Universal Lesion Detection Utilising Cascading R-CNNs and a Novel Video Pretraining Method

Shahin Amiriparian[1], Alexander Meiners[1], Daniel Rothenpieler[1],
Alexander Kathan[1], Maurice Gerczuk[1], Björn W. Schuller[1,2]

Abstract— **According to the WHO, approximately one in six individuals worldwide will develop some form of cancer in their lifetime. Therefore, accurate and early detection of lesions is crucial for improving the probability of successful treatment, reducing the need for more invasive treatments, and leading to higher rates of survival. In this work, we propose a novel R-CNN approach with pretraining and data augmentation for universal lesion detection. In particular, we incorporate an asymmetric 3D context fusion (A3D) for feature extraction from 2D CT images with Hybrid Task Cascade. By doing so, we supply the network with further spatial context, refining the mask prediction over several stages and making it easier to distinguish hard foregrounds from cluttered backgrounds. Moreover, we introduce a new video pretraining method for medical imaging by using consecutive frames from the YouTube VOS video segmentation dataset which improves our model's sensitivity by 0.8 percentage points at a false positive rate of one false positive per image. Finally, we apply data augmentation techniques and analyse their impact on the overall performance of our models at various false positive rates. Using our introduced approach, it is possible to increase the A3D baseline's sensitivity by 1.04 percentage points in mFROC.**

I. INTRODUCTION

As the second leading cause of death globally, cancer represents one of the most dangerous diseases of our time. According to the WHO[1], 9.6 million people died as a result of cancer in 2018, showing a growing trend with 10 million cases in 2020 [1]. Early detection of cancer drastically increases the chance of survival, enabling an early intervention to prevent further spreading. However, in about $50\,\%$ of all cases, cancer is still only detected at an advanced stage, resulting in a considerably worse course of the disease [2]. According to Crosby et al. [2], one of the five biggest challenges to enable an earlier detection is to develop systems that are able to recognise biological changes such as tissue alterations in a timely and accurate manner. To this end, recent work focused on developing and improving approaches for detecting lesions, an area of abnormal tissue that can be either benign or malignant.

Lesions are usually detected using medical imaging methods such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray or ultrasound [3]–[5] and subsequently, are examined by a medical doctor based on their shape, size, and location, following the commonly used Response Evaluation Criteria in Solid Tumors (RECIST) guidelines [6], [7]. However, studies have shown that human error can lead to inaccurate conclusions in up to $3\,\%$ of all cases, resulting in serious consequences for patients [8]. In this regard, computer-aided disease screening approaches can support physicians, contribute to earlier detection, and improve the overall detection rate.

Computer vision for medical image analysis is a well-established field of research and evolved considerably in the last years. Convolutional Neural Networks (CNNs) demonstrated their capacity to identify patterns in images and formed the basis for architectures, such as R-CNN, enabling a precise object detection [9], [10]. One limitation of these methods constitutes their inability to learn 3-Dimensional (3D) context from 2D input data which is crucial for discerning lesions [11]. To address this shortcoming, Yan et al. [11] proposed integrating 3D context into 2D regional CNNs, resulting in 3D context enhanced region-based CNNs for lesion detection. More recently, Transformer-based architectures have been utilised to effectively model 3D context found between slices [12]. Nevertheless, a drawback of 3D context models is that large 2D datasets (e. g., [13], [14]) cannot be utilised for pretraining. More recent 3D medical image approaches solved this challenge, introducing methods for combining 2D pretraining with 3D networks [15], [16]. In particular, these models are able to learn 3D representations while initialising their weights from 2D convolutional kernels. Further, Yang et al. [17] proposed an asymmetric 3D context fusion operator (A3D), a lesion detection framework that uses different weights for fusing 2D slices, resulting in a considerable performance increase and representing the current state-of-the-art on well-known datasets, such as the DeepLesion benchmark [18], and is still being expanded on, e. g., by adding Transformer-based slice attention modules [19]. In the early days, it was common to only recognise single types of lesions, such as skin [20] or liver [21] lesions. However, large-scale datasets such as the DeepLesion benchmark [18] tackled this challenge, enabling the detection of various lesion types using one model for the first time.

In this paper, we extend the studies of Yang et al. [17], proposing a novel R-CNN lesion detection approach evaluated on the DeepLesion dataset [4]. Our contribution is twofold. First, we introduce a lesion detection framework using cascading CNNs incorporated into an A3D architecture. Second, we present a new video pretraining approach

[1]Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Augsburg 86161, Germany `first.last@uni-a.de`

[2]GLAM – Group on Language, Audio, & Music, Imperial College London, London SW7 2AZ, UK

[1]https://www.who.int/health-topics/cancer

for medical imaging using consecutive frames from the YouTube VOS video segmentation dataset [22], thus taking into account the importance of pretraining, emphasised in previous works [17].

II. DeepLesion Dataset

For our experiments, we use the DeepLesion dataset [4] which is a large-scale, open-access dataset of medical images comprising $32,735$ types of lesions in $32,120$ CT slices derived from $10,594$ studies of $4,427$ unique patients.

While the raw RECIST annotations from the radiologists are provided, the National Institutes of Health (NIH) also published generated bounding boxes with five-pixel padding to the annotations. The patient's age and gender are also provided, as well as a flag for possible noisy scans. A majority of scans are 512×512 pixels in size, while others are 768×768 and 1024×1024. Additional slices $30\,\mathrm{mm}$ above and below the key slices are contained in the dataset for most CT images [4].

III. Methodology

Our approach for lesion detection with video pretraining is illustrated in Figure 1. In our experiments, we choose the state-of-the-art A3D as the base model, as it exhibits a promising slice fusion strategy. We use the truncated DenseNet-121 backbone and Feature Pyramid Network (FPN) from A3D and forward the generated feature map to the Region Proposal Network (RPN) component that creates bounding box proposals. Subsequently, we incorporate our model with the Hybrid Task Cascade (HTC) architecture composed of three Mask R-CNN branches for instance segmentation. A brief summary of the backbone, FPN, and HTC is provided in Sections III-A and III-B.

We start our approach by initialising the DenseNet-121 backbone with ImageNet [13] weights as done in A3D. A pretraining is performed using samples from the class "person" of the YouTube VOS dataset. Neighbouring frames are considered as an equivalent to adjacent CT scan slices (cf. Section III-C). The training process is then conducted as outlined in Section IV-A until sensitivities converge.

A. Context Fusion

Unlike 3D fusion operators which rely on the spatially symmetric transformation of 2D slices to ensure transformational equivariance, A3D acknowledges the uneven distribution of feature-relevant slices in medical imaging data [17]. A3D fuses the $D \times 512^2$ input features using an (asynchronous) fusion matrix $\in \mathbf{R}^{D\times D\times C}$, densely connecting over the slices' (D) features individually for each channel (C). We use $D \in \{3,7\}$ (slices) for our experiments. The fused features are further integrated into an FPN, which allows for small-scale lesion detection by integrating over features of different resolutions. The convoluted features of size $C_i \times \{128^2, 64^2, 32^2\}$, C_i denoting the channel size, are fused together, forming a feature map $\in 512 \times 128^2$ as input into the sequential RPN module. The FPN utilises channel sizes of $C_i \in \{256, 512, 1024\}$.

B. Instance Segmentation

We diverge from the A3D architecture by employing HTC for Instance Segmentation. Instead of relying on a linear R-CNN architecture, the HTC module utilises three interlaced Mask R-CNN branches (stages) to generate segmentation masks. Each Mask R-CNN branch integrates the updated bounding box and mask predictions of the previous branch. This allows for incremental predictions where the final branch is finetuned on the preceding stages' features.

C. Pretraining

Due to a lack of 3D imaging datasets, we aim to exploit the spatiotemporal information present in video data. Temporally close frames in videos tend to have a high correlation in pixels, and, as such, segmentation masks. In sliced medical imaging data, e. g., lesions, the area of interest spans over multiple neighbouring slices. We aim to learn this spatial dependency by pretraining our model on the Youtube VOS dataset, predicting segmentation masks on moving images. For pretaining, the models learn exclusively with the databases' "person" class. The dataset contains over $2,883$ videos total, sampled at $30\,\mathrm{fps}$ and segmentation mask annotations every fifth frame. Akin to the DeepLesion dataset, we create bounding boxes from the annotated segmentation masks, add padding of five pixels and convert the frames to greyscale. In this context, each frame corresponds to a slice of medical imaging data with an assigned thickness of $1\,\mathrm{mm}$. We employ a second pretraining approach denoted as Random Spacing which randomly skips 0 to 4 frames and adjusts the image thickness accordingly.

D. Data Augmentation

As the asynchronous nature of the models fusion process has the potential to impair its translational equivariance, we aim to improve on the data augmentations done in A3D [17]. We found that additionally applying Random Cropping [23] with ratios between 0.9 and 1 with a chance of $30\,\%$ worked best for this model, while rotational augmentation led to worse results.

IV. Experimental Settings and Results

A. Training Implementation

Our proposed model combines the A3D architecture with HTC to generate refined segmentation mask proposals. The A3D [17] module uses anchor ratios of $(0.5, 1.0, 2.0)$ and anchor scales of $(16, 24, 32, 48, 96, 192)$. We set the base learning rate to 0.04 with a weight decay ratio of 0.1. According to the linear scaling rule [24], this equals a learning rate of $2 \times \frac{0.04}{4\times 8} = 0.0025$. The model is initialised with pretrained weights from ImageNet. All experiments were performed using the official DeepLesion [4] train, test, and validation splits.

B. Results

The results of our experiments are listed in Table I. Due to the computational costs, we execute our experiments mainly using a slice depth of $D = 3$. Our best results

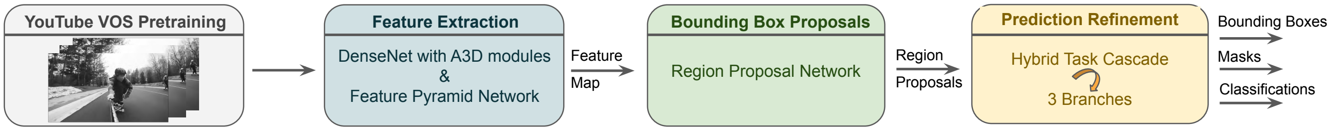

Fig. 1. High-level overview of our proposed approach comprising (i) video pretraining (YouTube VOS dataset), (ii) feature extraction (A3D + Feature Pyramid Network), (iii) bounding box proposals (Region Proposal Network), and (iv) prediction refinement (Hybrid Task Cascade with three branches) components. A detailed account of the approach is given in Section III.

TABLE I

COMPARISON OF ALL PROPOSED TECHNIQUES FOR THREE AND SEVEN INPUT SLICES. THE BEST RESULT FOR EACH FP IS BOLDFACED AND THE BEST MEAN RESULTS (mFROC[0.5, 1, 2, 4, 8, 16]) ARE MARKED WITH GREY SHADING. $\text{AUG}_1 = \{$FLIP, SHIFT, RESCALE, ROTATE$\}$, $\text{AUG}_2 = \text{AUG}_1 \cup \{$CROP$\}$.

		FP@						
Method	**Slices**	**0.5**	**1**	**2**	**4**	**8**	**16**	**mFROC**
A3D [17] + Aug_1 replicated (baseline)	3	72.59	80.99	87.07	90.92	93.77	95.95	86.88
A3D + Cascade R-CNN + Aug_1 ([17])	3	72.03	80.74	86.97	91.14	94.18	95.97	86.84
A3D + HTC + Aug_1	3	72.76	81.19	87.09	91.18	94.25	95.93	87.07
A3D + Cascade R-CNN + Aug_2 *(ours)*	3	72.60	80.76	87.07	91.43	94.40	96.29	87.09
A3D + Cascade R-CNN + Aug_2 *+ Pretraining (ours)*	3	73.23	81.56	87.58	91.53	94.50	96.15	87.43
A3D + HTC + Aug_2 *+ Pretraining (ours)*	3	73.47	81.60	**87.93**	91.92	94.79	96.25	87.66
A3D + HTC + Aug_2 *+ Pretraining + Random Spacing (ours)*	3	**74.06**	**81.78**	87.87	**92.10**	**95.13**	**96.60**	**87.92**
A3D [17] + Aug_1 replicated (baseline)	7	**78.54**	85.16	89.66	93.04	95.18	96.78	89.73
A3D + HTC + Aug_2 *+ Pretraining (ours)*	7	78.36	**85.24**	**90.41**	**93.44**	**95.79**	**97.07**	**90.05**

are achieved with the Random Spacing adjustment during pretraining. All models are evaluated at false positive rates between 0.5 and 16 by changing the classification threshold for the Mask R-CNN outputs. The experiments presented below all employ a modified version of the A3D architecture, using cascading R-CNNs' enhanced bounding boxes and segmentation mask prediction. We distinguish between the base augmentations $\text{Aug}_1 = \{$flip, shift, rescale, rotate$\}$ and our extension $\text{Aug}_2 = \text{Aug}_1 \cup \{$crop$\}$.

Using the original configuration provided in [17], we were not able to achieve the exact same results. Therefore, we compare our experiments with the replicated A3D results ($\sim \Delta 1\,\%$). As A3D is used universally throughout our experiments, we deem this to not have any effect on the relative changes in mean Free Response Operating Characteristic (mFROC) scores.

1) A3D + Cascade R-CNN + Aug_1*:* In this configuration, we substitute the Mask R-CNN component of the baseline [17] with the Cascade R-CNN and employ the set of augmentations (flip, shift, rescale, and rotate) introduced in [17]. The results show that this approach works better than the baseline when higher FPs (≥ 4) are considered.

2) A3D + HTC + Aug_1*:* Here, we replace the Mask R-CNN component with HTC and apply the original set of augmentations (Aug_1) [17] demonstrating an improvement of the sensitivity for all FPs except FP@16.

3) A3D + Cascade R-CNN + Aug_2*:* We observed that the augmentation with Random Cropping improves the mFROC by 0.25 percentage points compared to the original augmentations (random horizontal flip, shift, rescaling and rotation) [17].

4) A3D + Cascade R-CNN + Aug_2 *+ Pretraining:* Pretraining the initialised ImageNet weights on the Youtube VOS dataset allows the model to capture longer spatiotemporal dependencies. Accordingly, we further increase the mFROC by 0.55 percentage points compared to the *A3D + Cascade R-CNN +* Aug_2 model. The results for all FPs (except for FP@16) demonstrate the efficacy of using sequential, interconnected frames as a potential substitution for 3D imaging data within the pretraining process.

5) A3D + HTC + Aug_2 *+ Pretraining:* The culmination of previously mentioned methods leads to an overall increase of 0.78 mFROC.

6) A3D + HTC + Aug_2 *+ Pretraining + Random Spacing:* Utilising Random Spacing during pretraining we aim to make the model resilient towards variability in slice thickness. The results show an absolute increase of 1.04 mFROC.

7) A3D + HTC + Aug_2 *+ Pretraining (7 slices):* Also when a higher number of slices (7 instead of 3) are applied, our model outperforms the A3D base model at all FPs (except for FP@0.5).

V. CONCLUSIONS

We have introduced a novel pretraining and machine learning framework incorporating an A3D with HTC for universal lesion detection. We utilised the truncated DenseNet-121 backbone of A3D and pretrained it with videos from the YouTube VOS dataset [22]. We also employed Random Spacing as a second pretraining approach that adjusts the image thickness by randomly skipping 0 to 4 frames. Further, we applied Random Cropping for data augmentation. Finally, we have tested a set of combinations of all proposed methods and demonstrated their efficacy in improving the overall sensitivity. We believe that A3D is a highly promising approach also in terms of its adaptation to new components facilitating modular ablation studies.

VI. Limitations and Future Work

Currently, most Mask R-CNNs perform at high false positive rates, which restricts their real-world usage. It is sometimes necessary to choose a low classification threshold for high sensitivity in order to achieve high confidence scores for correct predictions. Since we pretrained on only one class of the YouTube VOS dataset, future work could make use of all available classes. Moreover, backbone optimisation, e. g., by swapping the A3D's backbone (DenseNet-121) with EfficientNet [25], or ResNeXt [26] could yield enhanced feature maps. A Feature Pyramid Network with feature map outputs at multiple stages can also be tested for further improving the detection of lesions of different sizes.

Zhang et al. [27] introduced Variable Dimension Transform (VDT) (this work was not available during the formation of this study), which seems to be a very promising approach. By using 2D semantic annotations, the proposed VDT [27] enables the learning of discriminative and invariant 3D feature representations for 3D medical imaging tasks.

For future research, personalised machine learning methods (e. g., [28]) should be considered for addressing the heterogeneity of medical data, characterised by different ages, genders, and physical characteristics of each individual.

References

[1] H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. ca cancer clin 2021; 71: 209-49," *CA Cancer J. Clin.*, vol. 71, pp. 209–249, 2022.

[2] D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M. Emberton, S. Esener, R. C. Fitzgerald, S. S. Gambhir, P. Kuhn, *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, p. eaay9040, 2022.

[3] M. Yun, W. Kim, N. Alnafisi, L. Lacorte, S. Jang, and A. Alavi, "18f-fdg pet in characterizing adrenal lesions detected on ct or mri," *Journal of Nuclear Medicine*, vol. 42, no. 12, pp. 1795–1799, 2001.

[4] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning," *Journal of Medical Imaging*, vol. 5, no. 3, pp. 1 – 11, 2018. [Online]. Available: https://doi.org/10.1117/1.JMI.5.3.036501

[5] T. G. Leighton, "What is ultrasound?" *Progress in biophysics and molecular biology*, vol. 93, no. 1-3, pp. 3–83, 2007.

[6] C. C. Jaffe *et al.*, "Measures of response: Recist, who, and new alternatives," *J Clin Oncol*, vol. 24, no. 20, pp. 3245–51, 2006.

[7] E. L. van Persijn van Meerten, H. Gelderblom, and J. L. Bloem, "Recist revised: implications for the radiologist. a review article on the modified recist guideline," *European radiology*, vol. 20, pp. 1456–1467, 2010.

[8] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?" *Insights into imaging*, vol. 8, no. 1, pp. 171–182, 2017.

[9] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[11] K. Yan, M. Bagheri, and R. M. Summers, "3d context enhanced region-based convolutional neural network for end-to-end lesion detection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I.* Springer, 2018, pp. 511–519.

[12] H. Li, J. Huang, G. Li, Z. Liu, Y. Zhong, Y. Chen, Y. Wang, and X. Wan, "View-Disentangled Transformer for Brain Lesion Detection," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Mar. 2022, pp. 1–5.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[15] J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, and B. Ni, "Reinventing 2d convolutions for 3d images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3009–3018, 2021.

[16] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.

[17] J. Yang, Y. He, K. Kuang, Z. Lin, H. Pfister, and B. Ni, "Asymmetric 3d context fusion for universal lesion detection," *Lecture Notes in Computer Science*, pp. 571–580, 2021. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-87240-3_55

[18] K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, and R. M. Summers, "Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9261–9270.

[19] H. Li, L. Chen, H. Han, and S. Kevin Zhou, "SATr: Slice Attention with Transformer for Universal Lesion Detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, ser. Lecture Notes in Computer Science, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 163–174.

[20] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[21] A. Ben-Cohen, E. Klang, A. Kerpel, E. Konen, M. M. Amitai, and H. Greenspan, "Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in ct examinations," *Neurocomputing*, vol. 275, pp. 1585–1594, 2018.

[22] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "Youtube-vos: A large-scale video object segmentation benchmark," *arXiv preprint arXiv:1809.03327*, 2018.

[23] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[24] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," 2017. [Online]. Available: https://arxiv.org/abs/1706.02677

[25] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[26] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[27] S. Zhang, Z. Li, H.-Y. Zhou, J. Ma, and Y. Yu, "Advancing 3d medical image analysis with variable dimension transform based supervised 3d pre-training," *arXiv preprint arXiv:2201.01426*, 2022.

[28] A. Kathan, S. Amiriparian, L. Christ, A. Triantafyllopoulos, N. Müller, A. König, and B. W. Schuller, "A personalised approach to audiovisual humour recognition and its individual-level fairness," in *Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge*. Lisbon, Portugal: ACM, 2022, pp. 29–36.