NOVEL DEEP LEARNING METHODS FOR OPTICAL RECOGNITION

Theses of Ph.D. dissertation

by:

Richárd Bence Rádli

Supervisor: László Czúni, PhD

University of Pannonia

Faculty of Information Technology

Doctoral School of Information Science

2026

1 Introduction

This dissertation tackles three distinct computer vision problems, each with significant practical applications in domains such as visual object recognition, anomaly detection, and model training.

The visual inspection of products is often inevitable in many manufacturing processes. Since anomalous items are often missing during training, unsupervised anomaly detection is the most suitable approach for building a defect detection system. Therefore, I investigated different ways to improve autoencoder (AE) networks in the first part of this work. I compared and modified an existing SSIM-AE [2] by increasing the number of convolutional layers of the network and turning the base model into a denoising AE. Furthermore, I investigated the integration of Spatial Transformer Networks (STNs) [4]. While simultaneous training of convolutional layers and STN weights proved ineffective for reconstruction, I discovered that pre-training the STN to normalize image orientation significantly improves the AE's defect detection performance.

In the second part of this study, I present a new approach to building deep neural models for classification tasks using an iterative, pseudo-inverse optimisation technique. Deep learning neural networks show a significant improvement over shallow ones in complex problems. Their main disadvantages are their memory requirements, the vanishing gradient problem, and the time-consuming solutions to find the best achievable weights and other parameters. Since many applications (such as continuous learning) would need fast training, one possible solution is the application of sub-networks that can be trained very fast.

The strategy involves building deep randomized structures in consecutive phases, with a high likelihood that classification performance will improve, or at least not decline, as the network depth increases. The generated structure's weights are determined in separate phases without applying backpropagation. Additionally, I

introduced a novel pruning and weight replacement scheme to further refine the network's structure and classification performance. The training time of the presented method, in most test cases, is typically a fraction of popular gradient-based optimization methods, such as Adam or SGD.

Medications are widely and increasingly used worldwide, especially among the elderly. In some countries, 66% of all adults use prescription pills, which is increasing with age: 75% of those aged 50 to 64, while 91% of those aged 80 and older use prescription pills [3]. The policy organisations claim that drug errors are the most frequent mistakes made in healthcare. As a result, there is a lot of interest in improving the safety of the medication use process. Optical pill recognition can be a beneficial solution to minimize these errors as a final check when giving pills to patients or in self-service. Although AutoML solutions offer an easy-to-use approach for pharmacists, these solutions have proven highly unreliable in clinical scenarios, highlighting the urgent need for more robust and accurate alternatives. Therefore, my third contribution targets aiding the pill dispensing process for pharmacists.

The main contribution here is the further development of a multi-stream, two-phase neural network framework for pill recognition. This method utilises metric learning, an effective method for few-shot object recognition, in order to identify pills. Building upon the previous pill recognition challenge-winning multi-stream network architecture [5], I modified the model to capture a wider range of pill features. Specifically, the original OCR stream was replaced with an LBP-based stream, and the impact of multi-head attention layers in the second phase was also investigated. A significant achievement is developing an improved triplet loss function, which integrates word embeddings or certain visual features to incorporate pill similarity into the model. This enhancement allows for finer-grained adjustments to the embeddings, surpassing the accuracy of traditional triplet loss methods. The proposed approach was validated using a novel pill dataset, which I created with the support of my colleagues [6].

2 Research methodology and utilized tools

The findings and proposals presented in this dissertation follow a unified research methodology over all thesis groups. Each investigation is motivated by a practical, real-world problem, which serves as the starting point for the subsequent analysis. This is followed by a comprehensive review of the relevant literature, with particular interest in state-of-the-art methods and their limitations in the given application field.

Building upon these insights, I designed and implemented novel or extended solutions specifically tailored to the given challenges of each problem. This process included the development of various custom models, the adaptation and refinement of existing architectures, as well as the creation of custom datasets. In addition, all experiments were supported by self-developed codebases to ensure reproducibility and flexibility in evaluation.

For each thesis group, I conducted systematic evaluations and tests in which the proposed methods were thoroughly compared against established baseline and state-of-the-art approaches. During the evaluation procedures, I made an effort to ensure fairness and consistency, typically involving standardised data splits, appropriate performance metrics, and repeated trials or cross-validation.

The computational experiments were carried out on a dedicated workstation with the following hardware configuration: an AMD Ryzen 5 5600X processor, an NVIDIA RTX 5000 GPU with 16 GB of VRAM, and 32 GB of DDR4 RAM. The utilized operating systems were Windows 11 Pro and Ubuntu 24.04 LTS.

The software environment was based on Python 3.8-3.11, utilising libraries for numerical computation (NumPy), deep learning (PyTorch), and computer vision (OpenCV). RayTune was employed for parameter search. Other useful libraries, such as Pandas, Matplotlib, Scikit-Image and Scikit-Learn, were also involved. The development was done within PyCharm IDE Professional.

The framework for **Thesis group I** was structured in two halves. In the first one,

four autoencoder architectures were evaluated: a standard SSIM-AE, an enhanced variant (SSIM-AEe) with increased convolutional layers, a denoising autoencoder (SSIM-DAE) trained with masked blocks, and its corresponding enhanced version (SSIM-DAEe).

To evaluate the performance of the autoencoders, I used three datasets: two texture image datasets (Texture 1 and Texture 2), sourced from [2], comprised 100 defect-free and 50 defective images each, showcasing diverse fault types. A custom CPU dataset, consisting of 60 images of CPU backplanes, was included. This dataset was partitioned into 48 training images and 12 test images, with the latter subjected to three distinct defect simulations:

- CPUa: I added extra components with a visually realistic appearance;
- CPUc: the test images were loaded with synthetic contamination;
- CPUm: some components (such as pins or capacitors) were removed.

These operations are carried out on all 12 test images. For training, the hyperparameter setups followed those in the original paper.

For the second phase of experiments, the Screw dataset from the MVTec collection [1] was added. This dataset comprised 320 defect-free images, each of size 1024×1024. The defective Thread-top class, consisting of 23 images, was utilized for the defect detection evaluations. Four distinct training scenarios were investigated:

1. A baseline AE was trained for reference.
2. An AE and an STN were jointly trained, both initialised with random weights.
3. The STN was pre-trained to normalise input pattern orientation, followed by AE training with the STN's localisation network frozen.
4. Similar to scenario 3, but with the STN's weights allowed to refine during AE training.

Across both experimental phases, the Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE) were computed to demonstrate the reconstruction

fidelity of the autoencoders.

During the experiments in **Thesis group II**, I compared six methods: two popular gradient descent optimisers (Adam and SGD), an autoencoder-based ELM named H-ELM, and three variants of my proposed method.

I used 13 datasets from the UCI Machine Learning Repository. The selection included five image datasets and eight traditional real-world, non-image datasets, each addressing classification tasks. The datasets were partitioned into train, validation, and test sets following a 70-15-15 split ratio. In all network configurations, the validation set was used for hyperparameter tuning. Additionally, for the Adam and SGD, the validation set was employed to support an early stopping mechanism during training. As part of the preprocessing procedure, each dataset was normalised by scaling the features between 0 and 1.

I executed ten tests for all networks on every dataset and averaged the results. The metrics I used were training and testing accuracy, training and testing F1-score, and training time.

In **Thesis group III**, I first examined the performance of existing AutoML platforms and the popular object detector YOLO11 for clinical pill identification using two laboratory-controlled datasets that I created, as well as real-world datasets. While platforms such as Google Vertex AI, Microsoft Azure Custom Vision, and Amazon Rekognition Custom Labels showed high accuracy in some tests, their performance was inconsistent and unreliable in clinical scenarios.

Motivated by these limitations, I modified and extended an existing multi-stream, two-phase neural network framework, which I compared against the previously mentioned platforms on the two laboratory-controlled datasets that I constructed in a one-shot learning scenario (i.e., the model received only a single reference image per class during testing). In this setup, the model was trained on 76 drug classes and tested on 30 previously unseen classes and their corresponding images.

Afterward, I analysed the performance of my proposed method on two datasets:

the publicly available CURE image database and a custom dataset, OGYEIv2, which I created with the help of my colleagues.

It is important to note that in these experiments I conducted one- and two-sided tests for the CURE dataset, while for OGYEIv2, the tests were only two-sided. The two-sided test considers both sides of a pill to belong to the same category, whereas the one-sided test treats the two sides as belonging to two separate classes, doubling the number of classes.

To increase the statistical robustness of the evaluation, I employed 5-fold cross-validation. Throughout the assessment process, I followed the conventional methodology used for metric learning models: the reference and consumer images underwent the metric embedding procedure, and then the generated vectors were compared to each other using the Euclidean distance to find the best match. Top-1 and Top-5 accuracies were calculated, averaged, and then converted to percentages for all five folds.

Two use cases were examined considering the selection of the reference set:

1. Traditionally, the reference set includes all drug classes in the dataset. In this case, I selected one fold from the consumer classes for queries during the inference process.
2. In the second type of reference set, I included only such classes that are present in the actual query fold, ensuring that the model compares identical consumer and reference classes.

To further analyse the model's performance, I carried out ablation studies, systematically removing specific stream networks to investigate the impact of domain-related features. Additionally, I performed experiments to evaluate the individual contribution of each stream. Another investigation revealed the relevance of different light sources, top or side, in pill recognition.

3 Thesis groups

Thesis group I *Enhancing autoencoder architectures to improve visual defect detection abilities*

1.1 If an autoencoder (AE) can reproduce (anomaly free) images with higher fidelity, it does not necessarily imply that it is better in anomaly detection.

I investigated the performance of different SSIM AEs for defect detection on repetitive and semi-repetitive structures across three datasets. In certain instances, despite achieving higher Structural Similarity Index (SSIM) values and lower Mean Squared Error (MSE) values, a model may exhibit poorer overall anomaly detection performance as measured by the Receiver Operating Characteristic Area Under the Curve (ROC AUC) compared to its counterpart. This observation suggests that while SSIM and MSE metrics indicate superior image fidelity and reconstruction accuracy, they do not consistently correlate with improved anomaly detection as reflected in ROC AUC.

1.2 I proposed a framework to show that Spatial Transformer Networks (STNs) can be trained to normalise the orientation of the input images. The combination of AEs with STNs can improve the anomaly detection efficiency of AEs.

During my investigation, I observed that simultaneous training of the convolutional layers of the AEs alongside the weights of STNs did not yield satisfactory reconstructions by the decoder. Instead, my approach focused on training the STN specifically to normalise the orientation of input images (as shown in Figure 1). To evaluate the effectiveness of this strategy, I assessed its performance across three classes of input patterns using reconstruction error and standard anomaly detection metrics.

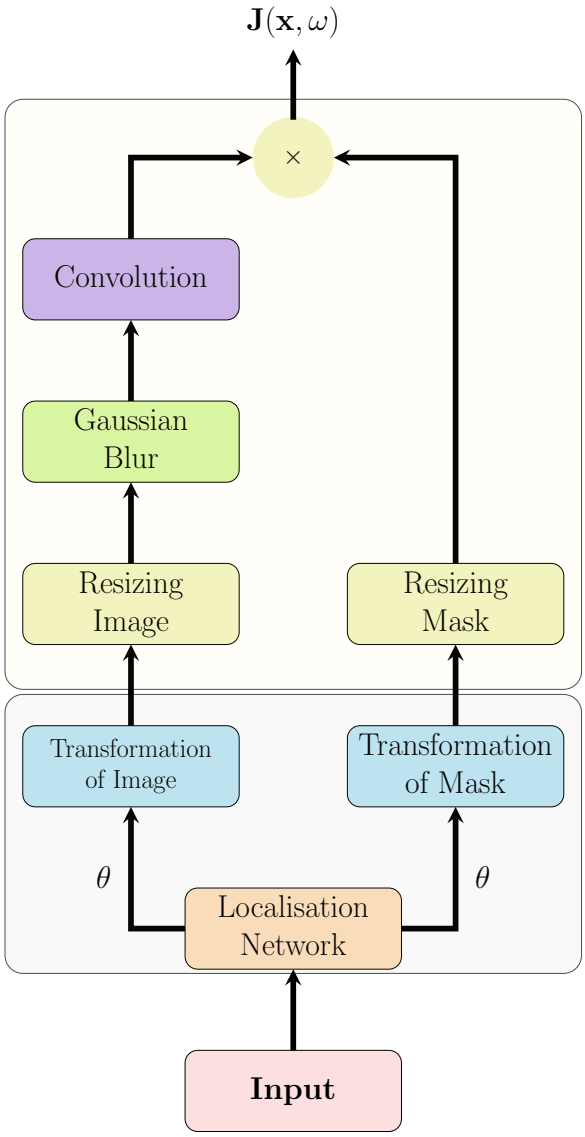

Figure 1: STN with custom loss function

The corresponding articles are [RR4] and [RR5].

Thesis group II *Fast and scalable deep randomized neural networks*

2.1 I proposed a framework where deep randomised structures are built in consecutive phases. The yielded network's training is significantly faster than backpropagation, while its classification is more accurate in most test cases with the same number of neurons.

I developed a framework that allows adding an arbitrary number of hidden layers to a single-layer feed-forward neural network. In each phase, new neurons are added using the following three strategies:

(a) direct reuse of previously trained output neurons;

(b) perturbation of previously learned output neurons via additive Gaussian noise;

(c) random sampling with orthogonal weights.

This approach allows efficient exploration of the parameter space while preserving useful representations learned in earlier phases. Important to note that in this framework the number of neurons decreases exponentially from layer to layer.

2.2 Built upon the previous framework, I proposed a scalable method to prune the model's least contributing neurons and replace them with the most contributing neurons of the auxiliary network(s) to further improve the previous results.

My proposed iterative refinement method enhances the network's overall classification performance by preserving beneficial random connections while systematically identifying and replacing suboptimal ones. The approach leverages auxiliary networks for evaluation, selecting the most effective connections and integrating them into the original network. The method is fully scalable, allowing for the creation of an arbitrary number of auxiliary networks and extension to additional layers.

The base and enhanced versions of the method were compared against the two most widely used gradient descent-based optimisers, Adam and SGD, and a method utilising autoencoders and the FISTA optimiser across 13 datasets and showed competitive results. A visual comparison of a fully connected neural network (with the Adam optimizer) and the proposed DEV-DRNN AUX 2 can be seen in Figure 2.

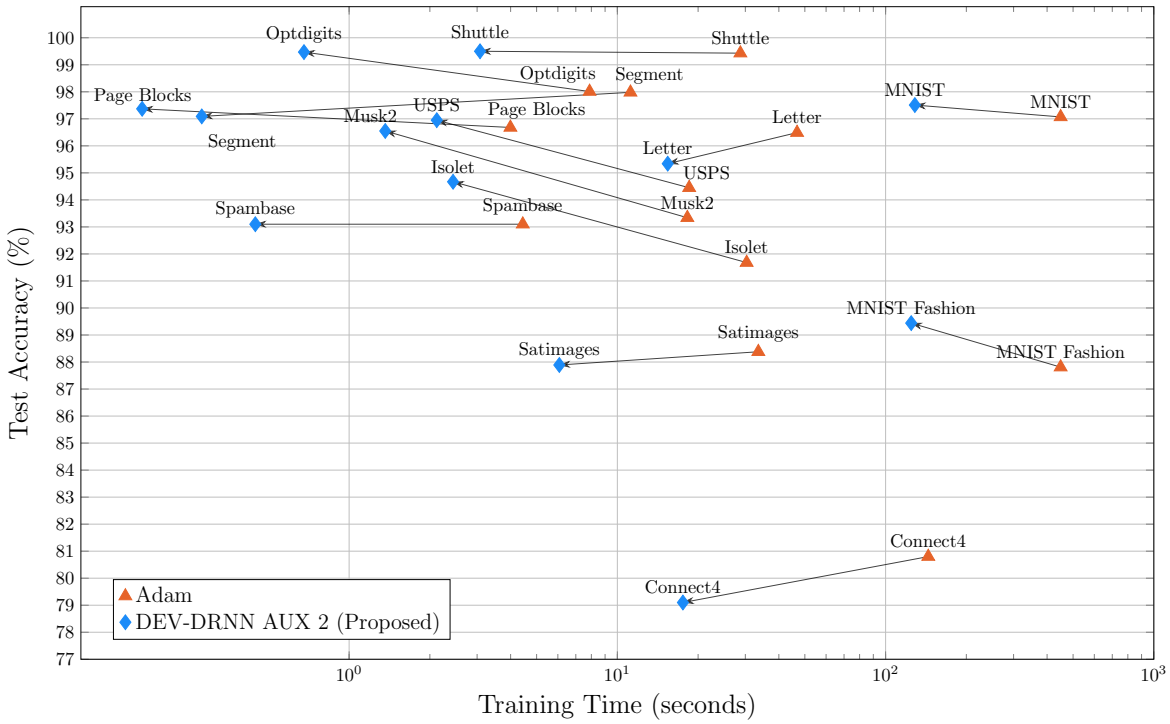

Figure 2: Comparison of the performance of fully connected networks (Adam optimizer) and the proposed DEV-DRNN AUX 2 method on 13 datasets (as the average of 10 experiments). The training time is shorter in all cases, while in 9 out of 13 cases, the test accuracy is equal or higher with the proposed solution. It should be noted that the training time is on the logarithmic scale, which shows the significant advantage of the proposed method.

The corresponding articles are [RR3] and [RR6].

Thesis group III *Metric-based pill recognition with the help of textual and visual cues and multi-head attention*

3.1 I proposed a multi-stream two-phase neural network by employing multi-head attention.

I created a multi-stream, two-phase neural network that integrates an EfficientNetV2-Small network as the backbone in the first phase and introduces multi-head attention modules in the second phase, as shown in Figure 3. Employing CNNs and multi-head attention modules can improve the pill recognition abilities of already existing frameworks. Unlike other approaches, this method does not require semantic networks such as OCR, thus making it more efficient and independent. Evaluations were carried out on two datasets, where the proposed network achieved superior results compared to the concurrent network.

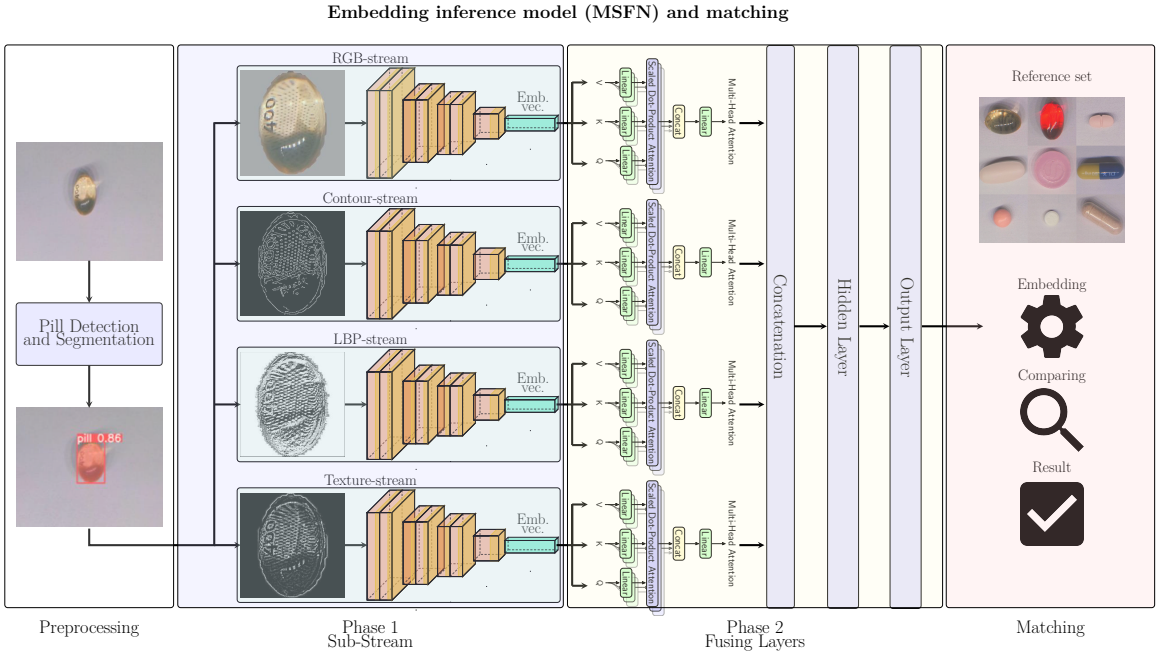

Figure 3: Overview of the proposed Multi-Stream Fusion Network.

3.2 The model's accuracy can be enhanced by injecting high-level cues, gained from free-text or visual appearance, into the training of multi-stream metric models.

Expanding upon the previous architecture, I demonstrated two methods for improving the model's classification accuracy. I generated discriminative cues: one from natural language text resources and another from high-quality reference images using visual descriptors. By integrating these high-level cues, the model is better able to classify visual differences between samples, improving its overall discriminative performance.

3.3 I compared AutoML platforms and YOLO11 with my proposed method and demonstrated its superior performance under a one-shot learning setting on controlled pill recognition datasets. While the evaluation was limited to this setting, the results suggest that the proposed method offer advantages more broadly.

I trained and evaluated three widely used AutoML platforms—Google Vertex AI, Microsoft Azure Custom Vision, and Amazon Rekognition Custom Labels—alongside the YOLO11 object detector on custom pill datasets without segregating classes during training and testing. Among the AutoML platforms, Google Vertex AI achieved the highest testing accuracy; however, all platforms exhibited substantial performance variability under clinical testing conditions. To further assess robustness and generalization, I evaluated the proposed multi-stream, two-phase neural network on two controlled test sets with fully segregated training and testing classes, using only a single reference image per class at inference time. Even under this constrained one-shot learning scenario, the model maintained superior performance compared to the alternative solutions.

The corresponding articles are [RR1], [RR2], [RR7], [RR8] and [RR9].

4 Applications of the scientific results

This research introduces practical solutions that apply to various Industry 4.0 technologies and industries like manufacturing and healthcare.

The autoencoder networks developed in **Thesis group I** have proven versatile and effective across various defect detection applications. Notably, these networks were directly implemented within a system as part of the collaborative project 2018-1.3.1-VKE-2018-00048 (2019-21), conducted in partnership with Foxconn Hungary Kft. and Eclipse Automation Hungary Kft.

The method in **Thesis group II** excels in scenarios demanding quick training, offering a highly efficient classification module that can be easily integrated as a final decision layer into larger, more complex network architectures.

The pill recognition methodologies detailed in **Thesis group III** are immediately applicable to real-world scenarios. My solutions were successfully implemented in a nurse trolley as part of the 2020-1.1.2-PIACI-KFI-2021-00296 (2022-24) project, collaborating with IVM Zrt. This product's innovative design was awarded the National Award for "Quality Innovation 2023." Furthermore, ongoing collaborations with the University of Pécs are exploring additional utilisation for the developed metric-learning network.

5 Related publications

Journal publications

[RR1] Rádli, R., Vörösházi, Z., & Czúni, L. (2024). Metric-based pill recognition with the help of textual and visual cues. *IET Image Processing*, 18(14), 4623–4638.

[RR2] Ashraf, AR., Rádli, R., Vörösházi, Z., & Fittler, A. (2026). Code-Based Versus AutoML Methods for Pill Recognition in Clinical Settings: Comparative Performance Study. *JMIR Medical Informatics*, 14 (1), e79160

[RR3] Rádli, R., & Czúni, L. (2026). Faster and Accurate: Developmental Deep Randomized Networks. *Neurocomputing*, Submitted to review.

Full papers in conference proceedings

[RR4] Rádli, R., & Czúni, L. (2021). About the Application of Autoencoders for Visual Defect Detection. In *29. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, Václav Skala - UNION Agency, pp. 181-188.

[RR5] Rádli, R., & Czúni, L. (2022). Improving the Efficiency of Autoencoders for Visual Defect Detection with Orientation Normalization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, SciTePress, pp. 651-658.

[RR6] Rádli, R., & Czúni, L. (2023). Deep Randomized Networks for Fast Learning. In *International Conference on Learning and Intelligent Optimization*, Springer International Publishing, pp. 121-134.

[RR7] Rádli, R., Vörösházi, Z., & Czúni, L. (2023). Multi-stream pill recognition with attention. In *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, IEEE, Vol. 1, pp. 942-946.

[RR8] Rádli, R., Vörösházi, Z., & Czúni, L. (2023). Pill Metrics Learning with Multihead Attention. In *Proceedings of the 15th International Joint Conference*

on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR, SciTePress, pp. 132-140.

[RR9] Rádli, R., Vörösházi, Z., & Czúni, L. (2024). Word and Image Embeddings in Pill Recognition. In *In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, SciTePress, pp. 729-736.

Other References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[2] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[3] Lee Shirey Emily Ihara, Laura Summer. Prescription Drugs, September 2002. `https://hpi.georgetown.edu/rxdrugs/`.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28:2017–2025, 2015.

[5] Suiyi Ling, Andreas Pastor, Jing Li, Zhaohui Che, Junle Wang, Jieun Kim, and Patrick Le Callet. Few-shot pill recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9789–9798, 2020.

[6] Richárd Rádli, József Bene, and Zsolt Vörösházi. OGYEIv2, 2024. `https://www.kaggle.com/dsv/8417535`.