colour, shape, convexity, edge, and imprint or engravings. In the resulting text dataset, the distribution of words between the different classes shows that the class with the highest word count contains a maximum of 28 words, while the class with the lowest word count comprises a minimum of 5 words.
The word embedding procedure is described in Subsection 4.3.

Figure 1: Pill images are captured with different lamps. First line: engraved text is only visible with side-light. Second line: same pills but with different poses, where the white imprint is only visible in the right photograph.

4 ENHANCING OBJECT RECOGNITION WITH TEXTUAL CLASS INFORMATION

4.1 Overview

A schematic representation of our proposed model is depicted in Figure 2. Addressing the limitations of metrics embedding - which does not inherently solves the challenge of object detection and localization - the initial step is inference using a state-of-the-art YOLOv7 object detection model trained for detect-

ing pills in an image. It draws bounding boxes around the tablets it detects, which we later crop and feed to our multi-stream network.

Subsequently, in the initial phase of image embedding (Phase 1), four parallel data streams are used, characterized by closely aligned structures. In the second stage (Phase 2), which is trained independently, the information content of these distinct branches is fused.

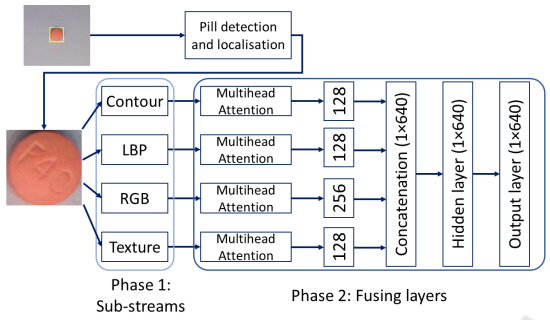

Figure 2: Overview of the proposed approach.

4.2 Image Streams

The implementation of different sub-streams is motivated by the intention to enforce the extraction and utilization of various image features that may be beneficial in different circumstances and for different types of pills. Our approach implements the following sub-streams:

1. Contour-stream: Contour images are produced by applying the Canny edge detector on a smoothed grayscale representation of the images, performed by a 7×7 Gaussian convolution kernel.
2. LBP-stream: LBP (Ojala et al., 1994), a widely adopted handcrafted local descriptor, finds application in numerous computer vision tasks, such as OCR (for both handwritten and printed text). We computed LBP representations from the grayscale inputs and integrated them into the same type of streams as the other descriptors.
3. RGB-stream: For RGB images, colour representations are directly fed into the embedding deep neural network.
4. Texture-stream: The images are generated by the subtraction of the smoothed grayscale representations from the original (grayscale) pill images.

In (Rádli et al., 2023) we analyzed the positive contribution of the LBP-stream and the attention mechanism, the ablation study of all sub-streams is omitted from this article due to size considerations. For all image based sub-streams we applied the well-proved EfficientNetv2 S (Tan and Le, 2021). The training mechanism is explained in Subsection 4.4.

4.3 Text Embedding

In our paper, we used the `hu_core_news_lg` CNN-based, large, pre-trained language model from the Python library called HuSpaCy (Orosz et al., 2023). The `hu_core_news_lg` model is trained on a large corpus of Hungarian Webcorpus 2.0 which includes more than 9 billion words based on newspapers and it is part of spaCy's language model pipeline. This model provides tokenization, sentence splitting, part-of-speech tagging, lemmatization, dependency analysis and named entity recognition, and includes pre-trained word vectors. The "lg" in the model name stands for "large," indicating that it is a relatively large-sized model with a broader vocabulary and potentially better performance on certain tasks compared to smaller models. It contains 200 000 unique word vectors and produces 300-dimensional floret vectors.

4.4 Training of Sub-Streams

To train the stream networks, we employ SNNs (siamese neural networks) featuring three inputs, where an anchor image is denoted by I_a, a positive example is represented by I_p, and a negative example is denoted by I_n (see the schematic illustration in Figure 3). In contrast to the approach employed by Ling et al. (Ling et al., 2020), where relatively light convolutional neural networks (CNNs) were utilized, our models take advantage of the more refined architecture of EfficientNetv2 S (small-scaled DNN model). EfficientNet (Tan and Le, 2019) is a generally applicable network for computer vision tasks, optimized for depth, width, and resolution. In contrast to the EfficientNet (v1) backbone EfficientNetV2 (Tan and Le, 2021) has several significant improvements. First, EfficientNetV2 extensively incorporates both MBConv (inverted residual block) and fused-MBConv, which has already proven to be very efficient in MobileNetV2 (Sandler et al., 2018). Secondly, one of the key difference is that EfficientNetV2 favors smaller expansion ratios in MBConvs, aiming to minimize memory access overhead associated with higher ratios. Moreover, EfficientNetV2 uses smaller 3×3 kernel sizes. To compensate this reduced receptive field more layers are utilized. Finally, another distinctive feature is the complete removal of the last stride-1 stage presented in the original EfficientNet. According to (Tan and Le, 2021) EfficientNetv2 outperformed its competitors in accuracy on ImageNet

ILSVRC2012, while learning significantly faster than others, other backbone networks could also be used for the proposed sub-streams. The CNN of (Ling et al., 2020) has 9M parameters for the RGB stream and 2.2M for the texture and contour streams, while EfficientNetV2 S, with multihead attention involved, 20.4M parameters.

Before the concatenation of the embedding vectors, we integrated a multihead attention module (Vaswani et al., 2017) within each stream. To gather the information from these data streams, the resulting output vectors were concatenated and a hidden layer and an output layer were combined to create the final embedding (see Figure 2). Notably, during the training of the fusion network the individual streams were kept frozen preserving their existing parameters.

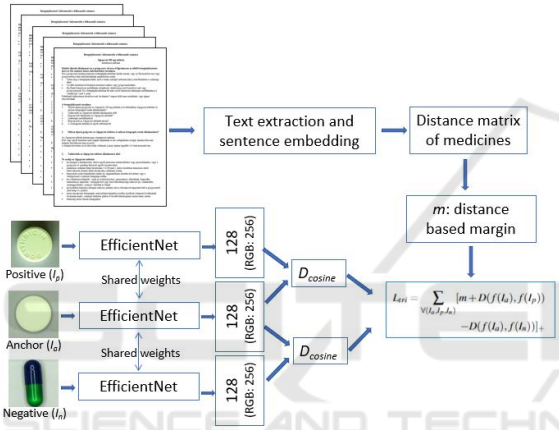

Figure 3: Overview of the training of a stream.

4.5 Loss Functions

The objective of metric embedding learning is to acquire a parametric function $f_\theta(I) : \mathbb{R}^F \rightarrow \mathbb{R}^E$, parameterized by θ (the parameters of the embedding network). This function is designed to map images into the embedding space where similar images are to be placed to metrically close positions, dissimilar images to distant points. This is achieved by the *Triplet loss* (Schroff et al., 2015) presented with the inputs I_a, I_p, and I_n. Formally:

$$L_{tri} = \sum_{\forall (I_a, I_p, I_n)} [m + D(f(I_a), f(I_p)) - D(f(I_a), f(I_n))]_+ \quad (1)$$

where D is the distance function, and margin m defines how far negative samples are to be placed.

4.5.1 Dynamic Margin Triplet Loss

The proposed dynamic margin triplet loss (DMTL) is responsible to maintain larger margins for less similar pills and smaller margins for more similar ones. It is achieved by changing the value of m in Eq.1 as a function of the distances of word embeddings:

$$m = \alpha \cdot d_i^{Norm} \quad (2)$$

where α is the default margin value (set to 0.5). The normalized distance (d_i^{Norm}) is based on the distance between the anchor and the negative word embeddings:

$$d_i^{Norm} = 1 + \frac{(u-1) \cdot (d_{min} - d_i)}{d_{max} - d_{min}} \quad (3)$$

where d_i is the distance value from the Euclidean distance matrix, d_{max} is the maximum distance, while d_{min} is the minimum distance in a row of the matrix, and u is the upper limit.

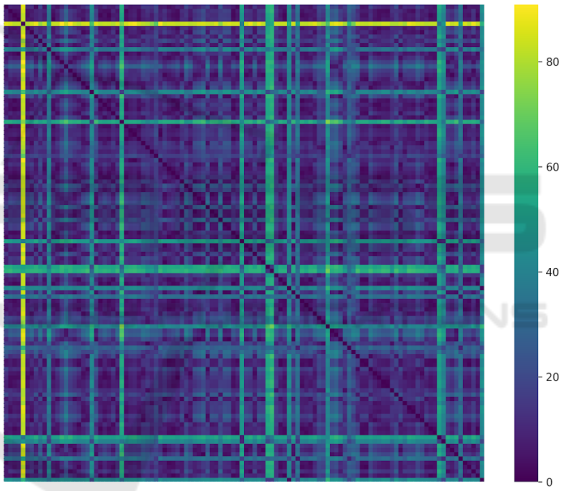

Figure 4: Visual illustration of pill-to-pill Euclidean distances based on word embeddings of the applied pre-trained language model.

Figure 4 shows the distance values (d_i) of pills based on the pre-trained language model embeddings of the official free-text description leaflets.

In Figure 5 we have selected the pill which is least similar (Algoflex Rapid - AR) to most of the others and the one that is very similar to others (Algopyrin - A). AR is responsible for the yellow vertical and horizontal lines in Figure 4. In the first row of Figure 5 the most similar (Tritace - T) and less similar pills (CalciKid - CK) are given with their distance values. The second row shows the pills most similar to Algopyrin - namely Dorithricin (D) and Naprosyn (N). Althogh we can intuitively agree on the distances within the rows, we can think that N is more similar to A than T is to AR, which is not represented by distances

based on word embeddings. Thus, the pre-trained embeddings contain usable information but could be improved in the future. The impact of using DMTL is revealed in the next section.

Figure 5: Visual illustration of examples of pill-pill Euclidean distances (d) based on word embeddings of the applied pre-trained model. First row: Algoflex Rapid and the two pills that are most and least similar to it. Bottom row: Algopyrin and the two most similar drugs to it.

Investigating the literature we found the most similar idea in (Zhou et al., 2020) called *Ladder-loss*. Ladder-loss was proposed as a loss function for visual-semantic embedding. There the usage of the triplet loss is to minimize the distance between a query image and its corresponding sentence while maximizing the distances between the query image and a set of irrelevant sentences. First, they calculate the relevance degrees between images and each sentence using similarity functions based on an NLP embedding model. After that, these relevance degree values are divided into L levels with predefined thresholds (the rungs of ladder). The similarities of a query image with texts is placed in an inequality chain, based on these level based classification of texts. Finally, these inequality chains define individual margin values planted into the triplet loss (see Eq. 1).

In contrast to Ladder-loss, in our approach, we are injecting semantic (textual) information into our purely visual embedding model. Thus we improve the training of the visual metrics model from a binary (similar/dissimilar) approach to a more refined scaled method. Moreover, we don't classify the samples into categories with distinct levels of relevance but apply the above defined normalization to obtain directly the continuous margin values as defined by Eq. 2 and Eq. 3.

5 EXPERIMENTS

In our paper, we conduct "two-sided" tests, where both sides of the pills are categorized into the same class, and we apply 5-fold cross-validation to get better statistical reliability.

During the evaluation procedures we adhered to the standard method where the query image underwent the embedding process and the resulting embedding vector was compared to the embedding vectors of randomly chosen reference pills using Euclidean distance. (In the future, we intend to test k-nearest neighbours matching, as there are typically several available reference images per class.) We ranked the results to determine the values of Top 1 and Top 5 accuracy, listed in Table 2, 3, and 4 at different values of u of Eq.3. We have also investigated all possible configurations to include the dynamic margin triplet loss at the two phases of our approach (see the first two columns of Tables 2- 4).

During the training of the sub-streams, the following hyper-parameters were set: the Adam optimizer was used, the learning rate was set to 1×10^{-4} for all four streams, weight decay regularization was applied with a coefficient of 1×10^{-5}, and the batch size was set to 32. The default value of the margin m was chosen to be 0.5 for both loss functions. Each model was trained for a total of 30 epochs and only the best weight file was kept.

As for the fusion phase (Phase 2), distinct hyper-parameter settings were implemented. The batch size was set to 32, the learning rate was adjusted to 2×10^{-4} for enhanced stability, and weight decay was initialized at 1×10^{-8}. In this phase, we introduced a learning rate scheduling mechanism where the initial learning rate was updated in every 5 epochs with a gamma of 0.1. The network was trained over 30 epochs and again only the file containing the best weighting factors was preserved.

Figure 6. displays training and validation loss curves of our model in Phase 2 (both phases used word embeddings, and the u parameter in our loss function was set to 4). The loss curves indicate a stable model where no signs of overfitting or underfitting are observed.

The training and testing processes were run on an NVIDIA Quadro RTX 5000 GPU card equipped with 16 GB of VRAM memory. The following tables show our experimental results:

Throughout the experiments, the use of DMTL had a clearly positive effect on the accuracy of the model. Specifically, when DMTL was applied in both Phase 1 and Phase 2, the model consistently achieved the highest performance, reaching a peak Top 1 accu-

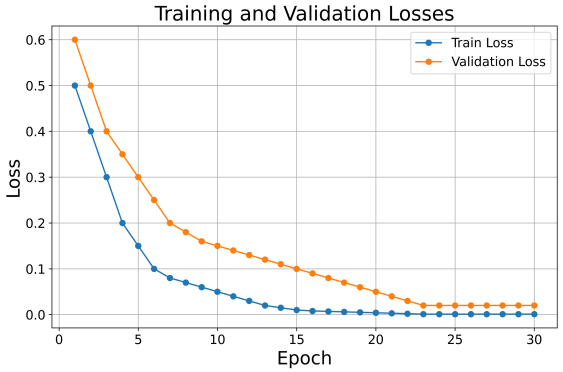

Figure 6: Evolution of loss functions during training.

Table 2: Results of ablation study of the dynamic margin triplet (DMTL) loss on the OGYEIv2 dataset, u=2.

Phase 1	Phase 2	Top 1 Acc.	Top 5 Acc.
w/o DMTL	w/o DMTL	91.72	98.125
w/o DMTL	w/ DMTL	91.72	98.125
w/ DMTL	w/o DMTL	92.34	98.43
w/ DMTL	w/ DMTL	**93.43**	**99.06**

racy of 93.56 % for u=3 and peak Top 5 accuracy of 99.84% for $u = 4$. This confirms the effectiveness of DMTL in enhancing the model's capabilities for pill recognition.

6 CONCLUSION

In the domain of pill recognition, we have adopted the metrics learning methodology of previously successful approaches (Zeng et al., 2017), (Ling et al., 2020), (Rádli et al., 2023), (Rádli et al., 23b). Beside, that our data streams are more uniform and can be trained in more straightforward ways than the previous models of (Zeng et al., 2017) and (Ling et al., 2020), we introduced a new triplet loss called dynamic margin triplet loss. With the help of DMTL we reached notable improvements (1.84 % in Top 1 accuracy, $u = 3$) compared the best previous model (Rádli et al., 23b). DMTL could inject textual information, generated by general language model embedding, into the visual model. Thus we learnt that even short free-text could add useful information to visual models in object recognition. We utilized k-fold cross-validation to test the robustness of our model, ensuring the evaluation of its performance across diverse subsets of the data. Additionally, we also created a novel pill dataset, named OGYEIv2.

In the near future, we are going to extend our tests to larger datasets, such as the NLM-NIH (Yaniv et al., 2016), and plan to perform not only two-sided but also one-sided tests.

Table 3: Results of ablation study of the dynamic margin triplet loss (DMTL) on the OGYEIv2 dataset, u=3.

Phase 1	Phase 2	Top 1 Acc.	Top 5 Acc.
w/o DMTL	w/o DMTL	91.72	98.125
w/o DMTL	w/ DMTL	92.81	99.21
w/ DMTL	w/o DMTL	92.96	99.21
w/ DMTL	w/ DMTL	**93.56**	**99.37**

Table 4: Results of ablation study of the dynamic margin triplet loss (DMTL) on the OGYEIv2 dataset, u=4.

Phase 1	Phase 2	Top 1 Acc.	Top 5 Acc.
w/o DMTL	w/o DMTL	91.72	98.125
w/o DMTL	w/ DMTL	92.5	99.06
w/ DMTL	w/o DMTL	92.96	99.06
w/ DMTL	w/ DMTL	**93.53**	**99.84**

ACKNOWLEDGEMENTS

This work has been partly supported by the 2020-1.1.2-PIACI-KFI-2021-00296 and the TKP2021-NVA-10 project of the National Research, Development and Innovation Fund. We also acknowledge the financial support of the Hungarian Scientific Research Fund grant OTKA K-135729. We are grateful to the NVIDIA corporation for supporting our research with GPUs obtained by the NVIDIA Hardware Grant Program. Last but not least, we would like to thank József Bene for his work in creating the dataset.

REFERENCES

Busta, M., Neumann, L., and Matas, J. (2017). 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.

Cronenwett, L. R., Bootman, J. L., Wolcott, J., Aspden, P., et al. (2007). *Preventing medication errors*. National Academies Press.

Heo, J., Kang, Y., Lee, S., Jeong, D.-H., and Kim, K.-M. (2023). An accurate deep learning–based system for automatic pill identification: Model development and validation. *J. Med. Internet Res.*, 25:e41043.

Ling, S., Pastor, A., Li, J., Che, Z., Wang, J., Kim, J., and Callet, P. L. (2020). Few-shot pill recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9789–9798.

Nguyen, A. D., Nguyen, T. D., Pham, H. H., Nguyen, T. H., and Nguyen, P. L. (2022). Image-based contextual pill recognition with medical knowledge graph assistance. In *Asian Conference on Intelligent Information and Database Systems*, pages 354–369. Springer.

Ojala, T., Pietikainen, M., and Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions.

In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585. IEEE.

Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., and Farkas, R. (2023). Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue*, pages 58–69, Cham. Springer Nature Switzerland.

Rádli, R., Vörösházi, Z., and Czúni, L. (23b). Pill metrics learning with multihead attention. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management SCITEPRESS - Science and Technology Publications (2023)*, pages 132–140.

Rádli, R., Vörösházi, Z., and Czúni, L. (2023). Multistream pill recognition with attention. In *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 942–946.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823.

Tan, L., Huangfu, T., Wu, L., and Chen, W. (2021). Comparison of RetinaNet, SSD, and YOLOv3 for real-time pill identification. *BMC Medical Informatics and Decision Making*, 21:1–11.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4.

Yaniv, Z., Faruque, J., Howe, S., Dunn, K., Sharlip, D., Bond, A., Perillan, P., Bodenreider, O., Ackerman, M. J., and Yoo, T. S. (2016). 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.

Zeng, X., Cao, K., and Zhang, M. (2017). 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.

Zhou, M., Niu, Z., Wang, L., Gao, Z., Zhang, Q., and Hua, G. (2020). Ladder loss for coherent visual-semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13050–13057.