Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
DOI: https://doi.org/10.35414/akufemubid.1475807
AKÜ FEMÜBİD **24** (2024) 065105 (1382-1394)
AKU J. Sci. Eng. **24** (2024) 065105 (1382-1394)

DAC: Differentiable Auto-Cropping in Deep Learning

DAC: Derin Öğrenmede Türevlenebilir Oto-Kırpma

Makale Bilgisi / Article Info
Alındı/Received: 30.04.2024
Kabul/Accepted: 01.09.2024
Yayımlandı/Published:02.12.202

Ahmet Esad TOP[1*], Mustafa YENİAD[2], Mahmut Sertaç ÖZDOĞAN[3], Fatih NAR[2]

[1] *ASELSAN, Corporate Management Directorate, IT Department, Ankara, Turkey*
[2] *Ankara Yıldırım Beyazıt University, Department of Computer Engineering, Ankara, Turkey*
[3] *Ankara Yıldırım Beyazıt University, Department of Prosthodontics, Ankara, Turkey*

Öz

Bir görüntünün sınırlarını ilgi alanına odaklanacak şekilde otomatik olarak ayarlama işlemi olan oto-kırpma, panoramik diş radyografilerinin teşhis kalitesinin iyileştirilmesi açısından çok önemlidir. Önemi, minimum bilgi kaybıyla farklı girdi görüntülerinin boyutunu standartlaştırma yeteneğinde yatmaktadır, böylece tutarlılık sağlanmakta ve sonraki görüntü işleme görevlerinin performansı iyileştirilmektedir. Çalışmaların birçoğunda CNN'ler yaygın olarak kullanılmasına rağmen, farklı boyutlardaki görüntüler için oto-kırpma kullanan araştırmalar sınırlı kalmaktadır. Bu çalışma, panoramik diş radyografilerinde türevlenebilir oto-kırpma kullanmanın potansiyelini araştırmayı amaçlamaktadır. Çalışmada, çoğunlukla 2836×1536 veya buna yakın çözünürlüklü, 3 diş hekimi tarafından beş farklı sınıfa bölünmüş 20.973 panoramik diş radyografisinden oluşan benzersiz bir veri kümesi kullanıldı; bu, önceki çalışmadaki aynı veri kümesidir (Top et al. 2023). Değerlendirme için bu veri kümesine en başarılı sonucu veren ResNet-101 modeli kullanıldı (Top et al. 2023). Varyansı azaltmak için, hem oto-kırpma olan hem de oto-kırpma olmayan eğitimlere 10 kat çapraz doğrulama kullanılarak model değerlendirildi. Daha doğru ve sağlam sonuçlara ulaşmak için veri artırma yöntemi de kullanıldı. Veri artırma, oto-kırpma olan eğitim için, oto-kırpma olmayan eğitime göre çok daha az etkili olacak şekilde ayarlandı. Veri kümesiyle ilgili sorunları azaltmak için geliştirilen önerilen oto-kırpma optimizasyonu sayesinde doğruluk %1,8 artarak %92,7'den %94,5'e çıktı. Makro ortalama AUC'si de 0,989'dan 0,993'e yükseldi. Önerilen oto-kırpma optimizasyonu, uçtan uca bir CNN'de eğitilebilir bir ağ katmanı olarak uygulanabilir ve diğer problemler için de kullanılabilir. Doğruluğu %92,7'den %94,5'e çıkarmak, iyileştirme için çok az alan kaldığından, azalan faydalar kanununa da bağlı olarak çok zorlu bir iştir. Sonuçlar, önerilen türevlenebilir oto-kırpma algoritmasının potansiyelini göstermekte ve farklı alanlarda kullanımını teşvik etmektedir.

Anahtar Kelimeler: *Bilgisayar destekli teşhis; CNN; Türevlenebilir kırpma; Gradyan yükselme; Panoramik radyografi.*

Abstract

Auto-cropping, the process of automatically adjusting the boundaries of an image to focus on the region of interest, is crucial to improving the diagnostic quality of dental panoramic radiographs. Its importance lies in its ability to standardize the size of different input images with minimal loss of information, thus ensuring consistency and improving the performance of subsequent image-processing tasks. Despite the widespread use of CNNs in many studies, research on auto-cropping for different-sized images remains limited. This study aims to explore the potential of differentiable auto-cropping in dental panoramic radiographs. A unique dataset of 20,973 dental panoramic radiographs, mostly with a resolution of 2836×1536 or close, divided into five classes by 3 dentists, was used, which is the same dataset from the previous study (Top et al. 2023). ResNet-101 model, which was the most successful network for the dataset (Top et al. 2023), was used for the evaluation. To reduce variance, the model was evaluated using 10-fold cross-validation for both non-auto-cropped and auto-cropped trainings. Data augmentation was also used to produce more accurate and robust results. For auto-cropped training, it was adjusted to be much less effective than the non-auto-cropped one. Accuracy was improved by 1.8%, from 92.7% to 94.5%, thanks to the proposed auto-crop optimization developed to reduce dataset-related issues. Its macro-average AUC was also raised from 0.989 to 0.993. The proposed auto-crop optimization can be implemented as a trainable network layer in an end-to-end CNN and can be used for other problems as well. Increasing the accuracy from 92.7% to 94.5% is a very challenging task due to diminishing returns, as there is little room for improvement. The results show the potential of the proposed differentiable auto-crop algorithm and encourages its use in different fields.

Keywords: *Computer aided diagnosis; Convolutional Neural Networks; Differentiable cropping; Gradient ascent; Panoramic radiograph.*

1. Introduction

There has been a significant transformation in the computer vision society in recent years, mostly driven by the widespread acceptance and application of Convolutional Neural Networks (CNNs) (LeCun et al. 1998). The CNNs have proven to be a fast, adaptable, scalable, comprehensive, and end-to-end learning approach that moves the field forward (Jaderberg et al. 2015). Input limitations on size are present in many current Neural Network (NN) models used in computer vision (He et al. 2016, Krizhevsky et al. 2012, Simonyan and Zisserman 2014, Zeiler and Fergus 2014). However, the conventional strategy is uniformly downsampling for

*Sorumlu Yazar/Corresponding Author: Ahmet Esad TOP e-posta/e-mail: ahmetesadtop@aselsan.com

the intended input size, which is lower than the images in the dataset (Recasens et al. 2018). In many circumstances, uniform downsampling is straightforward and efficient, but it can be lossy for tasks when the aspect ratio is crucial and the RoI projection is on a small portion. Significant advancements have been made in Computer-Aided Diagnosis (CAD) recently (Demir et al. 2023, Fidan et al. 2019, Katsumata 2023, Kemal and Kılıçarslan 2021, Kohinata et al. 2023, Yurttakal and Baş 2021). Manually extracted features along with learning methods were used in traditional CAD systems, including pattern recognition algorithms, to transform them into decisions (Shin et al. 2016). The evolution from rule-based to learning-based solutions was realized by Artificial Intelligence (AI) and further advanced by Deep Learning (DL) (Çelik and Çelik 2022, Kooi et al. 2017). This study applied the proposed auto-cropping method on a unique dental radiography dataset (Top et al. 2023) in a DL training for a CAD system.

When a few teeth are missing from the mouth, restorations like dental bridges and crowns are utilized to fill the gap (Sakaguchi and Powers 2012, Top 2023). Radiographs are an essential diagnostic tool since these restorations are not usually evident during a conventional clinical examination (Top et al. 2023, White et al. 2001). Restorations must be precisely detected and identified to stop several problems in oral health (Liedke et al. 2015, Top 2023). Panoramic X-ray is a useful radiography type and a valuable diagnostic tool for dental and oral health disorders as they provide a complete and exhaustive image (i.e., a full image of the whole jaw and teeth in a single image) of the oral anatomy (Corbet et al. 2009, Scarfe and Farman 2008).

Although panoramic radiography is a very valuable diagnostic tool, it is prone to significant and unforeseen geometric anomalies. Furthermore, the specific jaw curvature and posture of patients may result in differences in the generated image (Choi 2011, Top 2023). The NN training of panoramic radiographs presents a unique set of difficulties due to their comprehensive nature, capturing not only the teeth but also the chin, spine, and jaws, as mentioned in (Jader et al. 2018). As a result, it becomes difficult to set just one fixed position that applies to the entire dataset for the intended Region of Interest (RoI). One approach to overcome this difficulty is to extract the RoI from the image automatically. Thus, auto-cropping eliminates the problem of image variability and allows the model to focus on the most relevant parts of the images, which is crucial for improving model accuracy and diagnostic performance.

Additionally, while the dataset is extensive, data augmentation is necessary to enhance the robustness and generalizability of deep learning models. Data augmentation techniques, such as random rotations, shifts, and flips, introduce variability that helps the model learn to generalize better to unseen data. This is particularly important for reducing overfitting and improving the model's performance on the test set by simulating real-world uncertainties and variations.

This study aims to uncover the potential of the previously studied unique dataset (Top et al. 2023) by using differentiable auto-crop optimization, which is utilized in CNNs to detect the quantitative level of dental restorations. We propose a novel Differentiable Auto-Cropping (DAC) technique for automatically selecting the bounding box and then cropping the RoI by tweaking the gradient ascent optimization independently for each input picture. The proposed auto-crop algorithm was developed with the dataset (Top et al. 2023) in (Top 2023) to alleviate dataset-related issues. Hence, it aims to improve the training performance of the dataset by using auto-crop.

2. Related Work

In computer vision systems, auto-cropping is essential (Chen et al. 2016) because it allows for the autonomous separation of meaningful sections from images. A few methods have been proposed to address the challenge of end-to-end image cropping or downsampling after localizing (Dai et al. 2016a, Han et al. 2019, Jaderberg et al. 2015, Liang et al. 2022, Recasens et al. 2018, Riad et al. 2022, Rippel et al. 2015). In recent studies, Spatial Transformer Networks (STNs) (Jaderberg et al. 2015) were employed for image transformation and cropping operations, and they were usually trained in two stages. However, some end-to-end systems are also available, such as (Liang et al. 2022), which employed two stages. Locating the vertebrae and segmenting the entire spine were done in the first stage. Then, a regression network predicted the orientations of the localized vertebrae in the second stage. A differentiable cropping was employed for moving data between stages that were intended to closely connect both phases, and this was called as "inter-stage transfer method" in the study. The approach introduced in the study, which includes two steps with the localization of several bounding boxes inside a single image for cropping and sending to the regression network, cannot be applied to other tasks as it is designed as task-specific. It is based on supervised learning and requires ground-truth bounding boxes for training, but our study can operate under unsupervised learning as

well as supervised learning and does not require ground-truth bounding boxes for training. Additionally, the RoI Align pooling layer (He et al. 2017) was employed to differentiate the cropping process, as this layer gathers the information from the outputs of localization and segmentation networks rather than using the raw input. However, our proposed method works on the raw input.

To reduce computational complexity and provide some shift-invariance, CNNs frequently use a variety of downsampling operators, such as strided convolutions or pooling layers, that gradually reduce the resolution of intermediate features (Riad et al. 2022). According to (Riad et al. 2022), the first downsampling layer with differentiable strides was proposed by employing both the spatial and frequency domains, much like (Rippel et al. 2015). As opposed to (Rippel et al. 2015), which optimizes a fixed bounding box updated by a stride parameter, their solution employs backpropagation to learn the size of the bounding box. Our study is similar to (Riad et al. 2022), as it also determines the bounding box area itself. Additionally, (Dai et al. 2016a) presented one of the major advances in differentiable semantic segmentation. They created a differentiable pooling layer known as the "RoI Warping Layer" that uses interpolation on the feature maps of the preceding layer and reduces the size for the next layer. Instead of downsampling the input image right after the input layer, all three strategies (Dai et al. 2016a, Riad et al. 2022, Rippel et al. 2015) focused on intermediate layers. However, our proposed method downsamples (i.e., crops) the input image that is fed to the NN in the first place. (Han et al. 2019) conducted a study to develop a specialized layer (i.e., for detecting people) that can transform RoI in a differentiable and supervised manner. This layer aims to identify specific objects (i.e., individuals) from the input data. It is important to note that this approach relied on supervised learning and was primarily designed for object detection tasks. The detector employed in the approach was trained alone using state-of-the-art techniques (Dai et al. 2016b, Girshick 2015, He et al. 2017, Ren et al. 2015) and detecting the bounding boxes required supervised learning. The weakness of the need for supervision is the point that inspired this study, as our proposed method does not require such supervision.

The end-to-end auto-cropping applicability logic described in this study bears a striking resemblance to the end-to-end concept implemented by (Recasens et al. 2018). A saliency-based non-uniform distortion layer for CNNs that improves spatial sampling for specific tasks was introduced. This layer integrates seamlessly into existing networks and enhances task performance by selectively preserving important information from high-resolution data. However, it should be emphasized that our auto-cropping technique uniformly downsamples the input, and therefore the employed approaches are different. In their groundbreaking work, Jaderberg et al. introduced the notion of spatial transformers, a learnable module that empowers CNNs with the ability to manipulate data spatially (Jaderberg et al. 2015). Explicit transformations of feature maps, including scaling, rotation, translation, and even non-rigid deformations, can be done by the module. A spatial manipulation inside a CNN was accomplished by both the STNs and our algorithm but with a different focus. Our algorithm wasn't built to achieve broad (general) spatial invariance; instead, it was made for RoI selection and cropping. Moreover, a differentiable pre-processing stage (i.e., $max - min$ pooling) and the optimization part were brought together to feed downsampled and clarified inputs (i.e., in terms of regional variability (Liu et al. 2020)) to the optimization in our algorithm.

Jiang et al. described a major drawback of bilinear interpolation, which is quite localized by considering only the four nearest pixels (i.e., causes gradients to be affected only by the intensity differences between these nearby pixels) (Jiang et al. 2019), which was the essential part of the spatial transformer module (Jaderberg et al. 2015). In an effort to alleviate this limitation, (Jiang et al. 2019) introduced an approach that transforms the sampling part through the inclusion of randomly generated auxiliary sample locations (i.e., by making bilinear sampling on those locations later on), thereby achieving a more linear approximation. Despite providing a broader context for local transformations, the optimization remains unaffected by all pixels in the input image. In contrast, our proposed auto-crop optimization takes every pixel into account, some with minimal impact (but never completely ignored), as observed in the two related studies.

3. Materials and Methods

3.1. Dataset

The study utilized a retrospective dataset (Top et al. 2023) from Ankara Yıldırım Beyazıt University (AYBU) Tepebaşı Oral and Dental Health Training Hospital, which was conducted in accordance with approved protocols by the AYBU non-drug ethics committee with the permission date of 19/04/2019 (file number: 2019-12). The study followed the criteria established in the Helsinki Declaration of 1964 as well as the updated version from 2013.

Figure 1. A sample from each class (Top et al. 2023)

Table 1. The classes and details (Top 2023, Top et al. 2023)

# of Restorations	# of Samples (%)	Category Name
0	4248 (20.25%)	No restoration
1-3	4231 (20.17%)	Low-level-restoration
4-6	4227 (20.16%)	Mid-level-restoration
7-11	4253 (20.28%)	High-level-restoration
12-32	4014 (19.14%)	Very High-level-restoration

The collection contains 20,973 panoramic radiographs in total, and each one is unique even if taken from the same person, pose angle, exposure time, different contrast and brightness values, and passed time between the shots leads this uniqueness). The dataset was obtained retrospectively from the Tepebaşı Oral and Dental Health Training Hospital's Picture Archiving Communication System (PACS) using the Infinitt PACS, developed by Infinitt Co. in Seoul, Korea. Filming was done with two different Planmeca ProMax X-ray devices (Planmeca, Helsinki, Finland). Three dentists labeled the dataset into five classes in consensus according to fixed dentures that included crowns and bridges. Details about the categories and number of samples can be seen in Table 1. Also, a sample from each class is depicted in Figure 1.

The dataset does not consist of single-resolution images, but there are two sets of resolution, most of which are grouped there. 44% of the total images have a resolution of 2836×1536 and 37% of the total images have a resolution of 2860×1536. 17% of the images were stored in 16, 24, or 32-bit format, and they were converted into 8-bit depth grayscale images. The proposed auto-crop method finds the RoI and resizes the image to a desired size (i.e., using bi-linear interpolation and preserving the aspect ratio of the original input), so the variable resolution of the dataset is not so important. In other words, the proposed method somehow brings scale-invariance.

As mentioned, during the pre-processing phase, the images were converted to single-channel grayscale. For non-auto-cropped training, they are resized to 224×224 pixels (i.e., the aspect ratio was corrupted). In contrast, the experiment using the auto-crop algorithm reproduces the inputs at a fixed resolution of 200×370 pixels (i.e., according to the desired aspect ratio and resolution of the input images), where the width is 370px; so no additional resizing is needed.

Apart from these, image augmentation settings tried to be defined not to lose parts of the RoI (see Figure 2) for non-auto-cropped experiments. However, since each person and each X-ray shot is unique and there is no single RoI location for all, these settings still apply an excessive augmentation for some samples. Translating and rotating without knowing the midpoint of the RoI causes problems such as the appearance of no-data (black) regions as a result of excessive augmentation, which reduces the efficiency of the data set as very large regions are no-data.

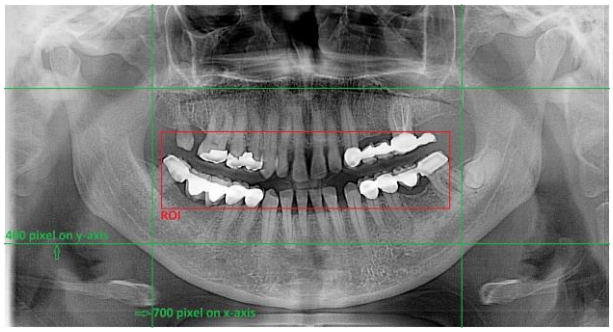


Figure 2. The image translation process and RoI (Top et al. 2023)

The auto-crop algorithm downsampled images into a single and desired resolution (i.e., 370×200), so the setting for the augmentation needed to be changed. Also, the reproduced dataset has lost some tissues (i.e., explained in the first section) other than the oral region, so the excessive augmentation (i.e., in terms of percentage) would damage the accuracy a lot. Therefore, a reasonable (i.e., not excessive) augmentation (i.e., random translation between -26 to 26 pixels on the x-axis and -14 to 14 pixels on the y-axis, random reflection on the x-axis, random rotation with an angle between -7 to 7 degrees, random scaling with a factor between 0.97 to 1.03, and random shearing over an angle between 0 to 2 degrees) was employed for the auto-cropped experiment.

In an X-ray film, teeth and bones appear lighter than cheeks and gums due to the absorption of the X-rays (Fitzgerald 2000, Top 2023). Similarly, dental restorations like crowns and bridges appear lighter than the teeth as they are radiopaque (Pröbster and Diehl 1992, Top et al.

2023). This feature makes the grayscale image of dental restorations learnable by CNN and detectable by the auto-crop algorithm (i.e., not just restorations but teeth as well).

3.2. Auto-Crop

The proposed auto-crop algorithm relies on gradient ascent optimization to maximize the cost function. This approach includes the dot product between a discrete signal (i.e., image) and a continuous function (i.e., rectangular function).

The features of the intended task are located in the oral region of the individual, which varies from person to person. Also, the exclusion of other tissues present in the panoramic radiograph is essential for enhancing the efficacy of the training process. In this context, the most practical approach is to crop the region from the images through an automatic procedure. To address this need, an auto-crop optimization was implemented to effectively find and crop the RoI from the images.

The process of evaluating the gradient by conducting a dot product between an image (i.e., discrete) and a continuous function is called numerical estimation of the derivative or convolution. In the inner product method, the derivative to be approximated is initially represented by the constructed continuous function. Subsequently, this function is sampled (i.e., discretized) to align with the sampling rate of the image, and the dot product can then be computed. The algorithm continuously updates the parameters to determine the RoI position. The necessity of "highlighting the prominent regions and preventing the less important ones" feature in the rectangular function has led to the use of the 2D Sigmoid function (see Figure 3), which can be described as:

$$\sigma(x, y, \gamma) = \frac{1}{1+e^{-x\cdot\gamma}} \cdot \frac{1}{1+e^{-y\cdot\gamma}} \quad (1)$$

where the default value of the sharpness (γ) is 1 as in Figure 3.

To search the RoI, rectangular area parameters (i.e., the center coordinates for the x and y axes, and a distance from the center) should be defined, and accordingly, the required function should be in the form of a rectangular function. Therefore, the rectangular function using 2D sigmoids is defined as:

$$R(x, y, x_c, y_c, r_{total}, \alpha, \gamma) = A \cdot B$$
$$A = \sigma(x - (x_c - r_{total}), y - (y_c - \alpha \cdot r_{total}), \gamma) \quad (2)$$
$$B = \sigma(-x + (x_c + r_{total}), -y + (y_c + \alpha \cdot r_{total}), \gamma)$$

where α is the aspect ratio ($height/width$), x_c and y_c are center coordinates of the region, r_{total} is the center-to-edge distance for the x-axis and '$\alpha \cdot r_{total}$' is for the y-axis, and γ is for controlling the smoothness (or sharpness) of the rectangular function. The default value of the sharpness (γ) is 1, but see Figure 4 for how the rectangular function changes when γ is changed (i.e., decreasing gamma increases the smoothness of the rectangular function, and increasing gamma results in a sharper function).

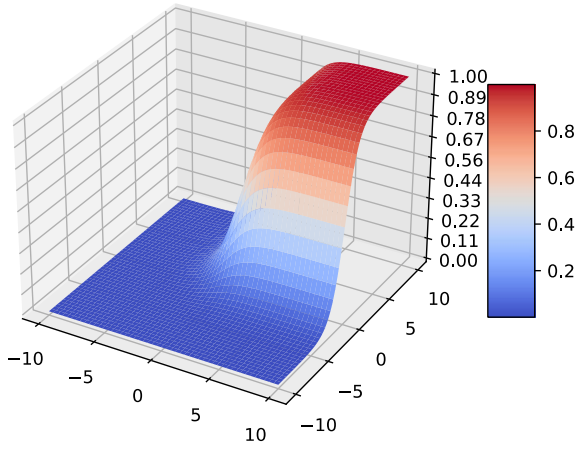
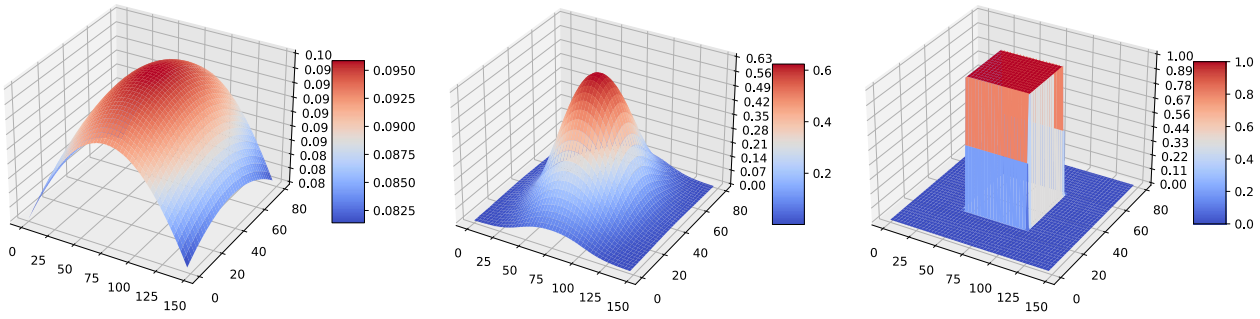


Figure 3. 3D plot of the sigmoid (Top 2023)



(a) For the sharpness (γ) $= 0.01$ (b) For the sharpness (γ) $= 0.1$ (c) For the sharpness (γ) $= 10$

Figure 4. 3D plots of the rectangular function for different γ values, where the function sampled for a 148×80 image and $(x_c,\ y_c) = (74, 40)$ with $r = 30$ (Top 2023)

A rectangular function of the same size as the input image, known as a full-size rectangular function, is employed for the optimization. The cost function utilizes a function derived from this rectangular function, not the identical one. The improved cost function is as following:

$$f(x_c, y_c, r, r_{min}, s, x, y, \alpha, \gamma) = \frac{f_{inside}}{f_{outside}+1}$$

$$f_{inside} = \frac{\sum_0^{n-1} s_i z_i}{\sum_{i=0}^{n-1} z_i} \quad (3)$$

$$f_{outside} = \frac{\sum_{i=0}^{n-1} s_i(1-z_i)}{\sum_{i=0}^{n-1} 1-z_i}$$

where z_i is the rectangular function (R) explained above and $r_{total} = r_{min} + r$.

Multiplying the input image by the rectangular function yields f_{inside}'s top portion, while the discretized rectangular function makes up the bottom portion of f_{inside}, which serves as a penalty to curb the increase of offset r forever. During the gradient ascent process, f_{inside} aims to maximize the desired inner portion of the rectangular function, while $f_{outside}$ aims to minimize the undesired outer part of the rectangular function (i.e., helps maximize f_{inside} in return). The inclusion of 1 in the denominator of the cost function (f) serves a purpose: when $f_{outside}$ falls within the range of 0 to 1 and approaches 0, the cost function, which we aim to maximize, goes towards infinity. Adding 1 ensures that the maximum value of the cost function that it can reach is at f_{inside}.

In the backward pass, the procedure calculates the gradients of x_c, y_c, and r (i.e., RoI parameters) iteratively through the dot product of the image and the rectangular function, which emphasize those with a strong resemblance. The algorithm updates the RoI parameters based on the gradients. This seeks to maximize the value of the cost function (i.e., better RoI selection) and is expressed as:

$$\underset{x_c, y_c, \mathrm{r}}{\operatorname{argmax}} f(x_c, y_c, r, r_{min}, s, x, y, \alpha, \gamma) \quad (4)$$

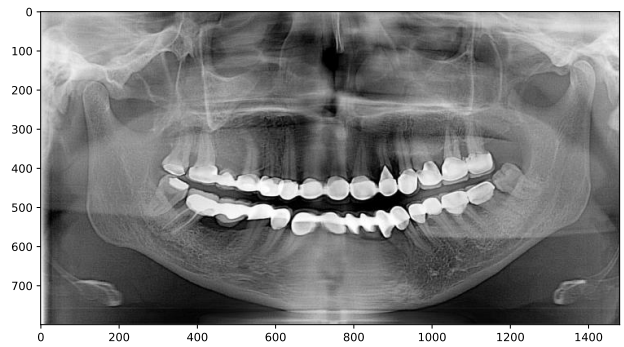
The procedure stops when convergence (i.e., when the vanishing gradient problem occurs or reaches a stable RoI position) is reached. Several preventive mechanisms, such as gradient clipping, center shifting, and the use of absolute values, were employed to avoid problems during convergence such as exploding gradient problem or RoI falling outside of the image after the update.

Before beginning the search, a differentiable pre-processing procedure (i.e., subtracting the min-pooling from the max-pooling) was executed on the input. The process inherently has differentiability due to the differentiability of max pooling. The min-pooling is simply the inverse version of max-pooling, so it is also differentiable. The operation can be stated as below:

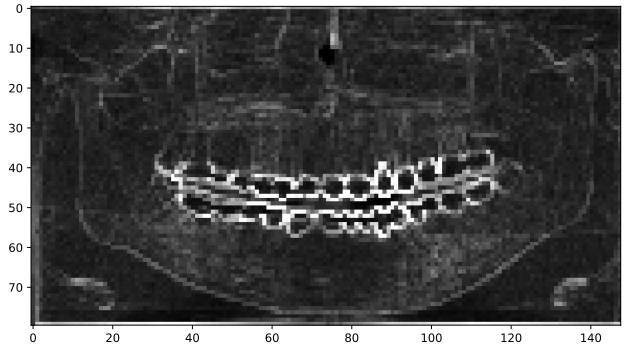
$$maxMinDif(x) = maxPool(x) - minPool(x) \quad (5)$$
$$minPool(x) = -maxPool(-x)$$

Instead of simply utilizing filtering, pooling processes were adopted to downsample the image, making edges more perceptible and reducing the resolution. Downsampling was desired to reduce the computation time of the following processes. The generated image (see Figure 5) was fed into gradient ascent optimization, making major variations in regional characteristics more evident (Liu et al. 2020) across all teeth.

PyTorch (Paszke et al. 2019) was used to implement the entire auto-crop process, leveraging its capabilities to efficiently run operations such as maximum pooling and dot product through the Torch library. PyTorch was considered ideal for providing end-to-end smart cropping, primarily because of its automatic gradient calculation capabilities on tensors (i.e., 2D tensors for our case).

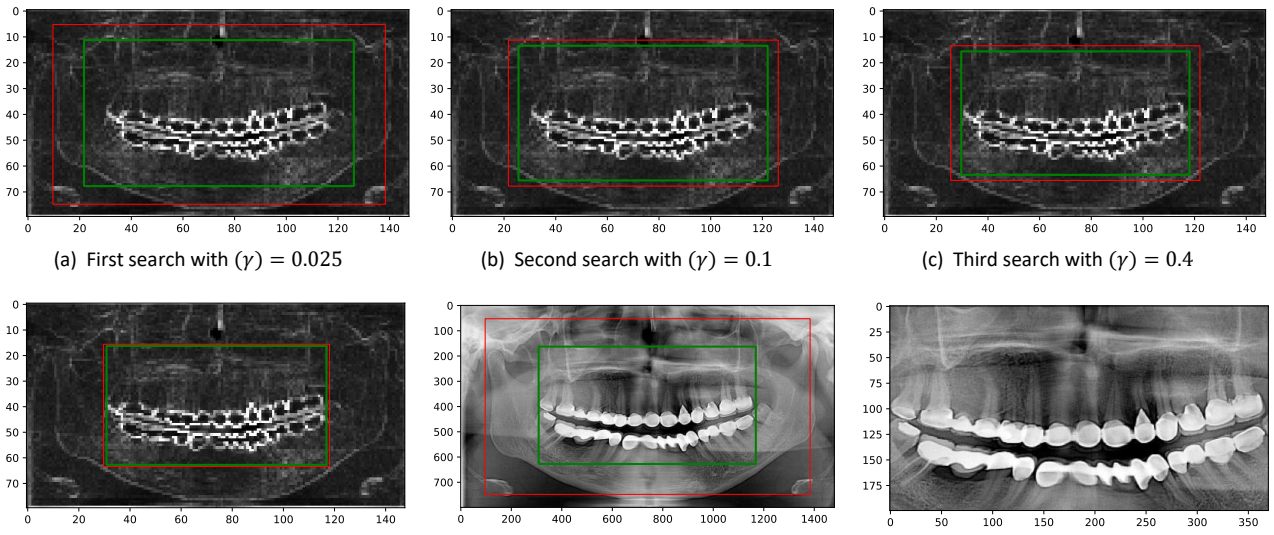


(a) The original input with a resolution 1480×800



(b) The generated image with a resolution 148×80

Figure 5. Result of $max - min$ pooling where the resolution was reduced 10x (i.e., window size and stride was 10) (Top 2023)



(a) First search with $(\gamma) = 0.025$ (b) Second search with $(\gamma) = 0.1$ (c) Third search with $(\gamma) = 0.4$

(d) Fourth search with $(\gamma) = 1.6$ (e) The initial & last bounding boxes (f) Cropped & resized result

Figure 6. Representation of the initial (i.e., red) and optimized (i.e., green) bounding boxes in the input image (Top 2023)

Algorithm 1. Adaptive sharpness γ and step size η

Input: γ, η and input image as s

Output: output image

1. Get $width$ and $height$ hence α information from image s
2. $s \leftarrow maxMinDif(s)$
3. $x, y \leftarrow meshgrid([0 \dots \text{width} - 1], [0 \dots height - 1])$ //$meshgrid$ returns a list of coordinate matrices from coordinate vectors
4. $x_c \leftarrow width/2$
5. $y_c \leftarrow height/2$
6. $r_{min} \leftarrow width/3.5$
7. $r \leftarrow (width/2.3) - (width/3.5)$
8. **for** $i \leftarrow 1$ to 4 **do**
9. $x_c, y_c, r \leftarrow \underset{x_c, y_c, \mathrm{r}}{\operatorname{argmax}} f(x_c, y_c, r, r_{min}, s, x, y, \alpha, \gamma)$
10. $\gamma \leftarrow \gamma * 4$
11. $\eta \leftarrow \eta/2$
12. **end for**
13. Crop the found RoI and resize it to 370×200
14. Save output image

Fine-tuning by increasing the sharpness (γ) and decreasing the step size after some rough searching was held to find the ideal solution. Sharpness started at 0.025 and the last fine-tuning value was 0.2, and the learning rate started at 96 and dropped to 12. Figure 6 depicts two rectangles (i.e., one for initial RoI and one for post-search) on the input image, allowing a visual inspection. As can be seen, when the sharpness was low (smoother rectangular function) and the step-size was high, the bounding box moved too much at the rough search (Figure 6a). On the contrary, fine-tuning stages (Figure 6b, Figure 6c, and Figure 6d) searched in small steps in an adaptive way. The fine-tuning algorithm can be seen in Algorithm 1.

The parameters of the rectangular function (x_c, y_c, and r_{total}) were not subject to the random initialization. Instead, a conscious approach was adopted that ensures a consistent starting point. As mentioned earlier, the (x_c, y_c) pair are center coordinates and hence initialized from the center. Also, the center-to-edge distance (r_{total}) was initialized in a way that the bounding box of the RoI was positioned close to the edges of the image. Then they were optimized by gradient ascent and found the appropriate RoI.

3.3. Training

AlexNet (Krizhevsky et al. 2012), VGG-16 (Simonyan and Zisserman 2014), ResNet-18, ResNet-50, ResNet-101 (He et al. 2016), and Inception ResNet V2 (Szegedy et al. 2017) were tested on the non-auto-cropped dataset in (Top 2023, Top et al. 2023). These networks were chosen based on previous performance improvements on the ImageNet dataset in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015). ResNets can still be considered one of the most successful networks around (Top et al. 2023). The selection also took into account the dataset's size, as it is a large dataset, and networks with a large number of parameters can exploit the full potential of the data. However, deep networks tend to face the vanishing gradient problem, and residual blocks have provided a solution to this problem (He et al. 2016). Accordingly, ResNet101 (see Figure 7) produced the best results at the time, so it was used in the auto-cropped trial.

The original network architecture underwent some minor modifications, such as reducing the input size from 3 channels to 1 channel (since the dataset consists of grayscale images) and the resolution from 224×224 to 200×370 (i.e., the input size could have been increased more but the limitation was the Video Random-Access Memory (VRAM) of the Graphics Processing Unit (GPU)), or decreasing the size of the classification layer (i.e., fully-connected layer) from 1000 to 5 as there are 5 categories available for the task.

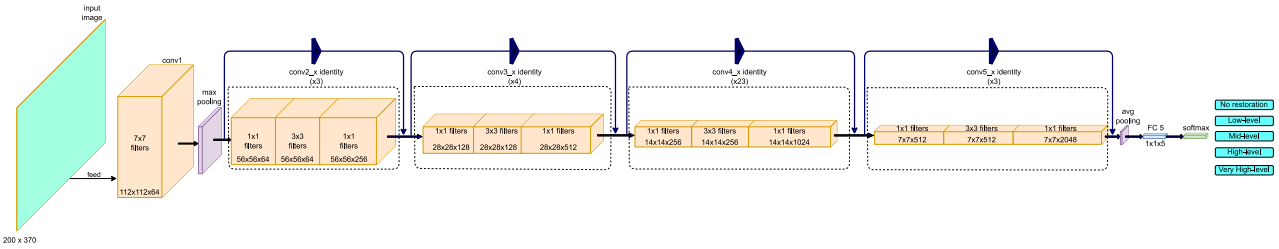


Figure 7. Representation of the structural design of the modified ResNet-101 architecture (Top 2023)

Transfer learning was not applicable because the dataset for the study consisted of grayscale images. The training aims to identify dental restorations (i.e., the objects it contains are different types of teeth, cheeks, gums, and bones, all present at the same time). The source domain of pre-trained networks is ImageNet (Russakovsky et al. 2015), and its purpose is to classify objects such as tigers, airplanes, or pineapples, so each model is trained from scratch.

We used 10-fold cross-validation to accurately assess model performance (by reducing variance) and reduce bias. This procedure ensured consistent and generalized results, where the average accuracy of the ten different trainings was used in the final evaluation. The "Adam" solver along with a batch size of 24 (i.e., to be able to fit the data into the VRAM of the GPU), and 1e-3 learning rate were the parameters of the training. A trial-and-error method was used for determining the best number of epochs. However, before we finished the development of the auto-crop algorithm, we tentatively trained and tested the idea (i.e., ad hoc) and found that 144 epochs solved the problem best, so when we finished development, we only used a 10-fold training with 144 epochs.

Since auto-cropping can deliver the unchanged aspect ratio, we modified the input size of ResNet101 as mentioned above (i.e., a 65.2% increase from 224×224 to 200×370), which significantly increased the total number of parameters in the network. The reason why the extra number of epochs is needed compared to the non-auto-cropped dataset is that the input layer has higher resolution and needs more information extraction or a deeper network.

4. Results and Discussion

For results, the average of the 10-fold cross-validation approach is reported, as explained in the previous section. On a single machine, the tests were carried out using a single Intel® Core™ i7-5960X CPU, 32 GB of Random Access Memory (RAM), and a single 8 GB NVIDIA® Quadro® M4000 GPU. The hardware capabilities help to decrease training time, but they are not necessary for practical usage in a clinical context. Even today's standard hardware (without a GPU) is adequate to run a test on the trained network. This highlights the method's high suitability for clinical applications, thanks to its significant results, robustness, and rapid execution time. Table 2 displays the average accuracies for these two approaches; DAC indicates that it is trained with our proposed method.

Table 2. Average accuracy results of 10-fold cross-validation and their corresponding number of epochs

Network (training type)	# of Epochs	Avg. Accuracy
ResNet-101 (non-DAC)	48	90.5%
ResNet-101 (non-DAC)	72	91.8%
ResNet-101 (non-DAC)	96	92.7%
ResNet-101 (non-DAC)	112	92.2%
ResNet-101 (DAC)	144	94.5%

Figure 8 shows confusion matrices to see accuracy results (i.e., precision and recall values can also be observed). The accuracy results indicate top-1 accuracy of the average of 10-folds. To conduct a more comprehensive performance comparison of the models, the Receiver Operating Characteristic (ROC) curve was generated by varying the threshold for each class (see Figure 9). Additionally, the Area Under the ROC Curve (AUC) was calculated for each class (see Figure 11), along with their macro-averages (see Figure 10). Log-scale on x-axis was used in Figure 9 and Figure 11 to illustrate better.

The highest accuracy achieved was 92.7% for non-auto-cropped training (i.e., underfitted before the 96 epochs and overfitted at the 112 epochs) and 94.5% for auto-cropped training, and the macro-average AUC scores were 0.989 and 0.993, respectively. The auto-cropped training showed a clear dominance over the non-auto-

cropped training in every aspect, as it is more successful in terms of accuracy, precision, recall, AUC for all classes, and macro-average AUC.

The classes "0", "7-11", and "12-32" tend to be simpler for the networks to learn, but the other classes "1-3" and "4-6" appear to be more difficult.

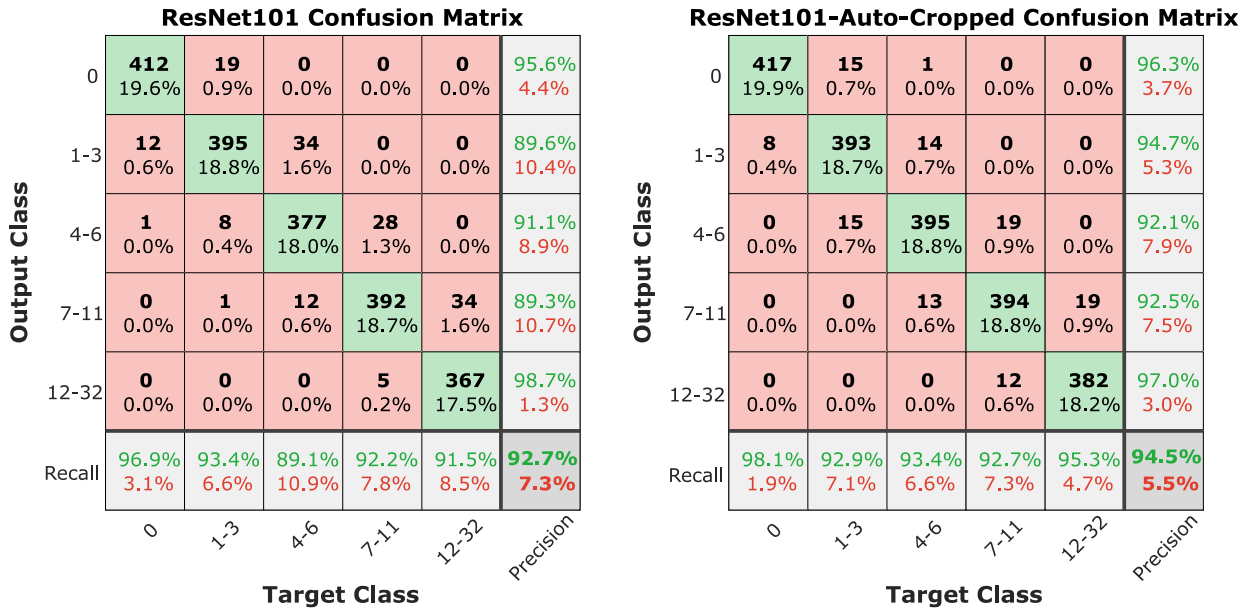


Figure 8. Confusion matrices for both training

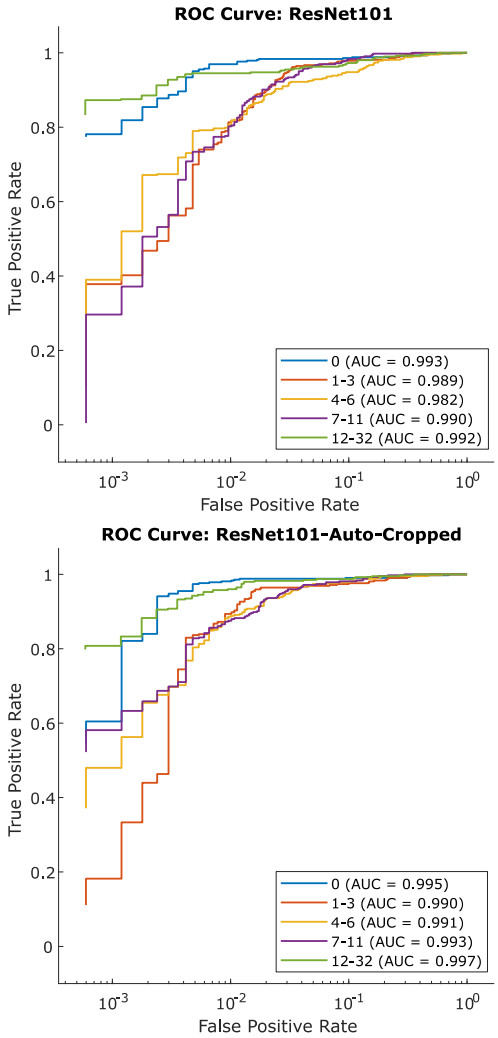


Figure 9. ROC curves present ability to separate positive from negative samples

The implementation of 10-fold cross-validation, which caused high computational costs during each training session, was found to be time-consuming throughout the training procedure. Additionally, fitting both the training data and all network parameters into the GPU's VRAM remained difficult, often requiring batch size reduction. Also, VRAM size has limited the increase of input size (i.e., we could have increased the input size more than 200×370). For these reasons, having access to a more powerful GPU with additional VRAM capacity might have helped with these tests. However, resizing a set of images of different sizes to a fixed size (i.e., 200×370) to fit in VRAM, and losing less information by finding RoI is possible thanks to our proposed auto-cropping algorithm.

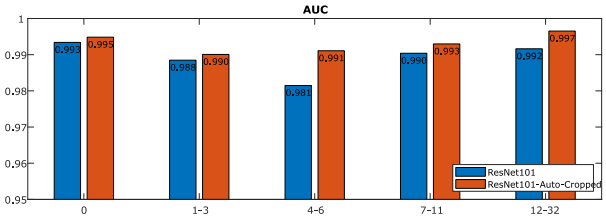


Figure 10. A bar comparison of the AUC scores in each class

Despite the limitations established on the augmentation parameters for non-auto-cropped training, excessive augmentation resulted in the loss of certain features. However, the use of auto-crop enabled the removal of noisy data, making extreme augmentation adjustments unnecessary.

Auto-cropping took advantage of changing the input size with a significant resolution increase of 65.2% (i.e., from

224×224 to 200×370). Using the higher resolution with auto-crop resulted in better accuracy, as the algorithm can identify and crop the region that holds the required features and reduces information loss that occurs with traditional resizing methods that preserve irrelevant features (i.e., the resulting image by resizing the original input contains parts not relevant to the classification task and loses more data compared to resizing after cropping). The cropping and resizing operations were accomplished without distorting the aspect ratio, which had previously been an issue when resizing to 224×224.

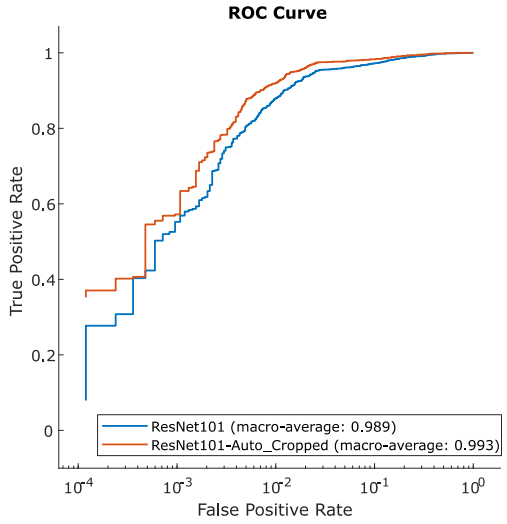


Figure 11. The macro-average AUC scores

The auto-crop operation improved the accuracy and produced more consistent results. Because the performance values of each fold at 10-fold cross-validation for auto-cropped training were closer to each other, while non-auto-cropped trainings were more unstable. Consistent results came into play because of the previous excessive augmentation approach, which involved noisy or unnecessary parts of the image.

5. Conclusion and Future Work

We have introduced an end-to-end differentiable optimization algorithm that finds the RoI with gradient ascent and crops it without corrupting the aspect ratio. We have demonstrated empirically that our method can provide better training performance than non-cropped training, where accuracy significantly increased from 92.7% to 94.5%. This improvement is particularly notable given the diminishing returns typically observed in high-accuracy regions. Also, the DAC method demonstrated its ability to improve the generalization capabilities of the network, allowing the model to continue effective training up to 144 epochs without overfitting. We have also shown our technique's effectiveness in locating and cropping the RoI after its search. The significant improvement in accuracy highlights DAC's effectiveness in refining input data and enabling better learning by precisely focusing on relevant features.

The efficiency of DAC method in resizing images while preserving critical information alleviated some of dataset challenges. By increasing resolution from 224×224 to 200×370, DAC leveraged higher resolution images to improve accuracy significantly compared to traditional resizing methods.The DAC method produced more consistent results across all folds of the 10-fold cross-validation, contrasted with the instability observed in non-DAC training. The consistency is attributed to the removal of noisy data and unnecessary augmentation, ensuring that only relevant features were emphasized during training.

Unlike other similar methods, our method does not need supervised learning, takes every pixel into account, and can be applied immediately after the input layer. The proposed method is simple to include in existing models and may be effectively trained end-to-end.

Auto-crop enables the network's input data to be cropped and resized, and it may be introduced as a network layer following the input layer (i.e., can be trained with the backpropagation and can be implemented using the Theseus layer (Pineda et al. 2022)) into current CNN architectures without additional supervision or adjustment to the optimization process (i.e., convenient due to gradient ascent). This also allows for the use of a dataset with varying resolution (i.e., multi-resolution). The auto-crop technique is not limited to tackling a specific problem; it may also be used for various Red-Green-Blue (RGB) and grayscale datasets.

Finding RoI can also lead to better data augmentation organization. For example, rotation-based augmentation creates black (no-data) regions in the augmented (rotated) image. With the introduction of the auto-crop, you can obtain an image without black regions if you rotate it around the center of the RoI with a properly selected rotation range.

Declaration of Ethical Standards

This study is derived from the doctoral thesis (thesis number: 836565) titled "Evaluation of Fixed Restorations on Panoramic Radiographs Using Deep Learning and Auto-Crop" completed on 20.09.2023 by Ahmet Esad TOP under the supervision of Asst. Prof. Dr. Mustafa Yeniad.

The authors declare that they comply with all ethical standards. They declare that they follow all ethical guidelines including authorship, citation, data reporting, and publishing original research in all processes of the paper and that they do not make any falsification on the data collected.

Credit Authorship Contribution Statement

Author-1: Conceptualization, Resources, Data curation, Methodology design, Formal analysis, Validation, Investigation, Visualization, Writing – original draft

Author-2: Conceptualization, Resources, Formal analysis, Study design, Validation, Investigation, Supervision, Writing – review and editing, Funding acquisition

Author-3: Conceptualization, Resources, Data curation, Validation, Investigation, Supervision, Writing – review and editing, Opinionator

Author-4: Conceptualization, Methodology design, Study design, Formal analysis, Validation, Investigation, Supervision, Project administration, Writing – review and editing, Opinionator

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors have no conflicts of interest to declare regarding the content of this article.

Data Availability Statement

The authors declare that the main data supporting the findings of this work are available within the article.

All data generated or analyzed during the study are subject to ethics committee approval.

Declaration of Ethics Committee Approval

Ethical approval for the study involving human participants was obtained in accordance with the guidelines set by the AYBU Institutional Review Board and adhering to the principles outlined in the 1964 Declaration of Helsinki and its subsequent amendments or similar ethical standards. The study received official approval and permission from the AYBU non-drug ethical committee with the approval date of 19/04/2019 (File Number: 2019-12).

Declaration of Informed Consent

Written informed consent was not required in this study due to retrospective data collection from all included patients. The informed consent exemption has been granted by the Institutional Review Board and is not expected to adversely affect the rights and health of patients whose data is used. The study ensures that all data is unidentified and anonymized during analysis and reporting, strictly adhering to guidelines to protect patient privacy and confidentiality.

Acknowledgement

The authors would like to express their gratitude to the AYBU Tepebaşı Oral and Dental Health Training Hospital for allowing them to gather the data.

6. References

Çelik, B., Çelik, M.E., 2022. Automated detection of dental restorations using deep learning on panoramic radiographs. Dentomaxillofacial Radiology 51, 20220244. https://doi.org/10.1259/dmfr.20220244

Chen, J., Bai, G., Liang, S., Li, Z., 2016. Automatic image cropping: A computational complexity study, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 507–515. https://doi.org/10.1109/CVPR.2016.61

Choi, J.-W., 2011. Assessment of panoramic radiography as a national oral examination tool: review of the literature. Imaging science in dentistry 41, 1–6. https://doi.org/10.5624%2Fisd.2011.41.1.1

Corbet, E., Ho, D., Lai, S., 2009. Radiographs in periodontal disease diagnosis and management. Australian dental journal 54, S27–S43.

Dai, J., He, K., Sun, J., 2016a. Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3150–3158. https://doi.org/10.1109/CVPR.2016.343

Dai, J., Li, Y., He, K., Sun, J., 2016b. R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems 29. https://doi.org/10.48550/arXiv.1605.06409

Demir, K., Aksakalli, I.K., Bayğin, N., Sökmen, Ö.Ç., 2023. Deep Learning Based Lesion Detection on Dental Panoramic Radiographs, in: 2023 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, pp. 1–6.

Fidan, U., Uzunhisarcıklı, E., Çalıkuşu, İ., 2019. Classification of dermatological data with self organizing maps and support vector machine. Afyon Kocatepe University Journal of Science and Engineering 19, 894–901. https://doi.org/10.35414/akufemubid.591816

Fitzgerald, R., 2000. Phase-sensitive x-ray imaging. Physics today 53, 23–26. https://doi.org/10.1063/1.1292471

Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Han, C., Ye, J., Zhong, Y., Tan, X., Zhang, C., Gao, C., Sang, N., 2019. Re-id driven localization refinement for person search, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9814–9823. https://doi.org/10.1109/ICCV.2019.00991

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969. https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., Oliveira, L., 2018. Deep instance segmentation of teeth in panoramic X-ray images, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, pp. 400–407.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. Advances in neural information processing systems 28. https://doi.org/10.48550/arXiv.1506.02025

Jiang, W., Sun, W., Tagliasacchi, A., Trulls, E., Yi, K.M., 2019. Linearized multi-sampling for differentiable image transformation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2988–2997. https://doi.org/10.1109/ICCV.2019.00308

Katsumata, A., 2023. Deep learning and artificial intelligence in dental diagnostic imaging. Japanese Dental Science Review 59, 329–333.

Kemal, A., Kılıçarslan, S., 2021. COVID-19 diagnosis prediction in emergency care patients using convolutional neural network. Afyon Kocatepe University Journal of Science and Engineering 21, 300–309. https://doi.org/10.35414/akufemubid.788898

Kohinata, K., Kitano, T., Nishiyama, W., Mori, M., Iida, Y., Fujita, H., Katsumata, A., 2023. Deep learning for preliminary profiling of panoramic images. Oral Radiology 39, 275–281.

Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312. https://doi.org/10.1016/j.media.2016.07.007

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Liang, Y., Lv, J., Li, D., Yang, X., Wang, Z., Li, Q., 2022. Accurate Cobb Angle Estimation on Scoliosis X-Ray Images via Deeply-Coupled Two-Stage Network With Differentiable Cropping and Random Perturbation. IEEE Journal of Biomedical and Health Informatics 27, 1488–1499. https://doi.org/10.1109/JBHI.2022.3229847

Liedke, G.S., Spin-Neto, R., Vizzotto, M.B., Da Silveira, P.F., Silveira, H.E.D., Wenzel, A., 2015. Diagnostic accuracy of conventional and digital radiography for detecting misfit between the tooth and restoration in metal-restored teeth. The Journal of prosthetic dentistry 113, 39–47.

Liu, S., Lu, Y., Wang, J., Hu, S., Zhao, J., Zhu, Z., 2020. A new focus evaluation operator based on max–min filter and its application in high quality multi-focus image fusion. Multidimensional Systems and Signal Processing 31, 569–590. https://doi.org/10.1007/s11045-019-00675-2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32. https://doi.org/10.48550/arXiv.1912.01703

Pineda, L., Fan, T., Monge, M., Venkataraman, S., Sodhi, P., Chen, R.T., Ortiz, J., DeTone, D., Wang, A., Anderson, S., 2022. Theseus: A library for differentiable nonlinear optimization. Advances in Neural Information Processing Systems 35, 3801–3818. https://doi.org/10.48550/arXiv.2207.09442

Pröbster, L., Diehl, J., 1992. Slip-casting alumina ceramics for crown and bridge restorations. Quintessence International 23.

Recasens, A., Kellnhofer, P., Stent, S., Matusik, W., Torralba, A., 2018. Learning to zoom: a saliency-based sampling layer for neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–66. https://doi.org/10.1007/978-3-030-01240-3_4

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28. https://doi.org/10.48550/arXiv.1506.01497

Riad, R., Teboul, O., Grangier, D., Zeghidour, N., 2022. Learning strides in convolutional neural networks. arXiv preprint arXiv:2202.01653. https://doi.org/10.48550/arXiv.2202.01653

Rippel, O., Snoek, J., Adams, R.P., 2015. Spectral representations for convolutional neural networks. Advances in neural information processing systems 28. https://doi.org/10.48550/arXiv.1506.03767

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252.

Sakaguchi, R.L., Powers, J.M., 2012. Craig's restorative dental materials-e-book. Elsevier Health Sciences.

Scarfe, W.C., Farman, A.G., 2008. What is cone-beam CT and how does it work? Dental Clinics of North America 52, 707–730. https://doi.org/10.1016/j.cden.2008.05.005

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics

and transfer learning. IEEE transactions on medical imaging 35, 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence.

Top, A.E., 2023. Evaluation of Fixed Restorations on Panoramic Radiographs using Deep Learning and Auto-Crop (PhD Thesis). Ankara Yıldırım Beyazıt Üniversitesi Fen Bilimleri Enstitüsü.

Top, A.E., Özdoğan, M.S., Yeniad, M., 2023. Quantitative level determination of fixed restorations on panoramic radiographs using deep learning. International Journal of Computerized Dentistry 26. 285-299 https://doi.org/10.3290/j.ijcd.b3840521

White, S.C., Heslop, E.W., Hollender, L.G., Mosier, K.M., Ruprecht, A., Shrout, M.K., 2001. Parameters of radiologic care: An official report of the American Academy of Oral and Maxillofacial Radiology. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology 91, 498–511.

Yurttakal, A.H., Baş, H., 2021. Possibility Prediction Of Diabetes Mellitus At Early Stage Via Stacked Ensemble Deep Neural Network. Afyon Kocatepe University Journal of Science and Engineering 21, 812–819. https://doi.org/10.35414/akufemubid.946264

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European Conference on Computer Vision. Springer, pp. 818–833.