

UNIVERSIDAD NACIONAL DE ASUNCIÓN
FACULTAD POLITÉCNICA



**CORRELATION BETWEEN IMAGE QUALITY METRICS OF
DIFFERENT IMAGE ENHANCEMENT TECHNIQUES APPLIED TO
PANORAMIC X-RAY IMAGES AND THE TEETH DETECTION
RESULTS OF DEEP LEARNING ARCHITECTURES**

CLAUDIA PATRICIA LEZCANO GIARDINA

Supervisors:

Prof. Horacio Legal Ayala, Dr.

Prof. José Luis Vázquez Noguera, D.Sc.

Prof. Luis Salgueiro, M.Sc.

A thesis submitted to the Polytechnic Faculty, National University of Asunción,
in fulfillment of the requirements for the degree of Master of Science in Computer Science

San Lorenzo - Paraguay

October, 2021

Datos Internacionales de Catalogación en la Publicación (CIP)
DE BIBLIOTECA CENTRAL DE LA UNA

Lezcano Giardina, Claudia Patricia.

Correlation between image quality metrics of different image enhancement techniques applied to panoramic X-ray images and the teeth detection results of deep learning architectures / Claudia Patricia Lezcano Giardina. - - San Lorenzo, 2021.
135 p.: il.

Tesis (Maestría en Ciencias de la Computación). - - Facultad Politécnica,
Universidad Nacional de Asunción, 2021

Bibliografía

1. Image enhancement. 2. Teeth detection. 3. Panoramic images. 4. Deep learning.

CDD 616.075 4

Hoja de aprobación de Tesis

**CORRELATION BETWEEN IMAGE QUALITY METRICS OF
DIFFERENT IMAGE ENHANCEMENT TECHNIQUES APPLIED
TO PANORAMIC X-RAY IMAGES AND THE TEETH
DETECTION RESULTS OF DEEP LEARNING
ARCHITECTURES**

Claudia Patricia Lezcano Giardina

Tesis de Maestría aprobada el 18 de Octubre de 2021 por los siguientes miembros del Jurado de Defensa:

Dr. Sebastian Grillo, Universidad Autónoma de Asunción
Dr. Jorge Mello Roman, Universidad Nacional de Concepción

Dr. Diego Stalder, FIUNA

Dr. José Luis Vazquez, FP-UNA, co-orientador de tesis.

MSc. Luis Salgueiro, Univ. Politécnica de Cataluña, España, co-orientador de tesis

Dr. Horacio Legal Ayala, FP-UNA, orientador de tesis



Prof. Dr. Horacio A. Legal Ayala

Coordinador Académico

Postgrado en Ciencias de la Computación

Facultad Politécnica

Universidad Nacional de Asunción



Prof. Dr. José Luis Vazquez N.

Co-Orientador

UNIVERSIDAD NACIONAL DE ASUNCIÓN
FACULTAD POLITÉCNICA



**CORRELATION BETWEEN IMAGE QUALITY METRICS OF
DIFFERENT IMAGE ENHANCEMENT TECHNIQUES APPLIED TO
PANORAMIC X-RAY IMAGES AND THE TEETH DETECTION
RESULTS OF DEEP LEARNING ARCHITECTURES**

CLAUDIA PATRICIA LEZCANO GIARDINA

A thesis submitted to the Polytechnic Faculty, National University of Asunción,
in fulfillment of the requirements for the degree of Master of Science in Computer Science

San Lorenzo - Paraguay
October, 2021

*In dedication to my grandmother Lilia and my uncle Vicente.
Always on my mind, forever in my heart.*

*For my wonderful mother Lilia, my sister Andrea, and my brother Manuel.
Thank you for always giving me all the support to follow my dreams.*

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Prof. Dr. Horacio Legal, Prof. D.Sc. José Luis Vázquez Noguera, and M.Sc. Luis Salgueiro for their motivation, encouragement, patience, knowledge, and guidance throughout the master's degree and writing of this thesis.

Besides my advisors, I would like to thank Dr. Vicente R. Fretes and Dr. Diego Defazio, for their contributions to the development of this research.

Finally, I would like to thank NIDTEC for the opportunity provided and CONACYT for the scholarship awarded.

Correlation between image quality metrics of different image enhancement techniques applied to panoramic X-ray images and the teeth detection results of deep learning architectures

Author: Claudia Lezcano Giardina

Supervisors: Prof. Horacio Legal Ayala, Dr.

Prof. José Luis Vázquez Noguera, D. Sc.

Prof. Luis Salgueiro, M. Sc.

ABSTRACT

Dental radiographic images complement the clinical examination, aiding the diagnostic process and enabling professionals to plan appropriate treatment for patients. Interpreting these images is an arduous task, performed mainly by dentists prone to misdiagnosis due to fatigue or lack of professional experience. In recent years, to reduce the workload of professionals, work has been done on the development of computerized systems for the analysis of radiographic images for clinical use. In the field of dentistry, deep learning approaches focus on the automatic detection and segmentation of teeth. Automatic teeth detection and segmentation are complex tasks, especially in panoramic images, which are one of the most requested exams for allowing the visualization of the maxillary and mandibular teeth along with the surrounding bone structure. The teeth detection results of automatic applications are held to high standards of accuracy for their clinical use. Achieving this level of accuracy is challenging for state-of-the-art object detection methods. In most of the recent works, researchers focus on post-processing techniques and fully neural network approaches to improve the results achieved by their teeth detection methods. The implementation of pre-processing techniques has proven to be a differentiating factor between deep learning algorithms with the same architecture. Nonetheless, selecting an image enhancement technique for improving the results of a deep learning architecture is not a trivial task, since on some occasions it can have the opposite effect. In this work, different image enhancement techniques are applied to panoramic images as a pre-processing step to determine which image characteristics correlate with improvements in the detection of teeth. For this purpose, five architectures were selected to train eleven models with a dataset composed of 300 images. For evaluating the enhancement techniques, various image quality metrics were considered, while the detection performance was evaluated using Accuracy, Precision, and Recall. The cross-correlation between the image quality metrics measured per image and the performance of the detection models was then computed. The selected enhancement techniques did not always improve teeth detection, with the Precision and Recall of the models being affected by the changes of different characteristics of the images. The presence and strength of the correlation vary according to the architecture, with Cascade R-CNN presenting the stronger correlation results.

Keywords: Image Enhancement, Teeth detection, Panoramic images, Deep Learning

Contents	Page
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF FIGURES	viii
LIST OF TABLES	xiv
LIST OF SYMBOLS	xv
LIST OF ACRONYMS	xvi
1 INTRODUCTION	1
1.1 Objectives	4
1.1.1 General Objective	4
1.1.2 Specific Objectives	4
1.2 Thesis organization	4
2 PRELIMINARY CONCEPTS	6
2.1 Panoramic Images	6
2.2 Dentistry terminology	8
2.3 Digital Image Processing	10
2.4 Deep Learning	12
2.4.1 Object Recognition	13
3 METHODOLOGY	16
3.1 Dataset	16
3.2 Image Enhancement techniques (IET)	18
3.2.1 Image Enhancement techniques (IET)	19
3.2.2 Image Quality Metrics (IQM)	32
3.3 Object detection architectures	39
3.3.1 Selected architectures	39
3.3.2 Detection evaluation metrics	45
3.4 Experimental setup	48
4 RESULTS AND DISCUSSION	52
4.1 IQM results	53
4.1.1 Discussion	55
4.2 Teeth detection	56
4.2.1 Discussion	64
4.3 Cross-correlation	65

4.3.1 Discussion	71
4.4 Inference on local panoramic images	73
4.4.1 Discussion	77
5 CONCLUSION AND FUTURE WORKS	78
5.1 Conclusion	78
5.2 Future work	80
6 SUMMARY IN SPANISH	82
BIBLIOGRAPHY	94
APPENDIX	100
A.1 VGG Image Annotator	100
A.2 Inference on challenging images	103
A.3 Presentation in International Conferences	109

2.1	Image a) is a bitewing X-ray image, b) shows a periapical X-ray image, and c) is a panoramic X-ray image. These images were taken in a local clinic in Asunción.	7
2.2	Restored teeth presence, marked with red boxes, in images a) and b) . Images c) and d) contain impacted teeth marked with blue boxes. The images in e) and f) show mouths with a fixed implant bridge and implants with and without abutment crowns marked with orange boxes. Lastly, in images g) and h) , we can see some teeth fragments marked with green boxes.	9
2.3	The figure shows the different shades of gray, where zero represents black in the rectangles and white represents the maximum gray level based on bit depth (Wainschenker et al., 2011).	11
2.4	Illustration of a DNN made using the figures of Greenewald et al. (2019) and Yirka (2019).	13
2.5	Illustration of a CNN by Kamarudin et al., 2017. The neurons of convolutional layers are known as feature maps or filters, which sizes are equivalent to the input image's size. Non-linear layers are used to applied non-linear functions, while the pooling layer enables the CNN to learn from features of different sizes (Kamarudin et al., 2017).	13
2.6	a) Image classification output example, b) Classification and object localization output example, c) Object detection output example, d) output of a Semantic segmentation algorithm, and e) output example of Instance segmentation task (Fei-Fei et al., 2018).	14
3.1	Images corresponding to categories 1, 4, 6 and 10. The image of category 1 shows a mouth with 32 teeth, appliance, and restored teeth. The image of category 4 is a mouth with 32 teeth. The image of category 6 shows a children mouth with 45 teeth. Last, in the image of category 10 we can see a mouth with only 14 teeth.	17
3.2	The original image, belonging to category 1, on top, which is the filter input, compared to the filtered image with the anisotropic diffusion below.	20
3.3	The original image on the top, which is the filter input, compared to the bilateral filtering image below.	22
3.4	The original image on the top is compared with the image generated by the Maximum or dilation filter.	23
3.5	The original image on the top is compared with the image generated by the Minimum or erosion filter.	24
3.6	Comparison of an original image, on top, with the image after applying the Edge Enhancement technique with the mask show in 3.4 below.	26
3.7	Comparison of an original image, on the left, with the image after applying the BBHE contrast enhancement technique.	27

3.8	Comparison of an original image, on the left, with the image after applying the CLAHE contrast enhancement technique.	28
3.9	Comparison of an original image, on the top, with the image after applying the GRMMCE contrast enhancement technique.	29
3.10	Comparison of an original image, on the top, with the image after applying the MTHT contrast enhancement technique.	30
3.11	Comparison of an original image, on the top, with the image after applying the QHELC contrast enhancement technique.	32
3.12	The upper left image shows the original image with $AMBE$ equal to zero, on the left, the same image with 50% more brightness has $AMBE = 0.226$. The images in the bottom present 100% and 150% more brightness than the original image, with $AMBE$ values of 0.302 and 0.410, respectively.	33
3.13	The original image of category 5 present an entropy of 6.70, the image on its right is an image enhanced with the edge enhancement technique that has an entropy of 6.90. The images in the lower part were improved with the BBHE and CLAHE techniques and present entropy values of 6.60 and 7.43, respectively.	34
3.14	An original image of category 8, in the upper left corner, is compared with its enhanced versions with SSIM values of 0.997, upper right corner, 0.847, bottom left corner, and 0.712 in the bottom right corner of the figure.	35
3.15	An original image of category 6, in the upper left corner, is compared with its enhanced versions presenting PSNR / PSNR-HVS values of 44.62 / 86.67, upper right corner, 28.64 / 70.34, bottom left corner, and 14.29 / 57.01 in the bottom right corner of the figure. From these values, we can conclude that PSNR and PSNR-HVS are directly proportional.	36
3.16	An original image of category 5, in the upper left corner, is compared with its enhanced versions with EPI values of 0.940, upper right corner, 0.576, bottom left corner, and 0.164 in the bottom right corner of the figure.	37
3.17	An original image of category 1, in the upper left corner, is compared with its enhanced versions with contrast values of 74.56, upper right corner, 52.21, bottom left corner, and 33.37 in the bottom right corner of the figure.	38
3.18	An original image of category 10, in the upper left corner, is compared with its enhanced versions with CIR values of 70.85, upper right corner, 29.98, bottom left corner, and 0.071 in the bottom right corner of the figure.	39
3.19	Illustration of Faster R-CNN architecture (Ren et al., 2015).	40
3.20	Illustration of Cascade R-CNN architecture (Cai & Vasconcelos, 2019).	41

3.21	RetinaNet illustration, where a) and b) are the backbone and the FPN, while b) and c) are the classification and the box regression subnetworks, respectively. The final output of the classification subnetwork has the form $W \times H \times KA$, because of K –length vector of classification targets assigned to each anchor A . Whereas, the box regression subnetwork attached a vector of the coordinates of box regression targets for each anchor, having an output form of $W \times H \times 4A$ (Lin et al., 2018).	43
3.22	Left: Prediction of the 4D vector (l, t, r, b) which encodes the location of a bounding box. Right: an ambiguous sample annotation (Tian et al., 2019).	44
3.23	FCOS architecture illustration. The backbone a) extracts features maps at three levels, while the FPN b) predicts features at five levels. The classification head c) , with an output of $W \times H \times C$, for an input of width W and height H with C classes. The center-ness head d) outputs the normalize distance from the center, whereas the regression head e) output is the 4D vector (l, t, r, b) (Tian et al., 2019).	44
3.24	YOLOv ₃ architecture illustration (Redmon & Farhadi, 2018).	45
3.25	IoU quantifies the area of overlap between a predicted box and a ground truth box. Different scenarios are shown in this illustration, where predicted boxes have IoU values between 0.95 and 0.13.	46
3.26	a) A predicted bounding box with the class label <i>cat</i> (red box) overlaps the ground truth box (blue box) with an IoU of 0.9, this is considered a TP. b) A predicted bounding box overlaps the ground truth box with an $IoU < 0.5$, which is considered a FP. c) a predicted bounding box with the class label <i>cat</i> does not overlap the ground truth box, this is a FP. d) there are not prediction for the object, this is a FN example.	47
3.27	The image in a) has the original size of 1991×1127 pixels, as the rest of the images in the dataset of Silva et al., 2018. The reference points are shown with the help of the white arrows, where 1 is the nasal septum, 2 is the inferior mandibular border, and 3 are the posterior mandibular ramus. In b) , with a dimension of 1886×829 , the trimmed sections of a) are highlighted in gray. In this image we can see that the detection of the reference points was successful except for the left posterior mandibular ramus. Because of this, the right part of the image is wider than expected.	48
3.28	Diagram of steps, followed to obtain the results of this work.	51

4.1	a) Original image; b) Anisotropic Diffusion output image; c) Bilateral Filtering output image; d) Edge Enhancement output image; e) Maximum output image; f) Minimum output image. The changes introduced by the b) Anisotropic Diffusion and the c) Bilateral Filtering are almost visually imperceptible, while the d) Edge Enhancement shows an image with highlighted edges. In e) , teeth have a thickened appearance, while in f) , the effect is the opposite.	54
4.2	a) Original image; b) BBHE output image; c) CLAHE output image; d) GRMMCE output image; e) MTHHT output image; f) QHELC output image. The BBHE image in b) , presents the greatest change in its contrast, followed by the CLAHE image in c) , both also having a visually perceptible change in their brightness. It can also be perceived a change in the e) MTHHT image, which presents a high value of CIR in comparison to the rest of the images in this figure. Lastly, the changes introduced by QHELC in f) , are practically imperceptible to the eye, which is translated to its IQM values, shown below the image.	55
4.3	Inference of Faster R-CNN: a) baseline model, b) Bilateral Filtering model, c) CLAHE model, d) MTHHT model. The undetected teeth are marked with the blue boxes, being two in a) , and one tooth in b) and d)	57
4.4	Comparison of teeth detection performed by RetinaNet a) baseline model, b) BBHE, c) GRMMCE, and d) Edge Enhancement on an image belonging to category 1. The red box in image b) are false positives, while the blue boxes indicates false negatives.	59
4.5	The above image shows the detection performed by Cascade R-CNN a) baseline and b) QHELC models in a challenging image from category 9 with 30 teeth when using 0.5 as a threshold. In a) , neighboring teeth were detected together in one box, while in b) , QHELC is able to detect these two teeth independently.	60
4.6	In these images, the inference performed by the baseline model a) of Cascade R-CNN, is compared with the GRMMCE b) and QHELC c) models while using an 0.75 of threshold. In a) , the model is unable to detect a crown, marked with a blue box, while the GRMMCE and QHELC models are able to detect it. . . .	61
4.7	Inference performed by FCOS a) baseline model and b) QHELC model in an image with 42 teeth. Both models failed to detect all 42 teeth, blue boxes marked the four and two undetected teeth in a) and b) , respectively.	63
4.8	Some undetected teeth, marked with a blue box, when using the baseline model of YOLOv ₃ with an 0.5 of threshold. The images a) , b) , and c) belong to category 1. In d) , an image of category 2 is shown, while e) and d) present images of categories 3 and 4, respectively. At the bottom, images g) and h) correspond to images of categories 5 and 9, respectively.	65

4.9	Correlation heat maps of Faster R-CNN models, where A was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and B is the heat map formed with the 0.75 detection results.	67
4.10	Correlation heat map of RetinaNet models, where A was formed using the detection results of the model when using the threshold of 0.5 and the IQM measured per image in the testing set, and B is the heat map formed with the 0.75 detection results.	68
4.11	Correlation heat map of Cascade R-CNN models, where A was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and B is the heat map formed with the 0.75 detection results.	69
4.12	Correlation heat map of FCOS models, where A was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and B is the heat map formed with the 0.75 detection results.	70
4.13	Correlation heat map of YOLOv ₃ models, where A was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and B is the heat map formed with the 0.75 detection results.	71
4.14	Above images: inference performed by the Faster R-CNN baseline model in a) is compared with the Maximum model in b). Bellow images: inference performed by the baseline model c) is compared against the Edge Enhancement model d).	74
4.15	Inference performed by Cascade R-CNN baseline model in a) is compared against QHELC model in b). Here we can see that while the QHELC model can detect all teeth in the image, the baseline model missed one of the teeth.	75
4.16	Correlation heat maps of Faster and Cascade R-CNN models, formed using the detection results of Table 4.8 and the IQM of each of the 20 images.	77
A.1	VIA getting started explanation.	100
A.2	The selection of the type of region for drawing in the images is chosen in the <i>Region Shape</i> area, marked with the red box number 1. In the <i>Project</i> area, marked with the number 2, the images added to the project can be seen and selected for the annotation. This area also enables removing and adding images to the project. The area in 3, <i>Attributes</i> , is where the name of the regions are added and edited, if necessary.	101
A.3	Set of annotated images of the training set.	102

A.4	View of the COCO format JSON files of the training and testing set. The code on the left shows the training set annotations files with 5622 teeth annotated in the 192 images of the set. The red box shows that the only attribute is the label tooth, which is given the ID number 0. In the code on the right, the annotations of the testing set with 1749 of teeth in the 60 images of the set, are shown with the green and yellow boxes highlighting the first and second boxes of the first image. These boxes show the ID, coordinates, are, and category ID of each bounding box. The category ID of both is 0, corresponding to the label tooth.	103
A.5	Inference performed by four models on an image belonging to category 6. a) shows the inference performed by the baseline model and b) the QHELC model of Faster R-CNN. While c) and d) shows the inference performed by the baseline and CLAHE models of RetinaNet, respectively. At the bottom, e) and f) are the FCOS baseline and QHELC models, respectively.	105
A.6	Inference performed by four models on an image belonging to category 9. a) shows the inference performed by the baseline model and b) the MTHT model of Faster R-CNN. While c) and d) shows the inference performed by the baseline and QHELC models of Cascade R-CNN, respectively. At the bottom, e) and f) are the RetinaNet baseline and CLAHE models, respectively.	106
A.7	Inference performed by six models on an image belonging to category 10. a) shows the inference performed by the baseline model and b) the MTHT model of Faster R-CNN. While c) and d) shows the inference performed by the baseline and QHELC models of Cascade R-CNN, respectively. At the bottom, e) and f) are the RetinaNet baseline and QHELC models, respectively.	107
A.8	Inference performed by six models on an image belonging to category 10. a) shows the inference performed by the baseline model and b) the MTHT model of Faster R-CNN. While c) and d) shows the inference performed by the baseline and GRMMCE models of RetinaNet, respectively. At the bottom, e) and f) are the images of the FCOS baseline and QHELC models, respectively.	108
A.9	Inference performed by the baseline models of FCOS with the threshold set to a) 0.5 and b) 0.75.	109

LIST OF TABLES

Page

3.1	Dataset image classification (Silva et al., [2018]).	17
3.2	Distribution of categories in the training, validation, and testing set.	18
3.3	Chosen pre-trained models per architecture.	49
3.4	Hyperparameters for each method.	49
3.5	Range of values tested during tuning.	50
4.1	Average of IQM measured on the testing group.	53
4.2	Teeth detection results of Faster R-CNN models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.	57
4.3	Teeth detection results of RetinaNet models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.	58
4.4	Teeth detection results of Cascade R-CNN models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.	59
4.5	Teeth detection results of FCOS models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.	62
4.6	Teeth detection results of YOLOv ₃ models.	64
4.7	Average of IQM measured on the 20 images.	73
4.8	Teeth detection results of Faster and Cascade R-CNN models trained using the 20 images belonging to a local clinic.	74

LIST OF SYMBOLS

$f(i, j)$	Digital image.
β_d	Domain spread of the bilateral filtering.
β_r	Range spread of the bilateral filtering.
$Hist()$	Histogram of a grayscale image.
k	Intensity level.
n_k	Number of pixels with intensity level k .
x_k	Normalized gray level of k .
$I(i, j)$	Input image.
$O(i, j)$	Output image.
$Dil(\cdot)$	Dilation.
$Ero(\cdot)$	Erosion.
SE	Structural Element.
mk	Mask image.
R_{mk}^{Dil}	Reconstruction by dilation.
ϕ_{SE}^R	Opening by reconstruction.
$WTHR$	White Top-Hat transform by reconstruction.
$\nabla^2(\cdot)$	Laplacian operator.
$\mu(\cdot)$	Statistical mean or mean luminance of an image.
$E(\cdot)$	Entropy.
M	Number of row in the image.
N	Number of columns in the image.
L	Maximum gray level.
(i, j) and (u, v)	Spatial coordinate representing a pixel of the image.
$l(\cdot, \cdot)$	Luminance comparison function.
$c(\cdot, \cdot)$	Contrast comparison function.
σ	Standard deviation.
$s(\cdot, \cdot)$	Structure comparison function.
cov	Covariance.
C	Constant.
$lc(i, j)$	Local contrast.
m	Arithmetic mean in the 3×3 window centered on k .
$p(\cdot)$	Probability of occurrence of an event.
$\Delta(\cdot)$	High pass filtered version of an image.
$CE(\cdot)$	Cross entropy
$FL(\cdot)$	Focal loss entropy
q	Class label.
γ	Focusing factor of FL.
W	Width of the output.
H	Height of the output.
K	Vector of classification targets.
A	Anchor.
(l, t, r, b)	Distance from a location to the four sides of the bounding box.
C	Number of classes.
P_q	Precision obtained for the class label q .

LIST OF ACRONYMS

ADA	A merican D ental A ssociation
AMBE	A bsolute M ean B rightness E rror
AC	A verage C ontrast
AP	A verage P recision
AR	A verage R ecall
BBHE	B rightness P reserving B i- H istogram E qualization
CE	C ross E ntropy
CIR	C ontrast I mprovement R atio
CLAHE	C ontrast L imited A daptive H istogram E qualization
CLHE	C ontrast L imited H istogram E qualization
CNN	C onvolutional N eural N etwork
DCT	D iscrete C osine T ransform
DNN	D eep N eural N etwork
EPI	E dge P reservation I ndex
FCOS	F ully C onvolutional O ne- S tage O bject D etection
FN	F alse N egative
FP	F alse P ositive
FPN	F eature P yramid N etwork
GRMMCE	G eodesic R econstruction M M C E
HVS	H uman V isual S ystem
IET	I mage E nhancement T echniques
IQM	I mage Q uality M etrics
IoU	I ntersection o ver U nion
mAP	m ean A verage P recision
MMCE	M ultiscale M orphological C ontrast E nhancement
MTHT	M ulti-scale T op- H at T ransform
PDE	P artial D ifferential E quation
PSNR	P eak S ignal to N oise R atio
QHELC	Q uadri- H istogram E qualization with L imited C ontrast
R-CNN	R egion-based C NN
R-FCN	R egion B ased F ully C NN
RoI	R egion o f I nterest
RPN	R egion P roposal N etwork
SD	S tandard D eviation
SGD	S tochastic G radient D escent
SSIM	S tructural S imilarity I ndex M easure
TP	T rue P ositive
VIA	V isual G eometry G roup I mage A nnotator
YOLO	Y ou O nly L ook O nce

CHAPTER 1

INTRODUCTION

Dental radiographic images complement the clinical examination by allowing the visualization of the teeth and surrounding bone structure. These images are essential tools that aid the diagnostic process and enable professionals to plan appropriate treatment (Jader et al., 2018). Moreover, dental X-ray images are also used in the forensics field for human identification by comparing an image obtained from a cadaver to its dental records (H. Chen et al., 2019). Bitewing, periapical, and panoramic X-ray images are three commonly used types of dental X-ray images (Jader et al., 2018). Unlike bitewing and periapical, panoramic images allow a 2-D visualization of the jaws, enabling the visualization of hidden caries, impacted and embedded teeth, cysts, tumors, bone loss, or bone abnormalities, among others (Betul, 2017; C.-W. Wang et al., 2016). In fact, panoramic radiography is one of the most requested exams for allowing the visualization of the maxillary and mandibular teeth and the surrounding bone structure, while exposing the patient to a small dose of radiation (Jader et al., 2018; Tuzoff et al., 2019). The reading and interpretation of radiographic images, especially panoramic images, is an arduous task performed mainly by dentists, prone to an incorrect diagnosis due to fatigue, bias, or lack of professional experience (H. Chen et al., 2019). Automatic analysis of dental radiographic images, as fundamental tools in the diagnosis and monitoring of oral health, acquires great importance to reduce the workload of professionals (Tuzoff et al., 2019).

In recent years, work has been done on the development of computerized systems for the analysis of radiographic images for clinical use. The localization of anatomical structures, which can be performed by image-processing and machine learning-based algorithms, is a crucial part of this analysis (Chung et al., 2021). Among the machine learning-based approaches, deep learning-based methods are becoming an increasingly chosen option since groundbreaking results have been achieved using Convolutional Neural Networks (CNN), especially for detection and classification tasks (Betul, 2017; Chung et al., 2021). Deep learning is a type of learnable

algorithm that is used to automatically extract and learn from a set of input data, without the need for hand-crafted feature engineering, for its later use in the analysis and interpretation of unseen data. Deep CNNs are among the most widely used deep learning algorithms for image recognition tasks, as they enable the extraction of specific features from the image to effectively represent and learn characteristics using multiple levels of abstraction (LeCun et al., 2015; Tuzoff et al., 2019).

In the field of dentistry, neural networks are being mainly used for the detection and segmentation of teeth, with a few works focusing on the classification of dental pathologies as caries and periodontal bone loss (Schwendicke et al., 2019). However, an automatic application for teeth detection or segmentation is held to high standards of accuracy for its clinical use, achieving this is challenging for state-of-the-art object detection methods (Chung et al., 2021). In addition, dental radiographic images are affected by permanent and temporary artifacts that represent an extra challenge for object detection methods. Tooth fillings, implants, restorations, metallic materials placed after trauma, among others, represent the permanent artifacts that will affect specially the localization of teeth in panoramic images. While dental appliances, impacted and embedded teeth, overlapping teeth, and missing teeth, represent the temporary artifacts that difficult the automatic localization of teeth (Amer & Aqel, 2015; Ehsani Rad et al., 2013). Recent deep learning-based work for tooth detection and classification used post-processing algorithms or fully deep learning method approaches to improve the detection results achieved by their methods. Tuzoff et al. (2019) used Faster R-CNN (Ren et al., 2015) architecture for teeth detection and numbering in panoramic radiographs, achieving a recall of 99.41% and a precision of 99.45% in the detection task. While achieving a specificity of 99.94% and a recall of 98.8% in the teeth numbering by using a heuristic method as post-processing to improve the results. While the work presented by Muramatsu et al. (2020) achieved a recall of 96.4% when using DetectNet (Barker et al., 2016) on panoramic images, which are first trimmed to exclude toothless regions. The method classifies the teeth into four tooth types, molar, premolar, canine, and incisor. The teeth detection algorithm, proposed by H. Chen et al. (2019), consists of three post-processing techniques to improve the Faster R-CNN detections and numbering of teeth on periapical images, achieving precision and recall of 98.8% and 98.5%, respectively. In the work presented by Betul (2017), an AlexNet (Krizhevsky et al., 2012) architecture is used for tooth detection and classification into molars, premolars, and canine-incisors in panoramic images. A mouth gap detection is implemented as a pre-processing step to decrease the search space in the images, achieving 94.32%, 91.74%, and 92.47% of accuracy of molars teeth, premolars teeth, and canine-incisors, respectively. Chung et al. (2021) proposed a point regression-based object detection method for tooth detection and classification using a cascaded neural network in panoramic images. A popular contrast enhancement algorithm, Contrast Limited Adaptive Histogram Equalization (CLAHE), was applied to the images to reduce the contrast variance of the images taken with four different X-ray machines. The algorithm achieves a precision of 99.7% and a recall of 97.2% in detecting teeth and missing teeth, than are then classified into 32 classes. Miki et al. (2017) proposed a method for automatically locating teeth on cone-beam

computed tomography images using a fully convolutional network based on AlexNet architecture. The average detection rate of the method, trained with 42 images and tested on 10, is 77.4% with 5.8 false positives per image. In the work presented by Zhang et al. (2018), a combination of Faster R-CNN and Region-Based Fully Convolutional Neural Networks (R-FCN) (Dai et al., 2016) is used for detecting and classifying teeth on periapical images. The detection and classification are improved by using a novel label tree and a cascade network structure, achieving a precision of 98.0% and a recall of 98.3% in teeth detection. The method not only detects and classifies healthy teeth using a dental notation universal system, but it can also detect missing teeth, decayed teeth, and filled teeth, achieving a precision of 95.8% and a recall of 96.1%.

The success of a deep learning method does not only depend on the architecture; achieving good performance is also related to aspects outside the network such as pre-processing, data augmentation, and hyperparameter optimization. Implementing pre-processing or data augmentation techniques has proven to be a differentiating factor between deep learning algorithms with the same architecture (Litjens et al., 2017). However, selecting an image enhancement technique with the purpose of improving the detection performance of a method is not an easy task. Xiaoran (2019) used five different image enhancement algorithms in order to explore the effect on the performance of two CNN architectures when classifying images belonging to two sets of images. One of the datasets was composed of RGB images of cats and dogs, while the second contained grayscale Chest X-ray images. According to the results, only one enhancement algorithm improved the performance of the first architecture, LeNet5 (LeCun et al., 1998), while when using a pre-trained VGG16 (Simonyan & Zisserman, 2014), four of the enhancement algorithms reduce the performance of the model. Rahman et al. (2021) explored the effects of applying various image enhancement techniques for the classification of Chest X-ray, with and without pneumonia caused by COVID-19, using six different CNN. One technique achieved the best results in classifying the chest X-rays.

In this work, different Image Enhancement Techniques (IET) are applied to panoramic images as a pre-processing step to determine which image characteristics correlate best with improvements in teeth detection results. Five object detection deep learning architectures were selected to be trained and tested using a dataset of 300 panoramic images. The IET were applied to the images and the detection performance of the models trained with the enhanced images was compared with the performance of the model trained with no pre-processing applied to the images, the baseline model. The output of each IET was evaluated with various Image Quality Metrics (IQM), whereas the detection performance of the models was evaluated by measuring the Accuracy, Precision, and Recall after considering as true detection those teeth that presented bounding boxes with an Intersection over Union(IoU) ≥ 0.5 and an $IoU \geq 0.75$. These results were then used to compute the cross-correlation between the performance of the detection models and the IQM measured per image.

The contributions of this work are summarized as follows:

- To the best of our knowledge, no study has analyzed the effect of various IET in detecting teeth in panoramic images and finding the relationship between images' characteristics and the detection performance of different object detection algorithms.
- The sensitivity of the detection performance towards the IET depends on the architecture.
- Some IQM show stronger relationships with a particular detection evaluation metric.

1.1 Objectives

1.1.1 General Objective

To determine the relationship between characteristics of panoramic images enhanced by different IET and the impact on the teeth detection results obtained by deep learning architectures.

1.1.2 Specific Objectives

- To select appropriate image enhancement techniques for applying to the panoramic X-ray images.
- To apply the selected image enhancement techniques to the panoramic X-ray images as a pre-processing step.
- To measure the outcome of the image enhancement techniques using the image quality metrics.
- To train the models using the chosen object detection deep learning architectures.
- To evaluate the models' performance computing the Accuracy, Precision, and Recall metrics.
- To compute the cross-correlation between the image quality metrics and the detection evaluation metrics.

1.2 Thesis organization

The rest of this document is organized as follows. Chapter 2 present a description of important concept related to panoramic images (section 2.1), dentistry terminology (section 2.2), digital image processing (section 2.3), and deep learning (section 2.4). In Chapter 3 section 3.1, the dataset used in this work is described. The Chapter continues with section 3.2, where the IET applied to the images and the IQM measured for evaluating the IET outcome are presented. The next section in this chapter, section 3.3, contains a general description of the selected

object detection architectures. Lastly, in section 3.4, the experimental setup followed in this work closes the Methodology Chapter.

Chapter 4 presents, in section 4.1, the IQM values obtained by measuring the IET outcome on the testing set. Of the following sections of this chapter, section 4.2 displays the object detection performance of each architecture, and section 4.3, the correlation between the IQM values and the detection results. The last section, 4.4, contains the results and discussion of an extra experiment developed with a local dataset. The conclusion of this work and the future works are presented in Chapter 5. Finally, Chapter 6 contains the summary of this work in Spanish.

CHAPTER 2

PRELIMINARY CONCEPTS

2.1 Panoramic Images

X-ray images are obtained using machines that generate X-ray with a certain energy to pass through the patient (Gonzalez & Woods, 2001) to then be captured in radiographic films or X-ray sensors. In the field of dentistry, there are two types of X-ray images, intraoral and extraoral. Intraoral images are obtained by inserting the radiographic film into the patient's mouth in the area to be examined, while extraoral images are taken with the patient between the X-ray source and the radiographic film or X-ray sensors behind the patient. Bitewing, periapical, and panoramic X-ray images are three commonly used types of dental X-ray images (Jader et al., 2018). Bitewing and periapical X-ray images are intraoral, while panoramic X-ray is an extraoral exam. Bitewing images allow the inspection of maxillary and mandibular teeth in a specific area of the mouth, and periapical images are taken when a whole tooth needs to be examined closely. On the other hand, panoramic images are taken to monitor the entire set of teeth and the jaws (C.-W. Wang et al., 2016). The images in Figure 2.1 show examples of these three types of X-ray images.

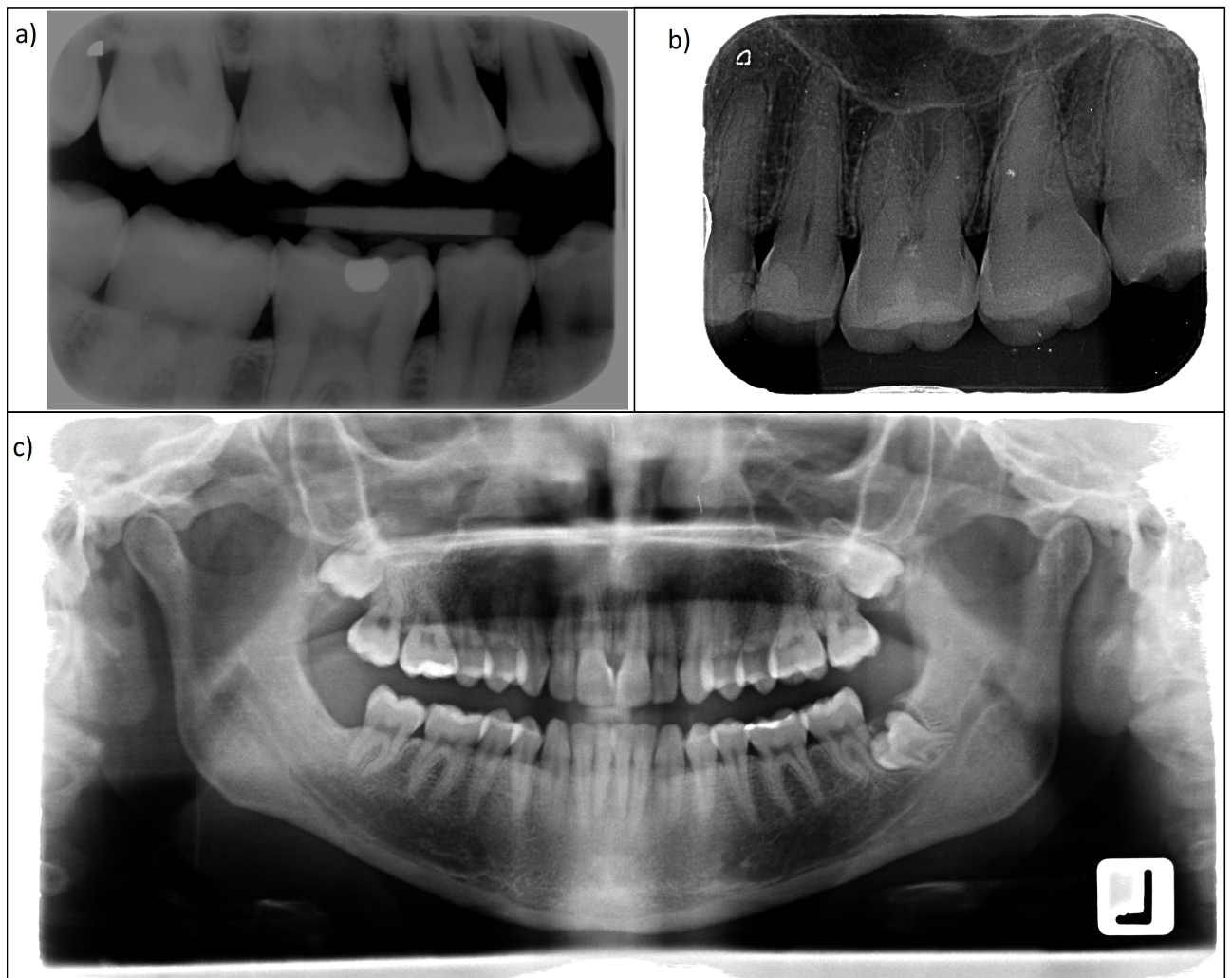


Figure 2.1: Image **a)** is a bitewing X-ray image, **b)** shows a periapical X-ray image, and **c)** is a panoramic X-ray image. These images were taken in a local clinic in Asunción.

Panoramic images are widely used since they allow a two-dimensional view of the nasal area, jaw joints, maxillary and mandibular teeth, and surrounding bone structure (Lira et al., 2010; Silva et al., 2018). It is a non-invasive, very quick exam that is used to plan treatment and can reveal cysts, tumors, hidden caries, bone loss, among others pathologies, while exposing the patient to a very low dose of radiation (Betul, 2017; Lira et al., 2010). Panoramic images can be taken using radiographic films or X-ray sensors, with sensors allowing the generation of high quality digital images.

2.2 Dentistry terminology

Radiographic images allow the visualization of healthy, restored, non-erupted, impacted, and embedded teeth. As well as implants, bridges, abutment crown, and tooth fragments. The definition of these dental terms, extracted from the Glossary of dental clinical terms of the American Dental Association (2021), are presented below:

- Dental restoration is the treatment, which a tooth undergoes after a cavity or trauma, to restore its full functionality.
- A non-erupted tooth is a tooth that has not emerge due to the presence of another tooth (impacted tooth), for lack of strength to erupt (embedded tooth), or to the lack of development.
- An impacted tooth is a not erupted or partially erupted tooth that is unlikely to completely erupt for the presence of another tooth, bone, or soft tissue.
- A dental implant is a material surgically placed used to support dental prosthesis like bridges and abutment crowns.
- Bridges are fixed partial dentures, which are prostheses placed for the replacement of one or more teeth attached to the natural teeth or implants.
- Abutment crowns are artificial crowns used for the retention or support of a dental prosthesis.

Figure 2.2 shows images with restored and impacted teeth, in images **a)** to **d)**, whereas in images **e)** to **h)** there are implants and teeth fragments.

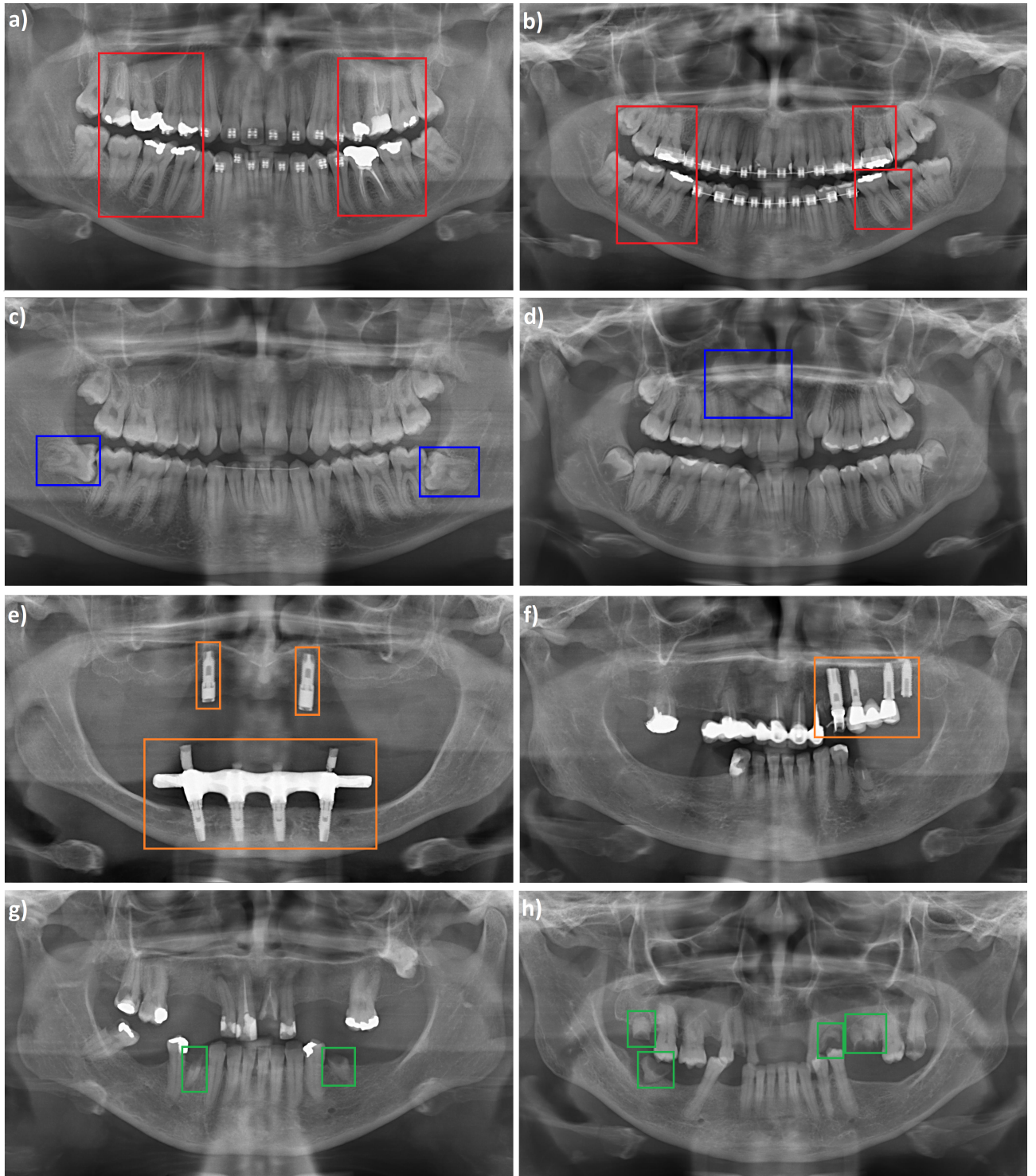


Figure 2.2: Restored teeth presence, marked with red boxes, in images a) and b). Images c) and d) contain impacted teeth marked with blue boxes. The images in e) and f) show mouths with a fixed implant bridge and implants with and without abutment crowns marked with orange boxes. Lastly, in images g) and h), we can see some teeth fragments marked with green boxes.

2.3 Digital Image Processing

Digital image processing is a field of digital signal processing, in which a digital image is processed by an algorithm in order to allow its storage, transmission, and representation for use by an autonomous machine perception, or to improve its representation according to human visual perception (Gonzalez & Woods, 2001). A digital image is defined as a two-dimensional function, $f(i, j)$, where (i, j) are the finite *spatial* coordinates of the finite *intensity* or *gray level* f . The elements of the image, each one with a specific location (i, j) and value $f(i, j)$, are commonly known as *pixels*. The set of all pixels locations, is known as the domain of the image, while the set of all intensity values is the range of the image (Gonzalez & Woods, 2001).

The bit depth, spatial resolution, and contrast are among the characteristics of a grayscale digital image that need to be introduced for a better understanding of the description of the IET selected in this work. The bit depth is the number of bits used for representing the image, which impacts the levels of grays L allowed for the pixels. The number of gray levels is commonly an integer power of two, for example, a black and white image has only two levels and needs one bit for its representation, the next bit depth would be 2, allowing $L = 2^2 = 4$ gray tones in an image. For most applications, images are represented with a bit depth of 8, which allows 256 levels of gray in them (Gonzalez & Woods, 2001). Figure 2.3 shows different gray levels in a rectangle, with the values of intensity of the black and white pixels for each of them. The spatial resolution refers to the number of pixels present in the image, given by $M \times N$, where M is the number of rows and N the number of columns, respectively, of the matrix of pixels used for representing the image (Gonzalez & Woods, 2001). Lastly, the contrast of an image can be defined simply as the difference between the maximum and minimum values of gray levels present in the image. An image is considered to have high contrast when there are an appreciable number of pixels with gray levels that cover a significant part of the gray scale, this is, when the image has a high dynamic range (Gonzalez & Woods, 2001).

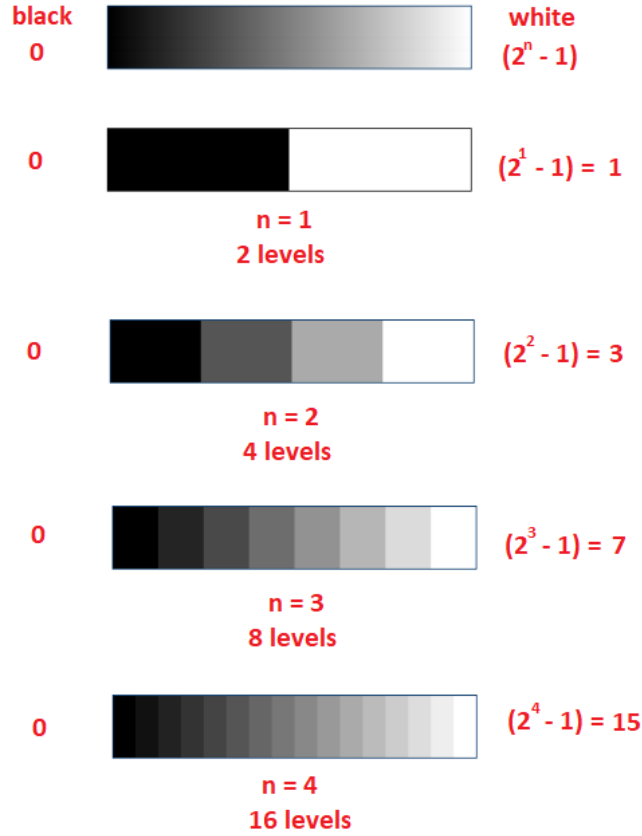


Figure 2.3: The figure shows the different shades of gray, where zero represents black in the rectangles and white represents the maximum gray level based on bit depth (Wainschenker et al., 2011).

The representation or information present in digital images can be improved by: altering the intensity of the pixels, known as the image brightness; removing noise, or the unwanted brightness variations mostly generated at the time of acquisition; sharpening the regions contours by edge enhancement; modifying the contrast of the images (Facon, 2002). IET, which can be divided into spatial domain methods and frequency domain methods, are used in order to make an image more suitable for a specific purpose or application (Gonzalez & Woods, 2001).

Spatial domain methods focus on the pixels of the image by manipulating the value at a certain location (i, j) . The pixel value at (i, j) is transformed by an operator T that is applied to a square or rectangular subimage, known as the pixel neighborhood, where the coordinates (i, j) are the center. In order to have the resulted image, this operator is applied to all the locations in the input image. The spatial domain process can be expressed by:

$$O(i, j) = T[I(i, j)] \quad (2.1)$$

where I and O are the input and the image resulted from the operation, respectively, and T is an operator on I , defined over some neighborhood of (i, j) . The transformation T consists of using a mask, also called filter or kernel, to alter the values of the pixel in an image. A mask is a 2-D array whose values modify the pixel value according to a certain purpose (Gonzalez & Woods, 2001).

Improving the contrast of an image is commonly done by histogram stretching or histogram equalization. The histogram of an image with gray levels in the range $[0, L - 1]$ is a discrete function $Hist(k) = n_k$, where n_k is the number of pixels that have the gray level k . The probability density function of an image is $p_x(x_k) = \frac{n_k}{n}$ where n represents the number of pixels in the image and x_k is the normalized gray level of k (Facon, 2002). While histogram stretching consist of modifying the intensity values so that the entire dynamic range of the image is filled, which is done by a per pixel transformation, histogram equalization modifies the intensity values of an image so that the histogram of the resulted image is, theoretically, flatten. Histogram equalization is applied using:

$$w_k = T(x_k) = \sum_{y=0}^k \frac{n_y}{n} = \sum_{y=0}^k p_x(x_y), \quad (2.2)$$

where the value w_k is scaled to the gray levels present in the image and the new number of pixels with a gray level k must be calculated to have the equalized histogram (Facon, 2002).

2.4 Deep Learning

The interpretation of medical imaging is a challenging task, mostly performed by experts, where the broad range of pathologies, the bias of the professionals, lack of experience, as well as fatigue could lead to misdiagnose (H. Chen et al., 2019; Shen et al., 2017). Computer-assisted intervention becomes important to help professionals with challenging and repetitive tasks such as reading and interpreting images. Machine learning techniques have been helping to improve computational medical image analysis, especially in recent years. However, the success of machine learning techniques strongly depends on the ability to extract meaningful features or the feature representation of the data (Shen et al., 2017). Habitually, experts were in charge of the feature engineering, which consists of designing meaningful features using knowledge about the problem's domain, making it difficult for the non-expert to use machine learning techniques for developing solutions on their own. Deep learning, which is a subfield of machine learning, requires less information related to the problem in order to solve it. Using only a set of data, with or without pre-processing, deep learning transforms the feature extraction step into a learning step, thus, the feature engineering stops being human dependable (Bengio, 2009; Shen et al., 2017). Automatically learning features requires learning complex functions, which can only be done using multiple levels of abstraction (Bengio, 2009).

Deep learning is based on artificial neural networks, whose design is inspired by the brain. The brain processes incoming information from the different senses by passing it through a series of cells capable to decode the signal. These cells, called neurons, communicate with each other through connections called synapses. In machine learning, an artificial neural network is a group of intercommunicated nodes designed to mimic the way in which the brain processes information and learns from it. Some types of neural networks, like CNN, are composed of

layers, a layer is a group of nodes with a certain purpose, specifically three of them: the *input layer* contains the input data, the *output layer* produces the result for the specific input after the computation performed by the *hidden layers* (Kamarudin et al., 2017).

The architecture of a neural network defines the number, type, and size of the layers in it. An architecture is designed for a specific task, such as object detection. Deep Neural Network (DNN) and CNN are two types of deep learning architectures with the ability to learn hierarchical feature representations only from the input data. This is possible because both architectures have numerous hidden layers, with CNN having fewer number of parameters to learn and presenting successful results in image recognition tasks (Kamarudin et al., 2017; Shen et al., 2017). The illustrations of a DNN and CNN are shown in Figures 2.4 2.5.

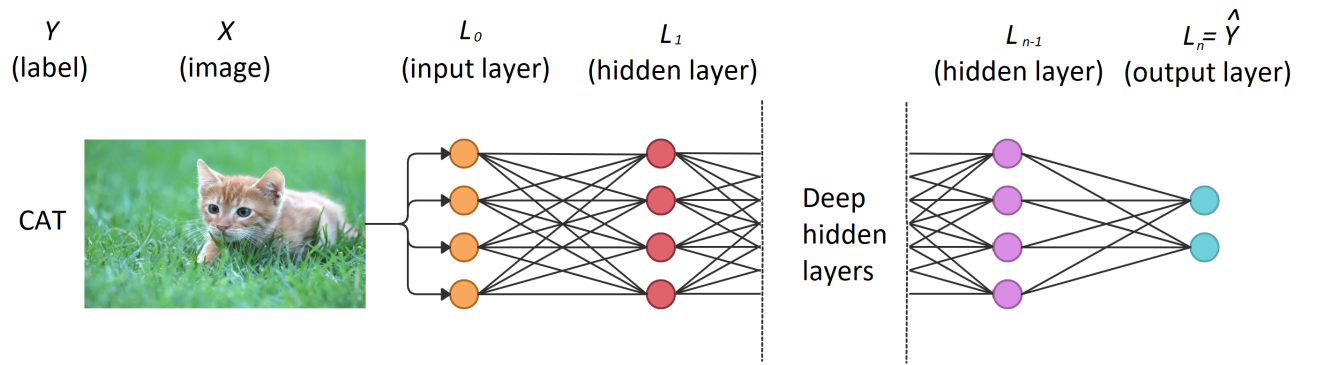


Figure 2.4: Illustration of a DNN made using the figures of Greenewald et al. (2019) and Yirka (2019).

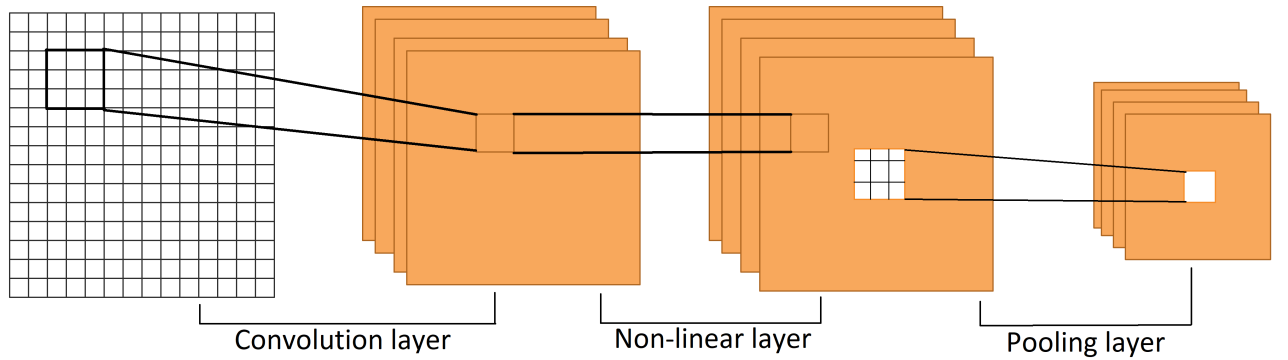


Figure 2.5: Illustration of a CNN by Kamarudin et al., 2017. The neurons of convolutional layers are known as feature maps or filters, which sizes are equivalent to the input image's size. Non-linear layers are used to applied non-linear functions, while the pooling layer enables the CNN to learn from features of different sizes (Kamarudin et al., 2017).

2.4.1 Object Recognition

Object recognition consists of identifying one or more objects of interest in an image or video. This problem can be solved using different computer vision tasks (Brownlee, 2021):

- **Image classification** consists of assigning a *class label* to an image. The algorithms have as an input an image with a unique object of interest and outputs a string which is

the *class label*, where the *label* indicates which object appears in the image. An example output is shown in Figure 2.6 a).

- **Object localization** is a task that involves framing one or more object of interest in an image with a *bounding box*. The *bounding box*, shown in Figure 2.6 b), is a rectangular box that should contain the object of interest. The input is an image with one or more objects of interest, while the output is one or more bounding boxes.
- **Object detection** is the task that combines classification and object localization, where every object of interest in an image has its own bounding box with a class label assigned to it. The input consists of an image with one or more objects of interest, while the output is one or more bounding boxes with a class label for each bounding box. Figure 2.6 c) shows an example of the output of an object detection algorithm.
- **Semantic segmentation** is a computer vision task used for the individual classification of the pixel in an image (Jader et al., 2018). An output example of this task can be found in Figure 2.6 d), where each pixel has been assigned a class label.
- **Instance segmentation** is a task that, combining object detection and semantic segmentation tasks, allows the classification, localization, and segmentation of each object detected in an image (Jader et al., 2018). In Figure 2.6 e), an output example of instance segmentation is shown. The image shows how the pixels belonging to the objects of interest have been colored according to the object to which they belong.

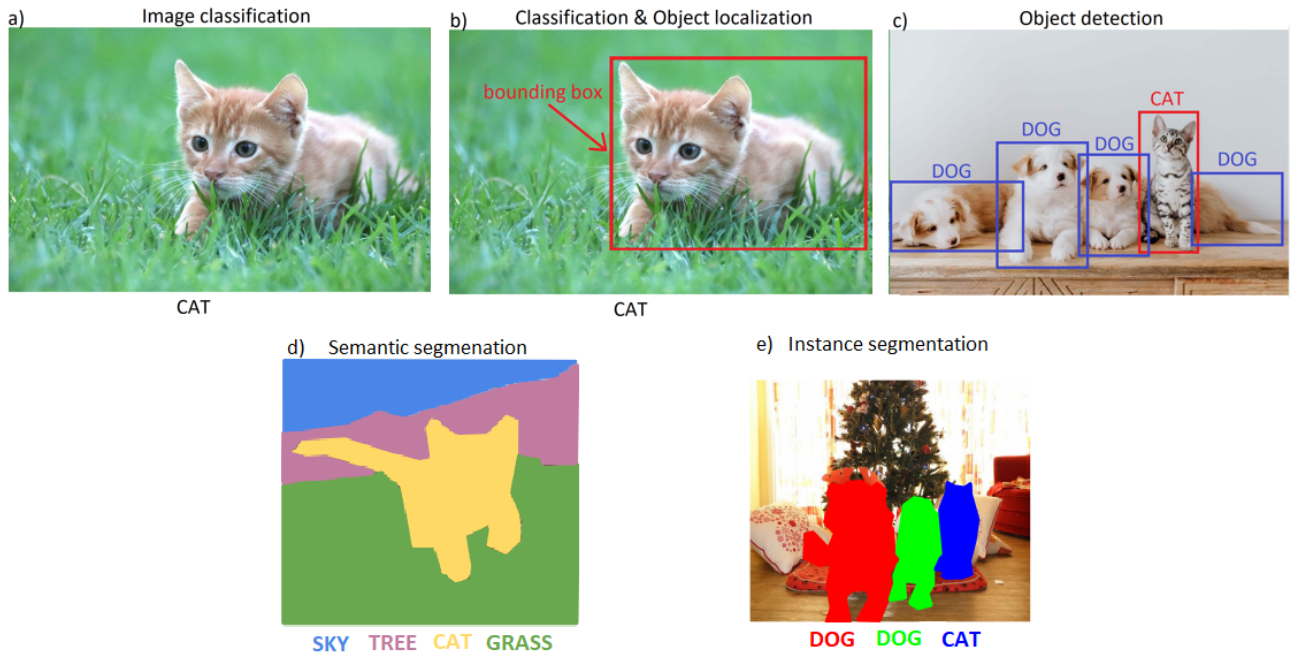


Figure 2.6: a) Image classification output example, b) Classification and object localization output example, c) Object detection output example, d) output of a Semantic segmentation algorithm, and e) output example of Instance segmentation task (Fei-Fei et al., 2018).

In this work, the experiments consist of detecting teeth in panoramic images, this is done by using object detection architectures and, since a bounding box will be classified as tooth or background, it is considered a binary classification problem.

CHAPTER 3

METHODOLOGY

This chapter presents in section 3.1 the characteristics of the dataset used for the experiments, in section 3.2 the IET applied to the images as a pre-processing step and the IQM used to evaluate the outcome of the IET. Section 3.3 displays the description of the object detection architectures and the detection evaluation metrics. Lastly, in section 3.4, the steps followed to perform the experiments are described.

3.1 Dataset

The panoramic images used in this work are digital images that belong to a dataset of 1500 panoramic X-ray images generated by Silva et al. (2018). This is currently the largest publicly free dataset available, originally conceived for semantic segmentation. The 1500 images are organized in 10 categories, not equally distributed, according to the number of teeth in an image, the presence or absence of restored teeth, dental appliances, and implants (Silva et al., 2018). In Table 3.1, the difference between the images belonging to each category is summarized, while Figure 3.1 shows examples of images of 4 of these categories.

Table 3.1: Dataset image classification (Silva et al., 2018).

Category	32 teeth	Restoration	Appliance	Implant
1	yes	yes	yes	no
2	yes	yes	no	no
3	yes	no	yes	no
4	yes	no	no	no
5	less	yes	no	yes
6	more	yes	no	no
7	less	yes	yes	no
8	less	yes	no	no
9	less	no	yes	no
10	less	no	no	no

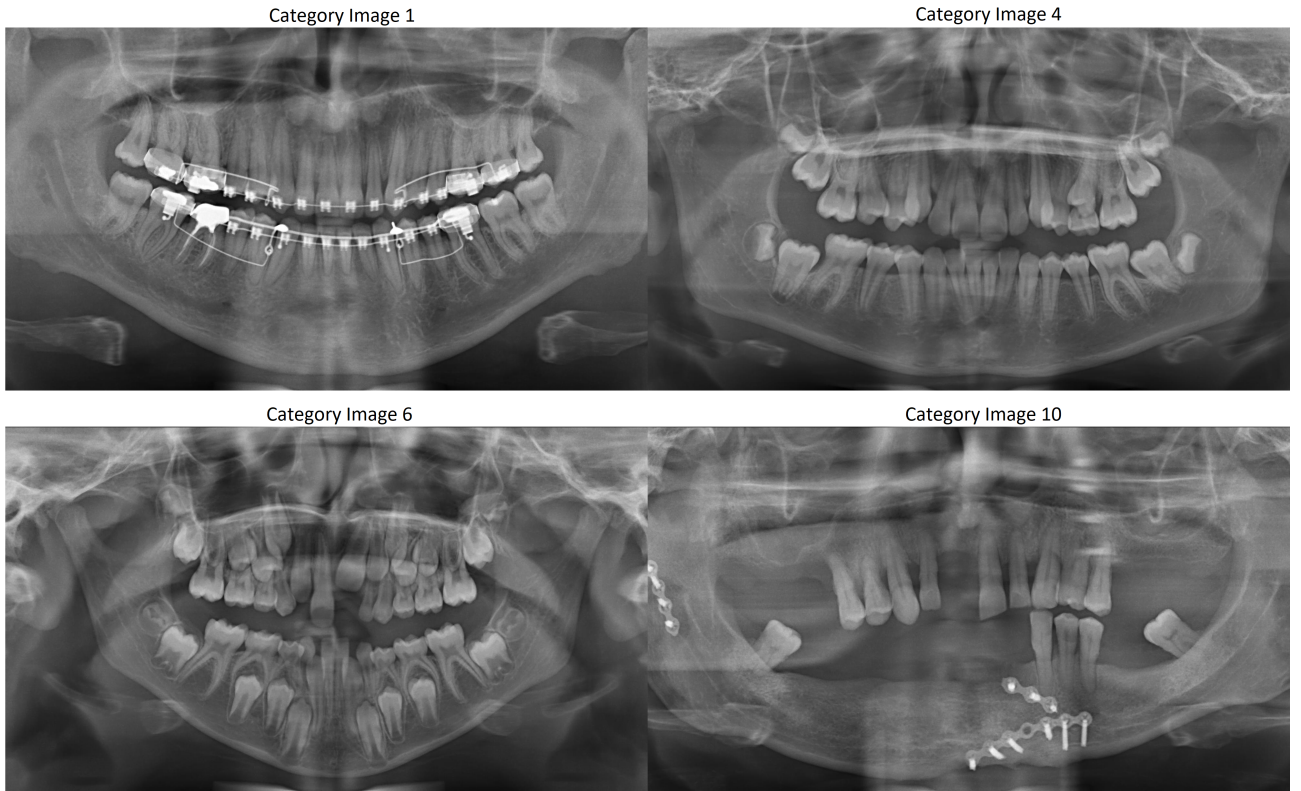


Figure 3.1: Images corresponding to categories 1, 4, 6 and 10. The image of category 1 shows a mouth with 32 teeth, appliance, and restored teeth. The image of category 4 is a mouth with 32 teeth. The image of category 6 shows a children mouth with 45 teeth. Last, in the image of category 10 we can see a mouth with only 14 teeth.

From the 1500 images of the dataset, 300 were selected and divided into 192 for the training set, 48 for the validation set, and 60 for the test set. The images in the dataset are numbered according to the category they belong to, the selected images correspond to the first 50 images in each category. The selection was made trying to maintain good variability in the dataset of 300 images. Also, the images of each category are almost evenly distributed in the training, validation, and testing set. The distribution of the categories in these sets, as well as the total of each category in the total dataset of 300 images, is presented in Table 3.2. In addition

to restorations, appliances, and implants, the condition of teeth vary through the dataset, in which there are also dental tooth fragments, impacted, and embedded teeth. These variations of teeth, as well as the implants, and appliances, were marked with a minimal-size bounding box and a unique label as the ground-truth. Regardless of the tooth condition or type, all bounding boxes were simply classified with the label *tooth*.

Table 3.2: Distribution of categories in the training, validation, and testing set.

Category	Training set	Validation set	Test set	Total
1	21	5	6	32
2	21	5	6	32
3	21	5	6	32
4	21	5	6	32
5	21	5	6	32
6	20	5	6	31
7	22	4	6	32
8	23	6	6	35
9	10	4	6	20
10	12	4	6	22
Total	192	48	60	300

In order to use the dataset for object detection purposes, the bounding box and label must be drawn and set using an annotation tool to later export this annotation in the appropriate format for training the models. The ground-truth annotations were done using VGG Image Annotator (VIA)(Dutta & Zisserman, 2019) and reviewed, edited, and validated by two experienced dentists¹. The annotation process was performed after trimming the panoramic images to suppress toothless regions. With the images trimmed and the annotations ready, ten IET were applied to each set of images. The trimming procedure is described in more detail in the Experimental Setup (see section 3.4), while in the next section, the IET applied to the images as a pre-processing step are explained.

3.2 Image Enhancement techniques (IET)

The automatic analysis of panoramic images is a complex task for being a two-dimensional view of a three-dimensional structure, which results in an image with overlapping anatomical structures, and for the similarity between the mouth tissues (Ehsani Rad et al., 2013; Jader et al., 2018). Furthermore, the final quality of the image depends on the equipment and the patient from whom the image is taken. For a deep learning algorithm, that needs as many images as possible to generalize the problem with an accuracy comparable to that of professionals, the variation patient to patient is also a challenge when the amount of images in the available dataset is small. In addition, dental radiographic images are affected by permanent and temporary artifacts that represent an extra challenge for object detection methods. Tooth

¹A more detailed description of the annotation tool and procedure can be found in the Appendix in A.1

fillings, implants, restorations, metallic materials placed after trauma, among others, represent the permanent artifacts that will affect specially the localization of teeth in panoramic images. While dental appliances, impacted and embedded teeth, overlapping teeth, and missing teeth, represent the temporary artifacts that difficult the automatic localization of teeth (Amer & Aqel, 2015; Ehsani Rad et al., 2013). In order to explore the effects on the detection of teeth using the chosen deep learning methods, ten IET are applied to the dataset as a pre-processing step.

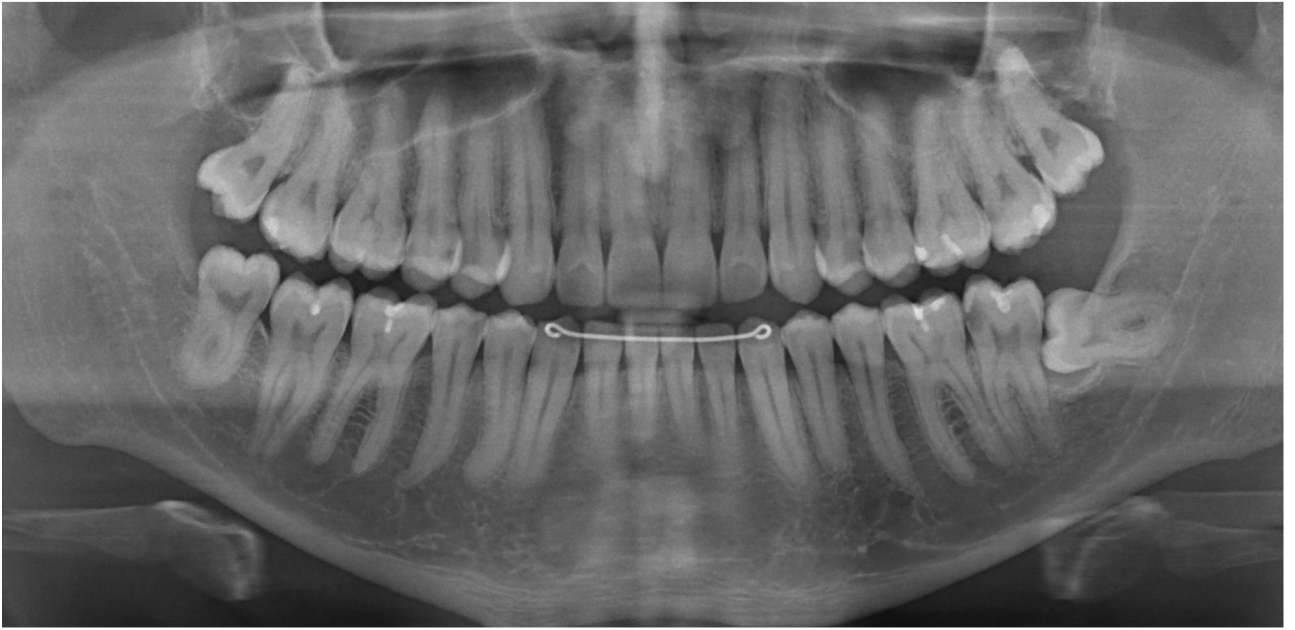
3.2.1 Image Enhancement techniques (IET)

The selected IET are presented in order, according to the purpose of the technique, where denoising or edge sharpening filters are first described, followed by five contrast enhancement algorithms:

- **Anisotropic diffusion:** this method is a smoothing - edge preservation filter in which the enhanced image is given by the solution of a Partial Differential Equation (PDE). Smoothing filters are commonly used for blurring and for removing noise from digital images. Blurring is used to remove small details from the image and to close small gaps between lines or curves. However, smoothing filters tend to alter essential information, like edges of objects, which can also be removed by these filters (Gonzalez & Woods, 2001). In the anisotropic diffusion method, the preservation of the edges is achieved by smoothing in a parallel direction across the region boundaries or edges, which is done by the formulation of a PDE (Sapiro & Ringach, 1996).

This smoothing filter was applied to the images of the dataset using the *anisotropic_diffusion* function provided by **MedPy**, which is a library for medical image processing in Python. This function allows to apply the Perona and Malik (1990) equation by setting the diffusion coefficient and the number of iterations. Figure 3.2 shows a comparison of an original image of category 1 with the filtered image with the anisotropic diffusion method. The difference between the images is almost imperceptible to the human eye, only with a closer look it can be perceived that the anisotropic diffusion image is a blurry version of the original image on the top.

original image



anisotropic diffusion image

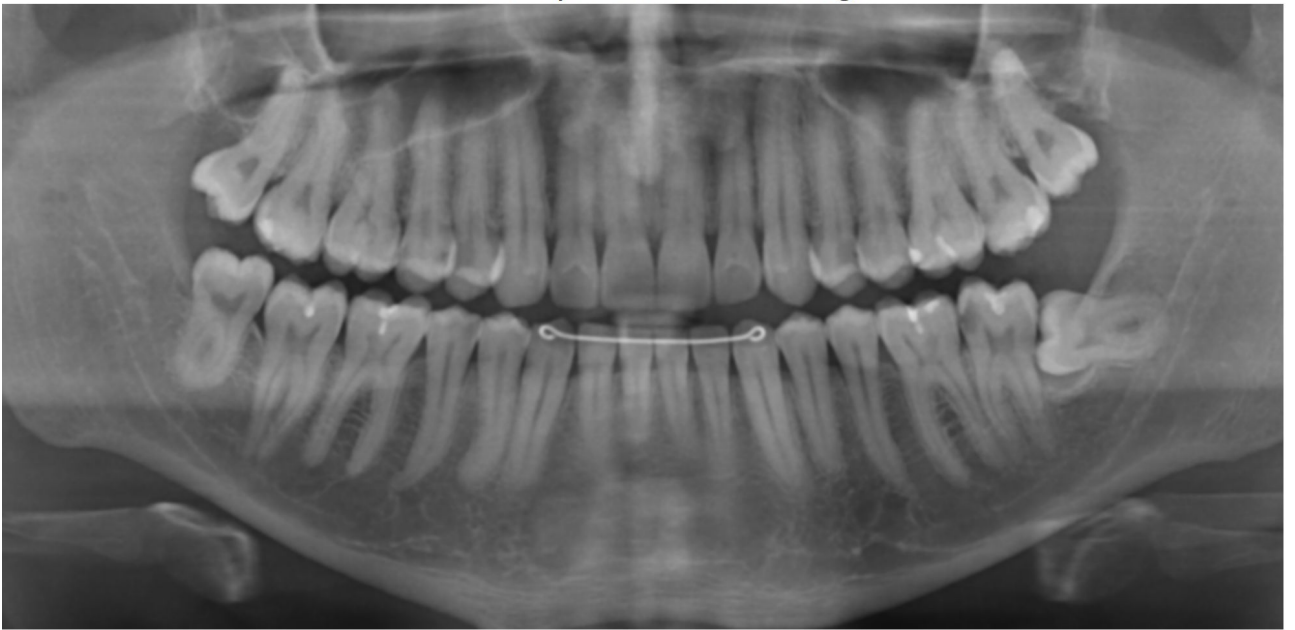


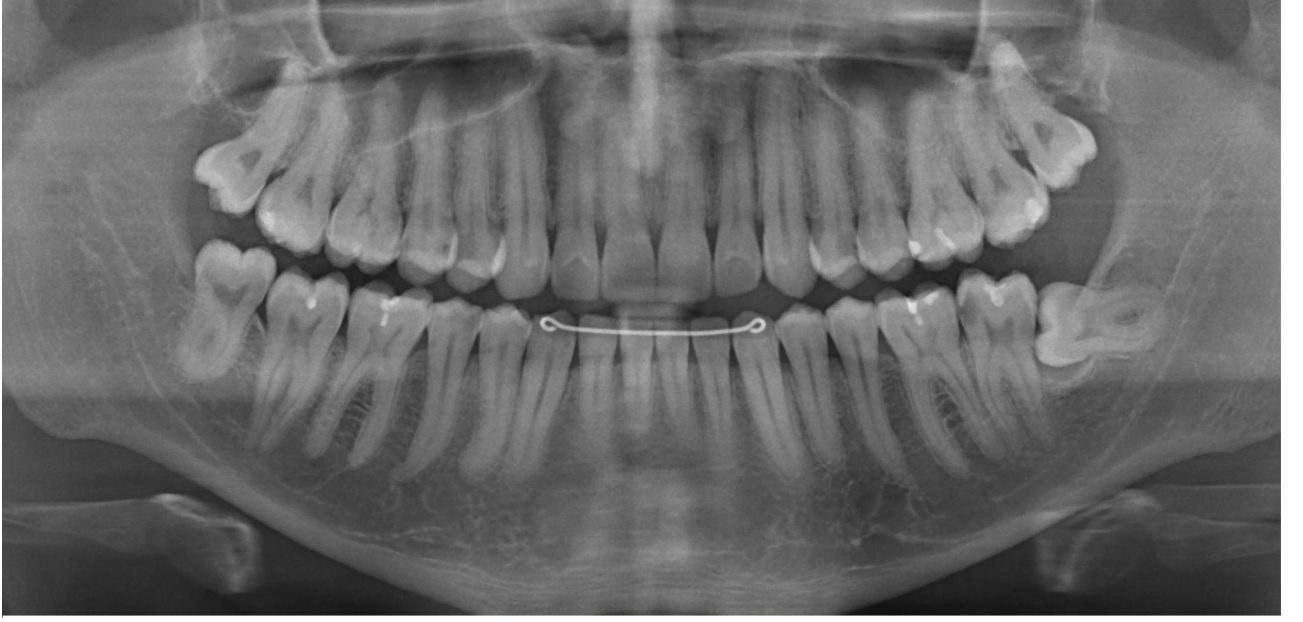
Figure 3.2: The original image, belonging to category 1, on top, which is the filter input, compared to the filtered image with the anisotropic diffusion below.

- **Bilateral filtering** is a non-iterative local filter for smoothing an image while preserving the edges (Tomasi & Manduchi, 1998). Most of the filters focus on the transformation of an image by changing the intensity value of a pixel in a given location using a function of the weighted values of the neighboring pixels. The values of these weights decay with the distance from the center pixel (Tomasi & Manduchi, 1998). Inspired by domain filtering, the authors define the range filtering, where the resulting intensity value of a pixel is average away with weights that decay with the dissimilarity between the center pixel and a nearby pixel. When the weights decay according to dissimilarity and not proximity, the

filter is able of preserving the edges on the image (Tomasi & Manduchi, 1998). Thus, bilateral filtering consists of simply applying both domain and range filtering. When using this combination, the domain filter removes the noise in smooth regions, while in sharp boundaries, the pixel value is replaced by an average between values of similar intensities (Tomasi & Manduchi, 1998).

This filter was applied to the images using the *bilateralFilter* method of the OpenCV library. The method is applied by choosing the size of the filter, the domain spread β_d , and the range spread β_r . The domain spread is chosen based on the desired or needed amount of blurring, where a large value of β_d , combines values at a greater distance from the central pixel, having a greater effect of blurring. In the same manner, the domain spread β_d is set accordingly to the desired level of combination of pixel values (Tomasi & Manduchi, 1998). The size of the filter was set to 7, while β_d , and β_r , were set to the same value, for simplicity, of 150. An original image is compared with the output of the bilateral filtering in Figure 3.3.

original image



bilateral filtering image

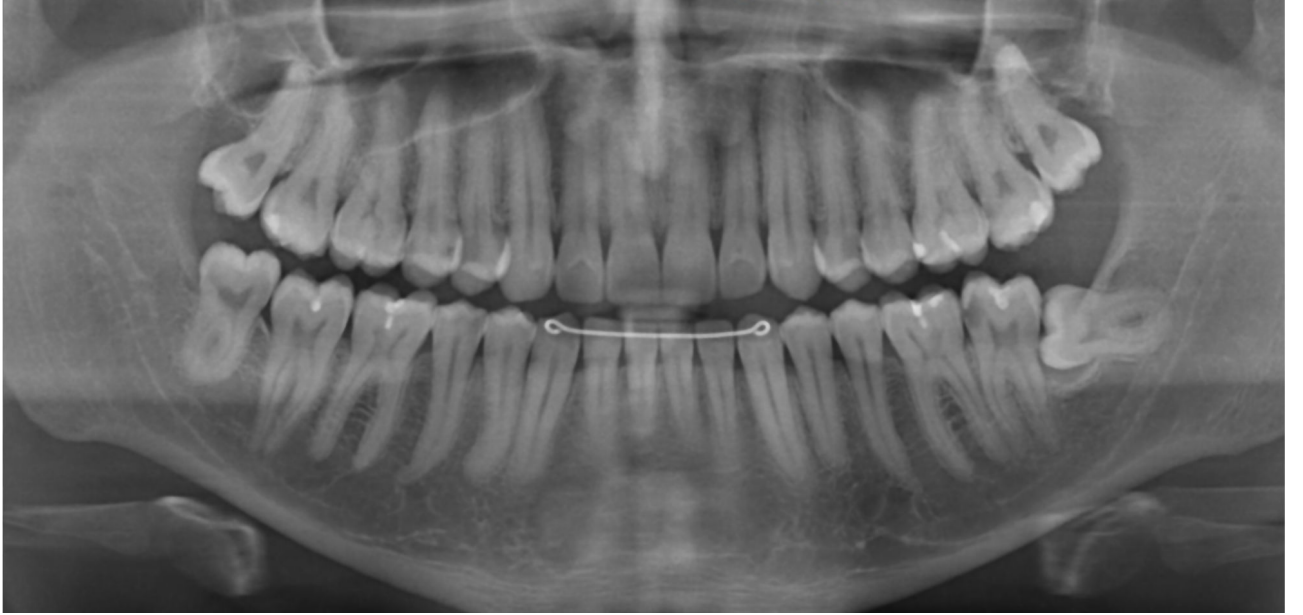


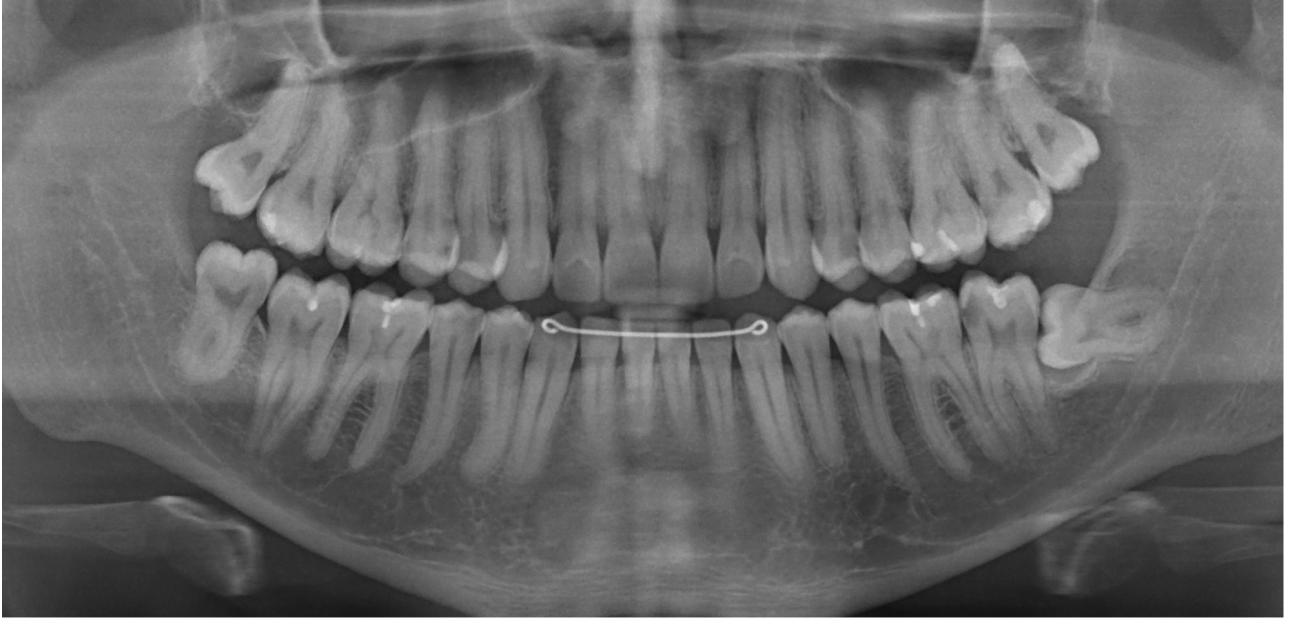
Figure 3.3: The original image on the top, which is the filter input, compared to the bilateral filtering image below.

- **Maximum** and **Minimum** filters, also known as dilation and erosion respectively, are morphological filters where a pixel value is replaced by the maximum or minimum intensity value found in the neighborhood of the pixel. The dilation has the effect of expanding the shapes in an image, while the erosion reduce them. The dilation and erosion for an image I and a structural element (SE) are defined as follows:

$$\begin{aligned} Dil_{SE}(I)(i, j) &= \max_{(u,v) \in SE}(I(i - u, j - v)) \\ \text{and } Ero_{SE}(I)(i, j) &= \min_{(u,v) \in SE}(I(i - u, j - v)) \end{aligned} \tag{3.1}$$

where (i, j) and (u, v) are the coordinates of the pixel in I and SE , respectively (Pierre, 2004). Figures 3.4 and 3.5 shows the comparison of an original image with the output of the Maximum and Minimum filters. The equation on 3.1 were applied twice in the images, this is, dilation and erosion were applied twice consecutively to obtain a visually perceptible change in the images.

original image



maximum image

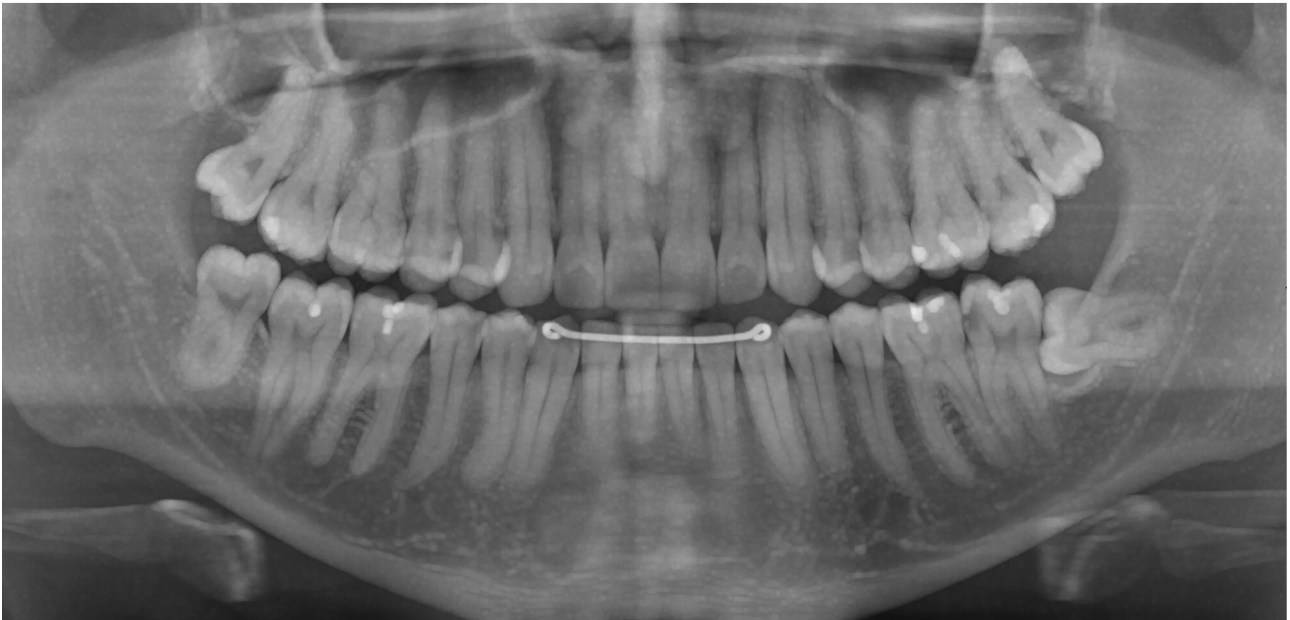
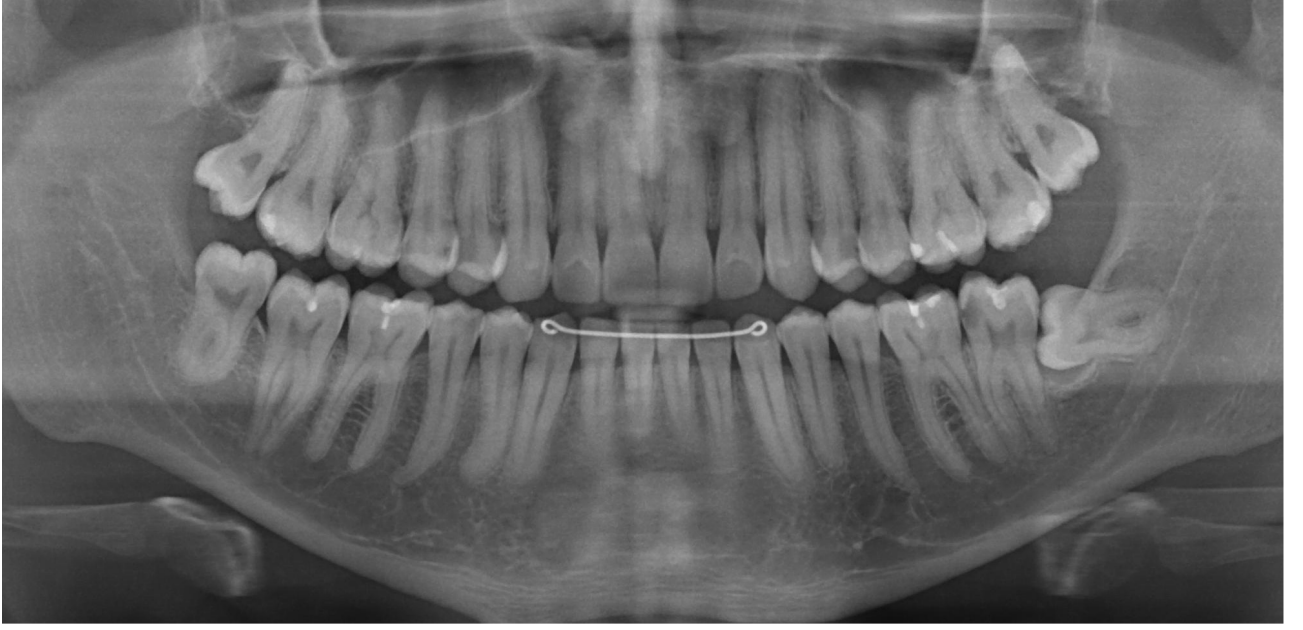


Figure 3.4: The original image on the top is compared with the image generated by the Maximum or dilation filter.

original image



minimum image

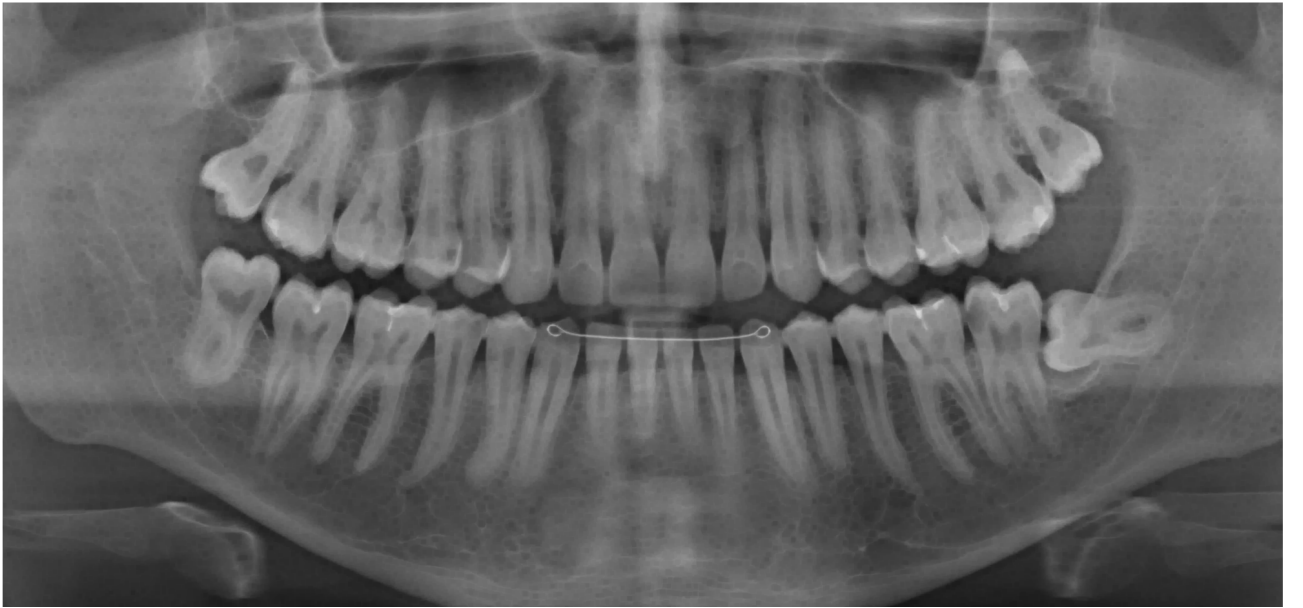


Figure 3.5: The original image on the top is compared with the image generated by the Minimum or erosion filter.

- **Edge enhancement** can be performed by both linear and non-linear techniques. The Laplacian is a commonly used linear operator to enhance edges on an image defined as:

$$\nabla^2 I = \frac{\partial^2 I}{\partial i^2} + \frac{\partial^2 I}{\partial j^2}, \quad (3.2)$$

where $I(i, j)$ is the input image and the second-order derivatives in the i -direction and the j -direction are $\frac{\partial^2 I}{\partial i^2} = I(i+1, j) + I(i-1, j) - 2I(i, j)$ and $\frac{\partial^2 I}{\partial j^2} = I(i, j+1) + I(i, j-1) - 2I(i, j)$, respectively (Gonzalez & Woods, 2001). The image with the edges enhance

$O(i, j)$ is obtained by adding (subtracting) the Laplacian to (from) the image $I(i, j)$:

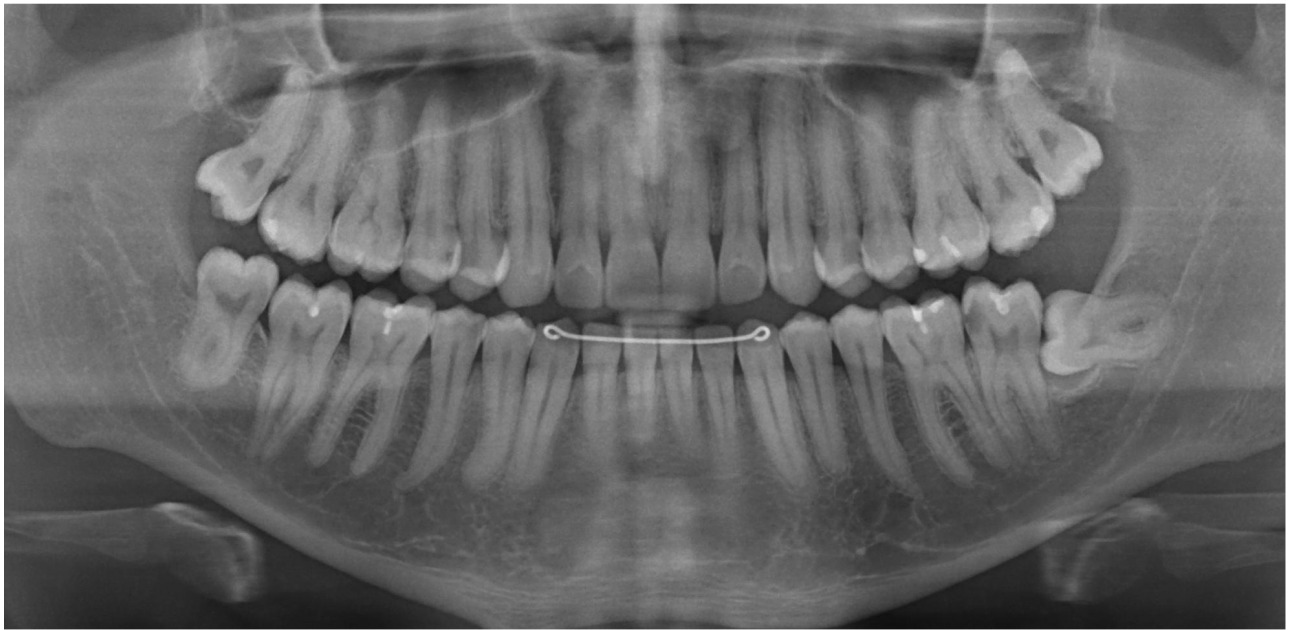
$$O(i, j) = \begin{cases} I(i, j) - \nabla^2 I(i, j) & \text{the center coefficient of the Laplacian mask is negative} \\ I(i, j) + \nabla^2 I(i, j) & \text{the center coefficient of the Laplacian mask is positive.} \end{cases} \quad (3.3)$$

The Laplacian operator is implemented by applying a single mask, whose coefficients are obtained from the resulted equation of substituting equation 3.2 in equation 3.3, to the image to be enhanced (Gonzalez & Woods, 2001). The Laplacian 3×3 mask used to enhance the edges of the panoramic images was:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (3.4)$$

Figure 3.6 shows the image after applying the Edge Enhancement, below, compared to the original image on the top.

original image



Edge enhancement image

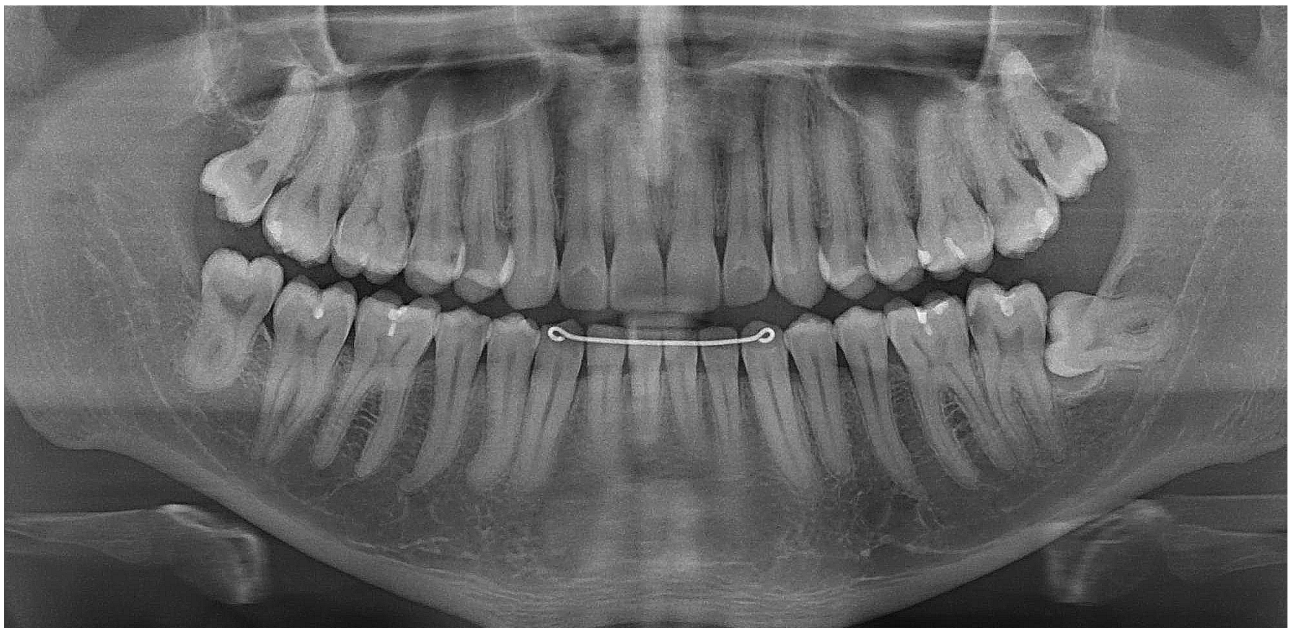


Figure 3.6: Comparison of an original image, on top, with the image after applying the Edge Enhancement technique with the mask show in 3.4 below.

- The **Brightness Preserving Bi-Histogram Equalization (BBHE)**, develop by Kim (1997), is a contrast enhancement algorithm that focus on the preservation of the brightness of the input image as a main goal. Firstly, the BBHE algorithm finds the mean intensity of an input image and uses it to split the input image into two sets of samples. The samples with less than or equal to the value of the mean are used to form a new first sub-image, whereas those values greater than the mean form a second sub-image. The algorithm then independently equalizes the respective histogram of these two subimages, the equalization over the range of samples from the minimum gray level up to the mean

corresponds to the first subimage and the equalization over the range of samples greater than the mean to the maximum gray level of the input image corresponds to the second subimage. The output image is finally formed by unifying the two equalized subimages. By bounding the images around the mean of the input image, the mean of the output image resulted from BBHE depends on the mean of the input image. Thus, preserving better the brightness compared to classic histogram equalization, making it a more suitable option for medical imaging and other applications (Kim, 1997). The comparison between an original image and its BBHE version is shown in Figure 3.7.



Figure 3.7: Comparison of an original image, on the left, with the image after applying the BBHE contrast enhancement technique.

- **Contrast Limited Adaptive Histogram Equalization (CLAHE)** is a popular method that has been used to enhance contrast of medical images with satisfactory results (Pisano et al., 1998; Shome & Vadali, 2011). CLAHE minimizes noise amplification, a problem commonly encountered when applying classic histogram equalization. The method consists of applying Contrast Limited Histogram Equalization (CLHE) to small, non-overlapping regions, called tiles or sub-images, that are obtained by dividing the input image (Sasi & Jayasree, 2013; Shome & Vadali, 2011). In CLHE, a threshold value is used to limit the maximum number of pixels in each interval of the histogram in order to avoid the increase of unwanted gray values, that is, to avoid the application of noise. The spatial resolution of the image is preserved by evenly redistributing the clipped pixels in the histogram. As a final step, the algorithm fused the resulting neighboring tiles by bilinear interpolation (Sasi & Jayasree, 2013). The image resulted from applying CLAHE can be found on the right image in Figure 3.8, where it is compared to the original image on the left of the figure.



Figure 3.8: Comparison of an original image, on the left, with the image after applying the CLAHE contrast enhancement technique.

- The **Geodesic Reconstruction MMCE (GRMMCE)** is a medical image enhancement technique proposed by Mello et al. (2020). The method is a variation of the Multiscale Morphological Contrast Enhancement (MMCE) algorithm proposed by Bai et al. (2012). MMCE uses the classic Top-Hat Transform to obtain the multiscale bright and dark regions of an image. This multiscale bright and dark image regions, along with the multiscale image details between neighboring scales, extracted by the structuring elements with a unique shape and increasing sizes of the Top-Hat Transform, are used to form the final bright and dark image regions. Then, the contrast between the final bright and dark image features is aggrandized to achieve image enhancement (Bai et al., 2012). The GRMMCE method proposes the introduction of the geodesic reconstruction in the Top-Hat transform with the purpose of enhancing the contrast and improving the details of the image while preserving its brightness and avoiding its distortion (Mello et al., 2020). Unlike classic dilation and erosion (see equation 3.1), which use an image and a structuring element as inputs, a geodesic transformation uses two input images and are iterative until stability is reached (Pierre, 2004). Next, the geodesic dilation (Dil_{mk}^y), the reconstruction by dilation (R_{mk}^{Dil}), and the opening by reconstruction (ϕ_{SE}^R), are defined to finally show the Top-Hat transform by geodesic transformation (Bai, 2013; Mello et al., 2020):

$$Dil_{mk}^y(I) = Dil_{mk}^1(Dil_{mk}^{y-1}(I)), \quad (3.5)$$

where mk is a mask image and Dil_{mk}^1 represents the minimum pixels between mk and the dilation of I ; R_{mk}^{Dil} is defined as the iterative geodesic dilation of I with respect to mk until stability is reached:

$$R_{mk}^{Dil}(I) = Dil_{mk}^y(I), \text{ where } Dil_{mk}^y(I) = Dil_{mk}^{y+1}(I). \quad (3.6)$$

The opening by reconstruction is defined as the reconstruction by dilation of I from the erosion of I by SE :

$$\phi_{SE}^R(I) = R_I^{Dil}(Ero_{SE}(I)), \quad (3.7)$$

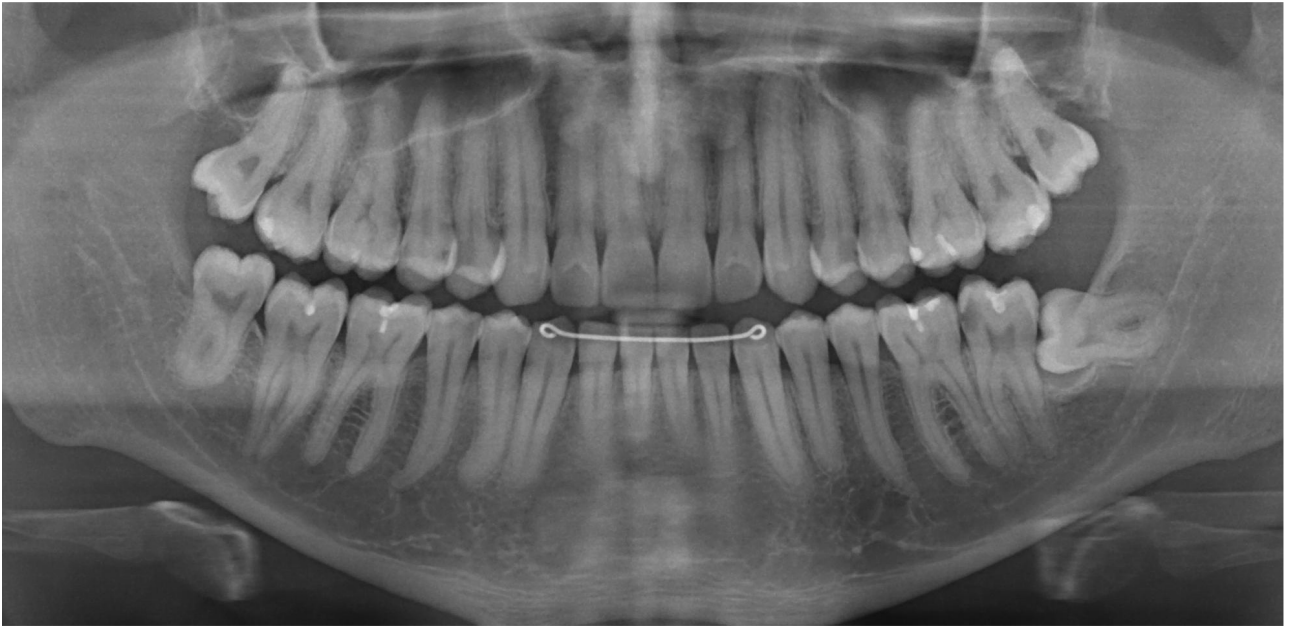
and lastly, by extracting the opening by reconstruction from I we have the white Top-Hat

transform by reconstruction ($WTHR$), which allows to extract the bright features from I :

$$WTHR(I) = I - \phi_{SE}^R(I). \quad (3.8)$$

The black Top-Hat transform, able of extracting the dark features of the image, can be found by defining the geodesic erosion, the reconstruction by erosion and the closing by reconstruction similarly as for the white Top-Hat transform above (Mello et al., 2020). An image enhanced with the GRMMCE technique is shown in the image below of Figure 3.9, while the original image is on top.

original image



GRMMCE image

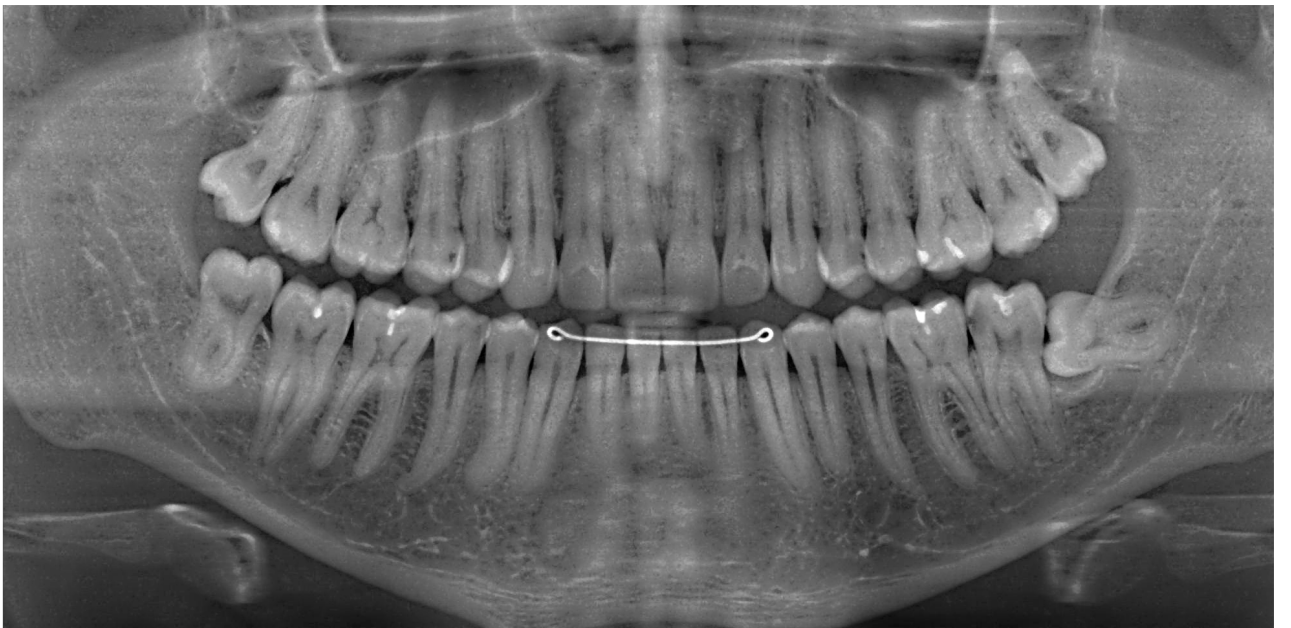
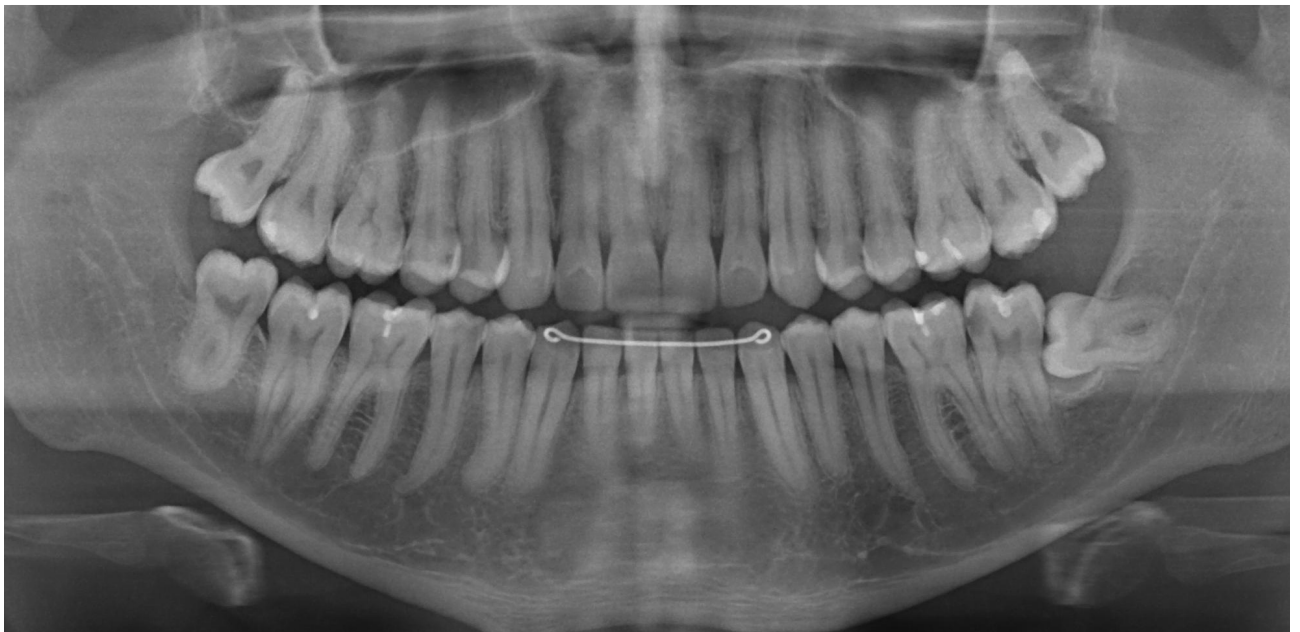


Figure 3.9: Comparison of an original image, on the top, with the image after applying the GRMMCE contrast enhancement technique.

- The **Multiscale Top-Hat Transform (MTHT)** is a contrast enhancement algorithm proposed by Mello et al. (2017) for color images. This method is also a variation of the MMCE algorithm proposed by Bai et al. (2012) for grayscale images. MTHT method uses a slightly different equation to extract the multiscale bright and dark image regions of the image, resulting in an image with a more uniform histogram than an image enhanced with MMCE, which ultimately results in an improvement of the contrast and greater definition of the image details (Mello et al., 2017). In Figure 3.10, a comparison of an original image and its enhanced MTHT version is shown.

original image



MTHT image

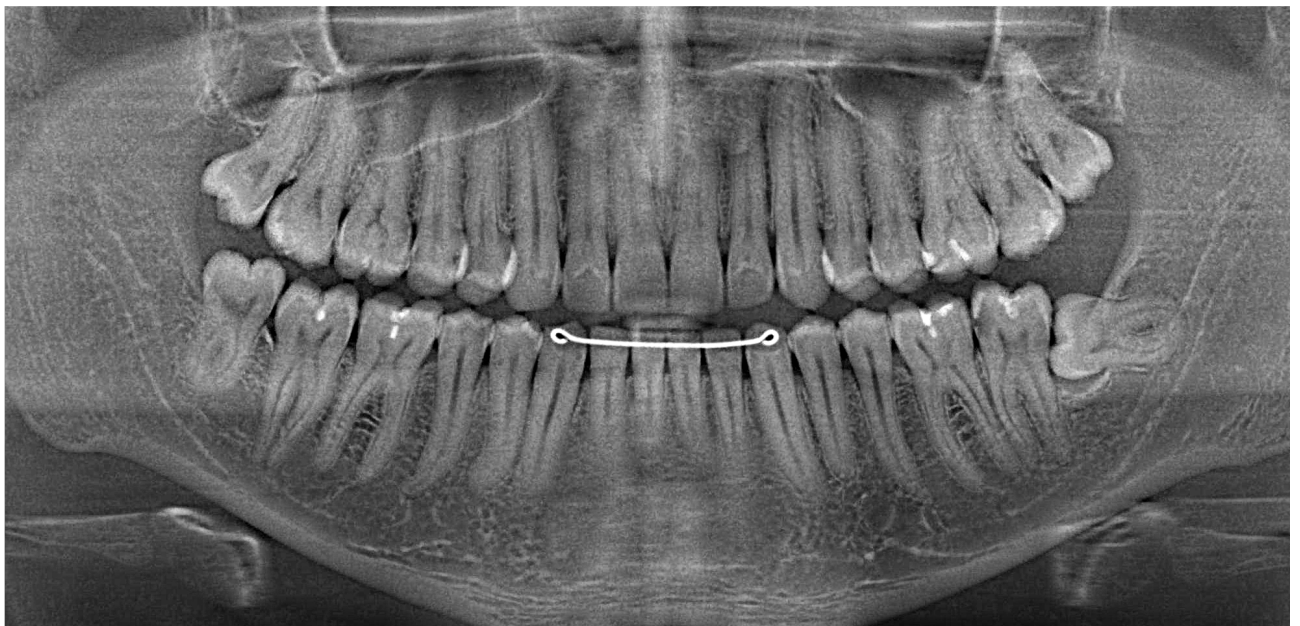
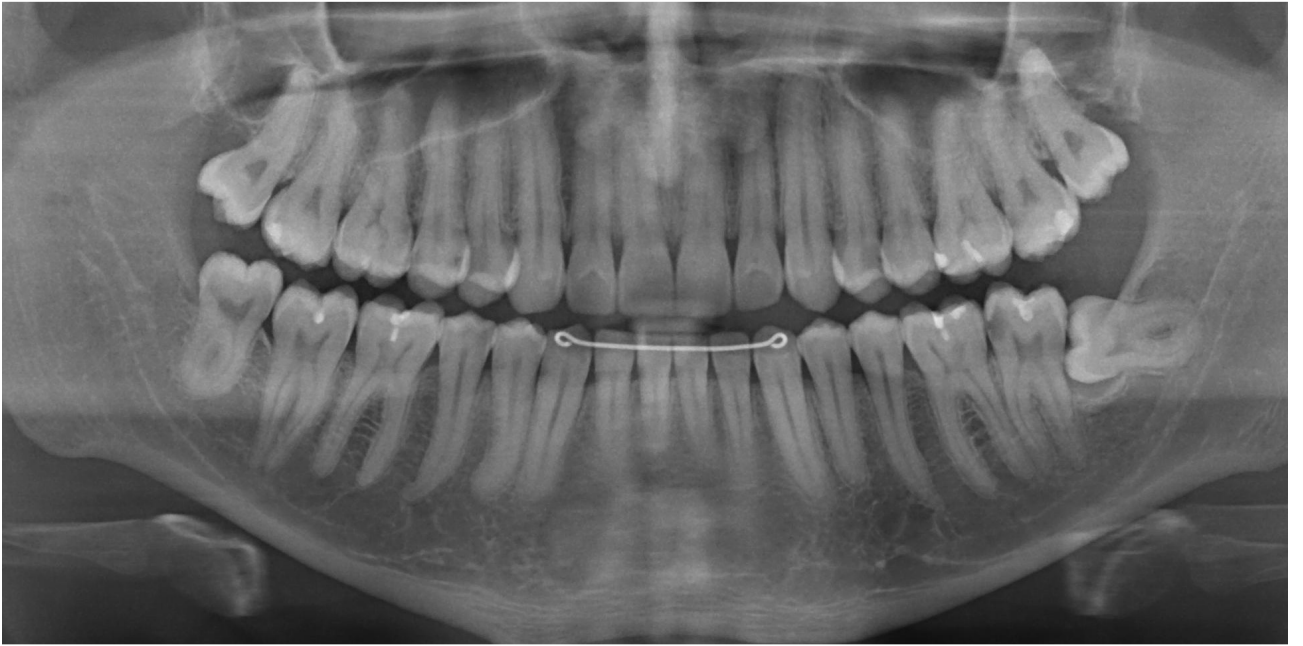


Figure 3.10: Comparison of an original image, on the top, with the image after applying the MTHT contrast enhancement technique.

- The **Quadri-histogram Equalization with Limited Contrast (QHELC)** is a method proposed by Brizuela et al. (2019) for improving the contrast of an image while avoiding the loss of information resulting from the alteration of the average brightness of the input image. For achieving the preservation of the brightness, the algorithm divides the histogram of an image into four sub-histograms before limiting the number of pixels in each bin of these sub-histograms by a certain threshold value or frequency. Then, the clipped values resulted from the mentioned process are evenly distributed along each sub-histogram before joining them again. The final image is obtained by equalizing the image resulting from the union of the clipped sub-histograms (Brizuela et al., 2019). The image resulted from QHELC technique can be found in the below image in Figure 3.8, where it is compared to the original image on the top of the figure.

original image



QHELC image

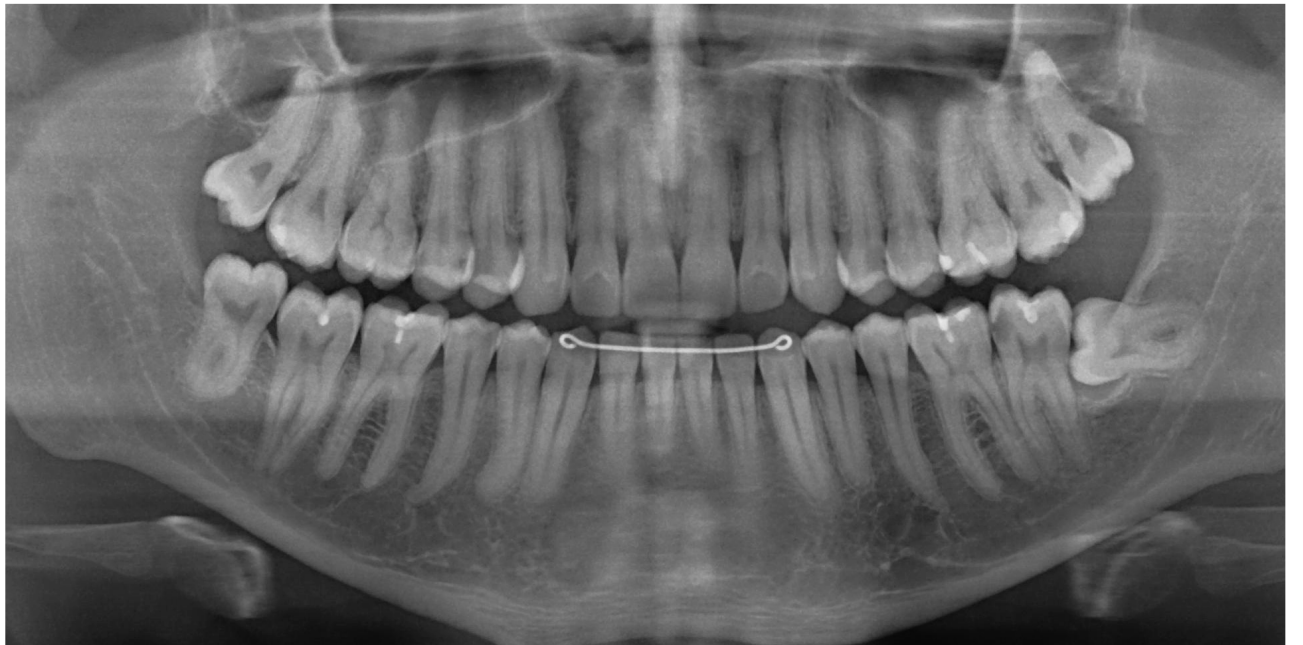


Figure 3.11: Comparison of an original image, on the top, with the image after applying the QHELC contrast enhancement technique.

3.2.2 Image Quality Metrics (IQM)

The IQM used for evaluating the outcome of each IET were Absolute Mean Brightness Error (AMBE), Entropy, Structural Similarity Index Measure (SSIM), Peak Signal to Noise Ratio (PSNR), PSNR with the Human Visual System (PSNR-HVS), Edge Preservation Index (EPI), Average Contrast (AC), and Contrast Improvement Ratio (CIR).

- **AMBE:** this metric is generally used to rate the performance of a method in preserving the original brightness of an image by measuring the absolute difference between the mean of the input image and the output image using the equation:

$$AMBE = |\mu_I - \mu_O|, \quad (3.9)$$

where I and O represents the input and output images, respectively, and $\mu_{(.)}$ denotes the statistical mean or average brightness of the image. A lower value of AMBE is interpreted as a better preservation of the original brightness of the image (S.-D. Chen & Ramli, 2003). The figure 3.12 shows different AMBE values when comparing an original image with three images with higher brightness.

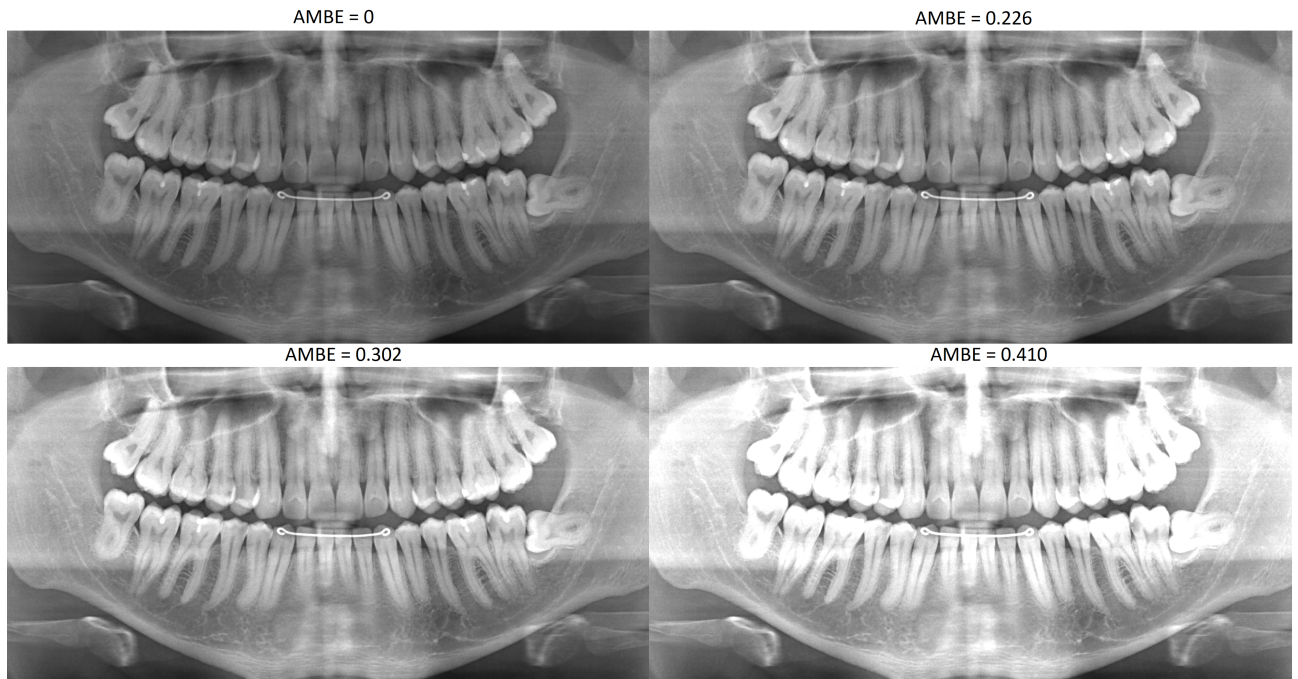


Figure 3.12: The upper left image shows the original image with AMBE equal to zero, on the left, the same image with 50% more brightness has $AMBE = 0.226$. The images in the bottom present 100% and 150% more brightness than the original image, with AMBE values of 0.302 and 0.410, respectively.

- **Entropy:** In the field of information theory, the level of uncertainty of a source of information is usually measured with Shannon's entropy. A source of information with a high level of disorder indicates that the events are less predictable and will result in a high entropy (Shannon, 1948). Shannon's entropy is given by:

$$E(I) = - \sum_{k=0}^{L-1} p(k) \cdot \log_2(p(k)), \quad (3.10)$$

where $p(k)$ is the probability of occurrence of the gray level or intensity level k , obtained from the normalized histogram of the image I that has L discrete gray levels. In image

processing, the discrete entropy allows to quantify the information content of an image by providing the number of bits needed to encode the image data. According to the value given by this metric, the more detailed an image is, the higher its entropy will be (Roman et al., 2019). Figure 3.13 an original image of category 5 is used to show how three IET change its entropy value.

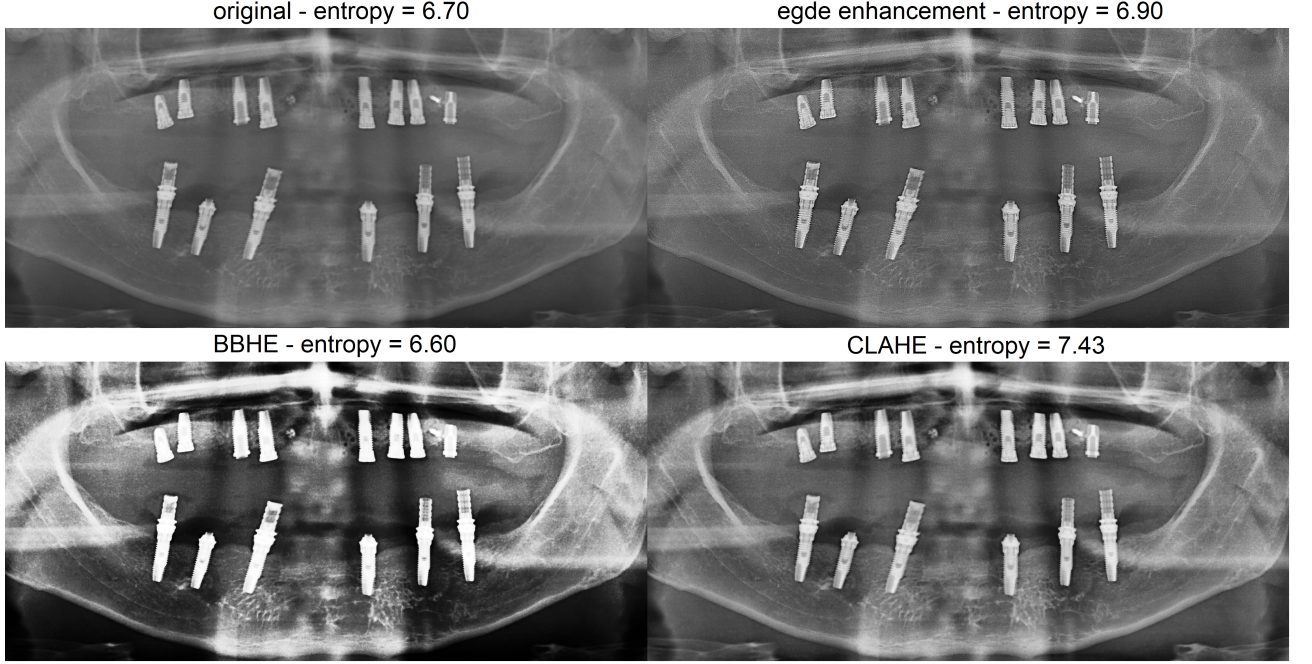


Figure 3.13: The original image of category 5 present an entropy of 6.70, the image on its right is an image enhanced with the edge enhancement technique that has an entropy of 6.90. The images in the lower part were improved with the BBHE and CLAHE techniques and present entropy values of 6.60 and 7.43, respectively.

- **SSIM:** The structural similarity between two images can be measured by the well-known SSIM, developed by Z. Wang et al. (2004). SSIM measures the distortion introduced into an output image by comparing the structure, the luminance, and contrast of the input image with the output image using:

$$SSIM(I, O) = l(I, O) \cdot c(I, O) \cdot s(I, O), \quad (3.11)$$

where

$$l(I, O) = \frac{2\mu_I\mu_O + C_1}{\mu_I^2 + \mu_O^2 + C_1}, \quad (3.12a)$$

$$c(I, O) = \frac{2\sigma_I\sigma_O + C_2}{\sigma_I^2 + \sigma_O^2 + C_2}, \quad (3.12b)$$

$$s(I, O) = \frac{cov(I, O) + C_3}{\sigma_I^2\sigma_O^2 + C_3}. \quad (3.12c)$$

The luminance comparison function, $l(I, O)$, measures the propinquity of the mean lu-

minance of two images, μ_I and μ_O . The value of $l(I, O) = 1$ if and only if $\mu_I = \mu_O$. The second term in equation 3.11 is the contrast comparison function, given by 3.12b, this function provides a value which represents the similarity between the contrast of the input image and the output image. $c(I, O) = 1$ only if the contrast of each image, given by the standard deviation σ_I and σ_O , are equal. The correlation coefficient between the images I and O is measured by the structure comparison function $s(I, O)$, given by 3.12c where $cov_{(I,O)}$ is the covariance between I and O . $SSIM \in [0, 1]$ where zero means no correlation between the images, and 1 means they are the same image. C_1 , C_2 and C_3 are positive constants present in the functions to prevent a denominator equal to zero (Horé & Ziou, 2010; Z. Wang et al., 2004). Figure 3.14 shows an original image and three enhanced versions presenting different values of SSIM.

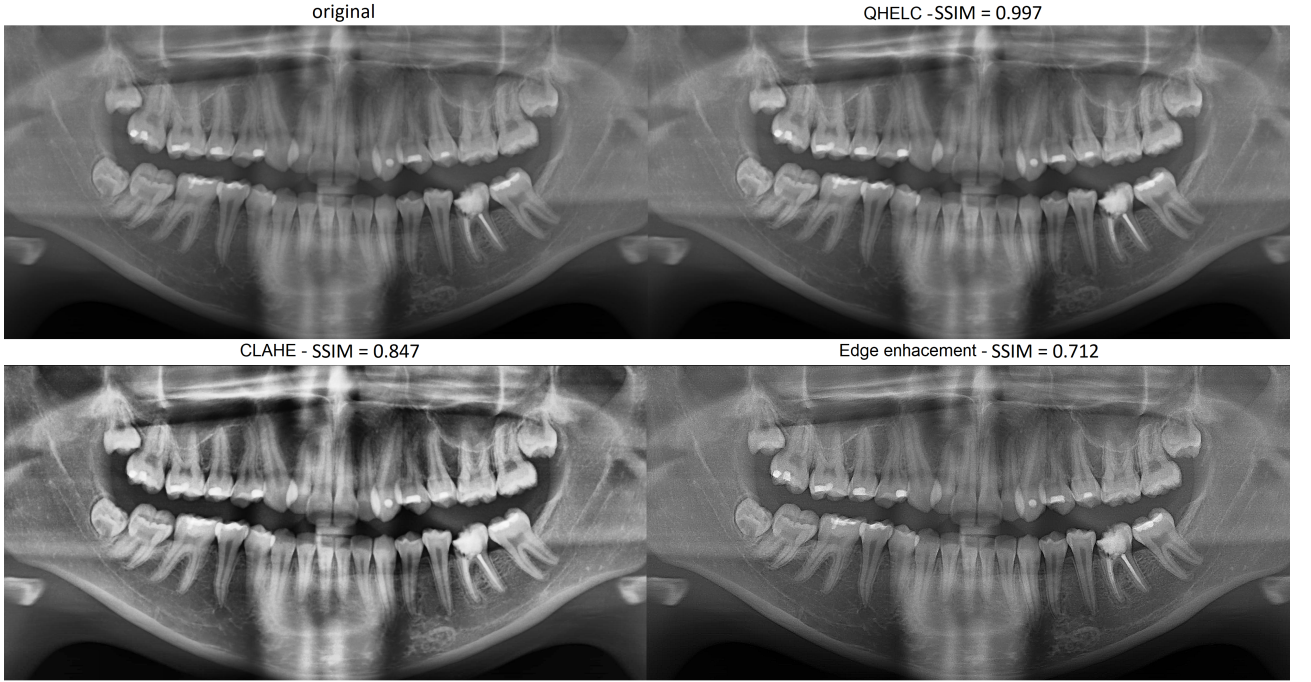


Figure 3.14: An original image of category 8, in the upper left corner, is compared with its enhanced versions with SSIM values of 0.997, upper right corner, 0.847, bottom left corner, and 0.712 in the bottom right corner of the figure.

- **PSNR:** The PSNR between an input image I and its output image O , both of size $M \times N$, is given by:

$$PSNR(I, O) = 10 \cdot \log_{10} \left(\frac{(L-1)^2}{MSE(I, O)} \right), \quad (3.13)$$

where $L - 1$ is the maximum gray level and MSE is the Mean Square Error:

$$MSE(I, O) = \frac{1}{M \cdot N} \sum_{i=1}^M \sum_{j=1}^N (I(i, j) - O(i, j))^2. \quad (3.14)$$

The introduction of a high level of noise to the output image would result in a low PSNR value, while if MSE approaches zero, PSNR will approach infinity. This means that higher

preservation of the image's quality would yield a higher PSNR value (Bai et al., 2012; Mello et al., 2020).

- The **PSNR-HVS**, proposed by Egiazarian et al. (2006), is a modified version of the PSNR metric where the HVS is considered. The HVS is sensitive to low frequency distortion, changes in the contrast and noise in images. Hence, the authors proposed PSNR-HVS and, according to the results shown in their work, it provides a better correlation with the Mean Observer Score (MOS) than PSNR. The MOS quantifies a subjective quality evaluation test, by averaging over the individual ratings given by a human for a certain stimulus, in this case, images. Equation 3.15 shows the average over a total of Y ratings:

$$MOS = \frac{\sum_{y=1}^Y rating_y}{Y}. \quad (3.15)$$

PSNR-HVS was computed using the formulation presented in Egiazarian et al. (2006), where the authors modified the MSE to take into account the HVS according to the approach described by Nill (1985). The images in Figure 3.15 present different values of PSNR and PSNR-HVS of images enhanced with three IET.

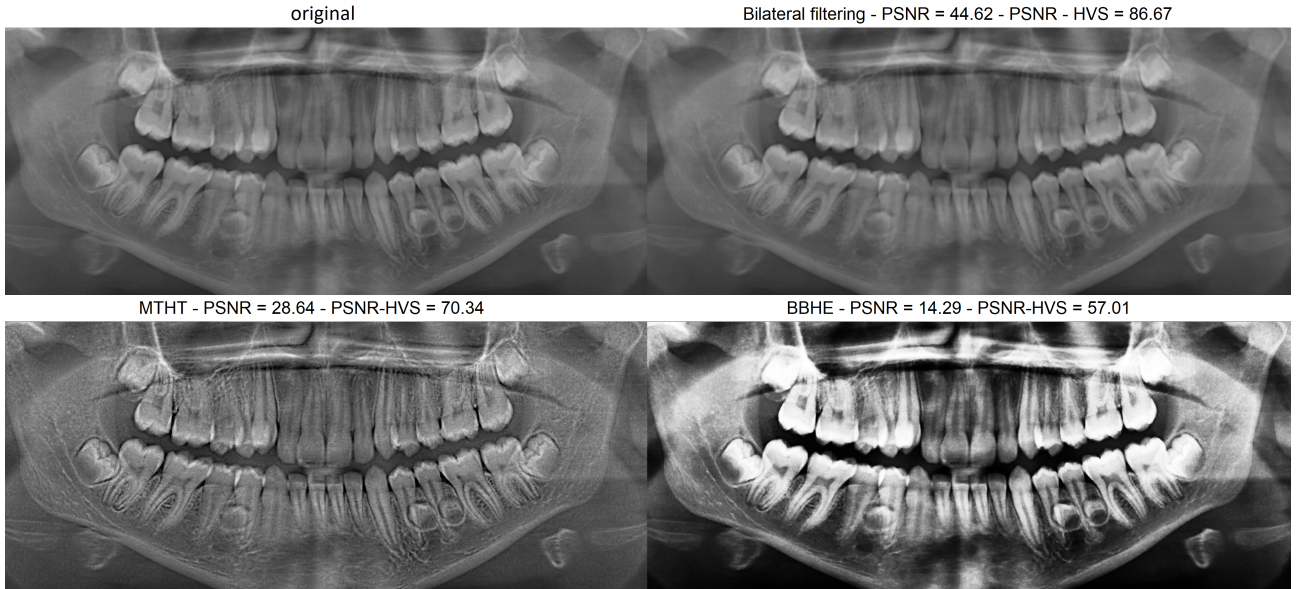


Figure 3.15: An original image of category 6, in the upper left corner, is compared with its enhanced versions presenting PSNR / PSNR-HVS values of 44.62 / 86.67, upper right corner, 28.64 / 70.34, bottom left corner, and 14.29 / 57.01 in the bottom right corner of the figure. From these values, we can conclude that PSNR and PSNR-HVS are directly proportional.

- **EPI** is an index used to ensure that the edges are preserved after applying a denoising filter to an image. Sattar et al. (1997) measured the preservation of the edges with:

$$\lambda = \frac{\Gamma(\Delta I - \mu_{\Delta I}, \Delta O - \mu_{\Delta O})}{\sqrt{\Gamma(\Delta I - \mu_{\Delta I}, \Delta I - \mu_{\Delta I}) \cdot \Gamma(\Delta O - \mu_{\Delta O}, \Delta O - \mu_{\Delta O})}}, \quad (3.16)$$

where ΔI and ΔO are the high pass filtered versions of I and O , obtained with a 3×3 pixel standard approximation of the Laplacian operator. $\mu_{\Delta I}$ and $\mu_{\Delta O}$ are the mean values in the Region of Interest (RoI) of ΔI and ΔO , respectively, and

$$\Gamma(I_1, I_2) = \sum_{(i,j) \in RoI} I_1(i, j) \cdot I_2(i, j). \quad (3.17)$$

The value of EPI will be closer to unity if the filtered or enhanced image is similar to the input image (Sattar et al., 1997). Figure 3.16 shows an image compared with its enhanced versions, which present different values of EPI.

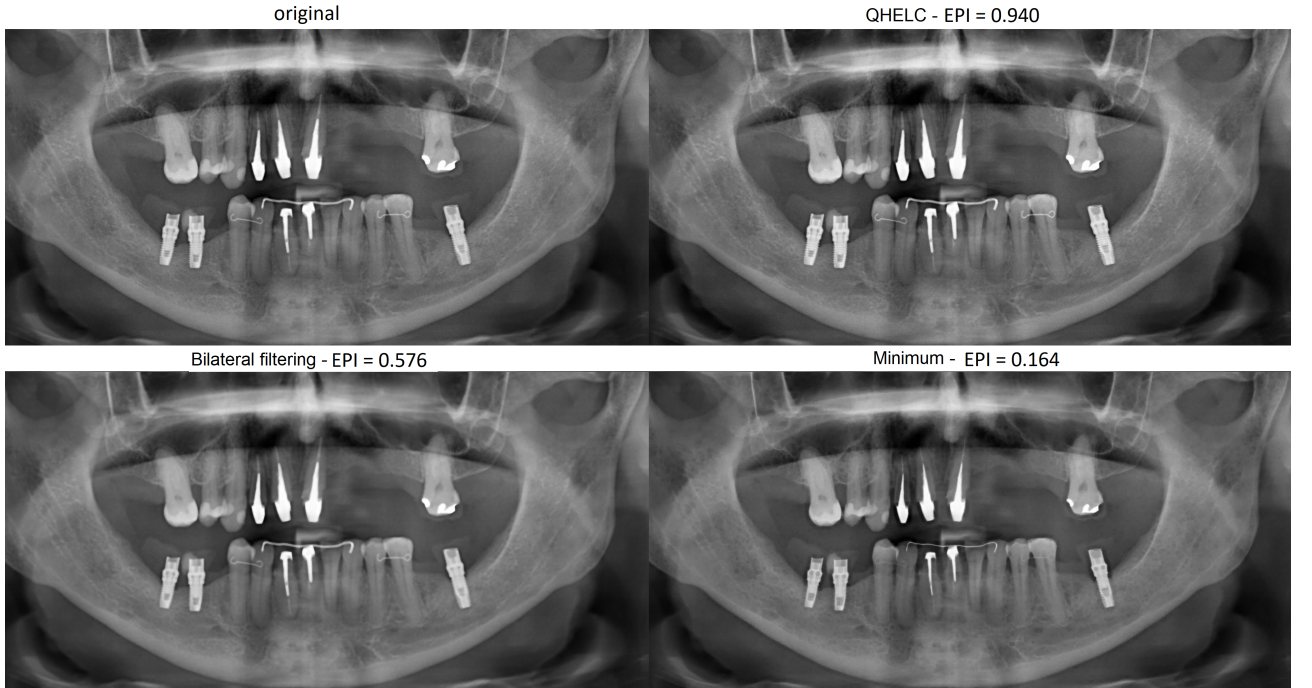


Figure 3.16: An original image of category 5, in the upper left corner, is compared with its enhanced versions with EPI values of 0.940, upper right corner, 0.576, bottom left corner, and 0.164 in the bottom right corner of the figure.

- **AC** is a metric commonly used to compute the global contrast in grayscale images after a contrast enhancement algorithm is applied. AC is such that the higher its value, the greater the contrast of the grayscale image is. This metric is defined as:

$$AC = \sqrt{\sum_{k=0}^{L-1} (k - \mu_O)^2 \times p(k)}, \quad (3.18)$$

where O denotes the output image, $L - 1$ represents the maximum gray intensity level for the image, μ_O is the average brightness of O and $p(k)$ is the probability of occurrence of the gray intensity k (Lim et al., 2013; Menotti et al., 2007). Figure 3.17 shows an image compared with its enhanced versions, which present different values of global contrast.

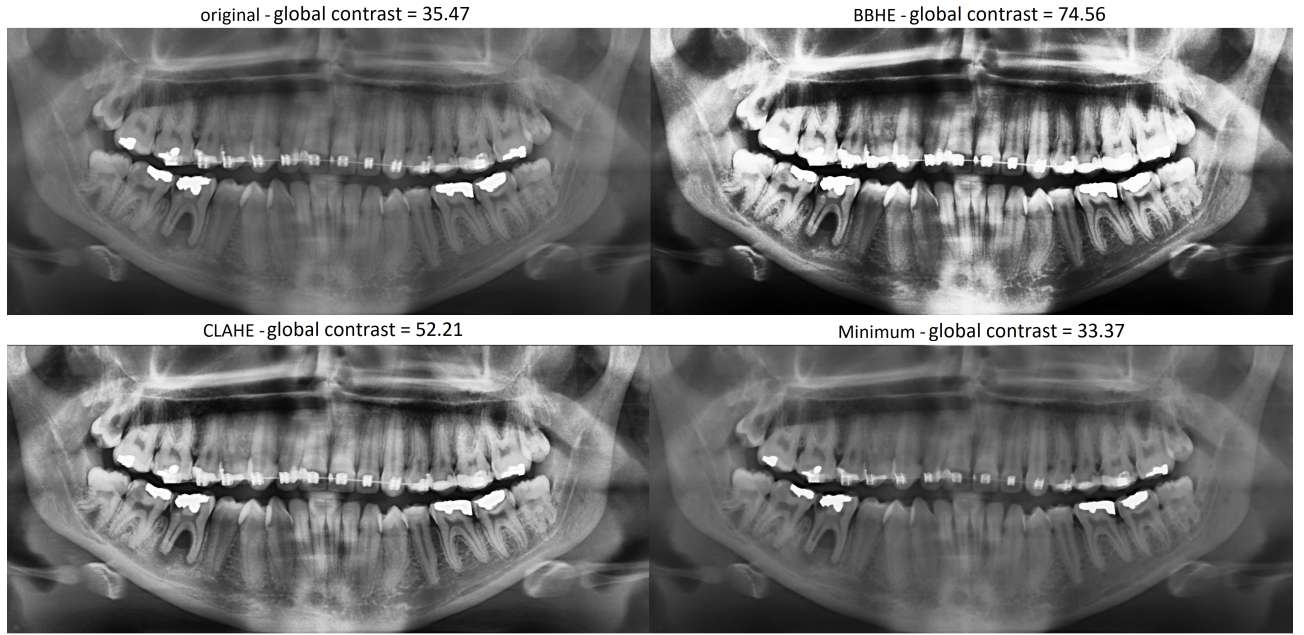


Figure 3.17: An original image of category 1, in the upper left corner, is compared with its enhanced versions with contrast values of 74.56, upper right corner, 52.21, bottom left corner, and 33.37 in the bottom right corner of the figure.

- For the local contrast, **CIR** was the selected metric. According to the definition proposed in (Gordon & Rangayyan, 1984), the local contrast $lc(i, j)$ is the difference between the intensity k of the pixel located at (i, j) and the average or arithmetic mean m between the 8 neighbors in the 3×3 window centered on (i, j) :

$$lc(i, j) = \frac{|k - m|}{(k + m)}, \quad (3.19)$$

CIR is defined as the ratio of the original image and the enhanced image within a region of interest RoI , which can be simple the whole image, as follows (Y.-P. Wang et al., 2003):

$$CIR = \frac{\sum_{(i,j) \in RoI} |lc(i, j) - \hat{lc}(i, j)|^2}{\sum_{(i,j) \in RoI} lc(i, j)^2}, \quad (3.20)$$

where lc and $\hat{lc}$ are the values of the local contrast of the original image and the image resulted from applying an enhancement technique, respectively (Y.-P. Wang et al., 2003). The value given by this performance metric will be such that the higher the difference between the local contrast of the images, the higher the value of CIR. Figure 3.18 shows an image compared with its enhanced versions, which present different values of CIR.

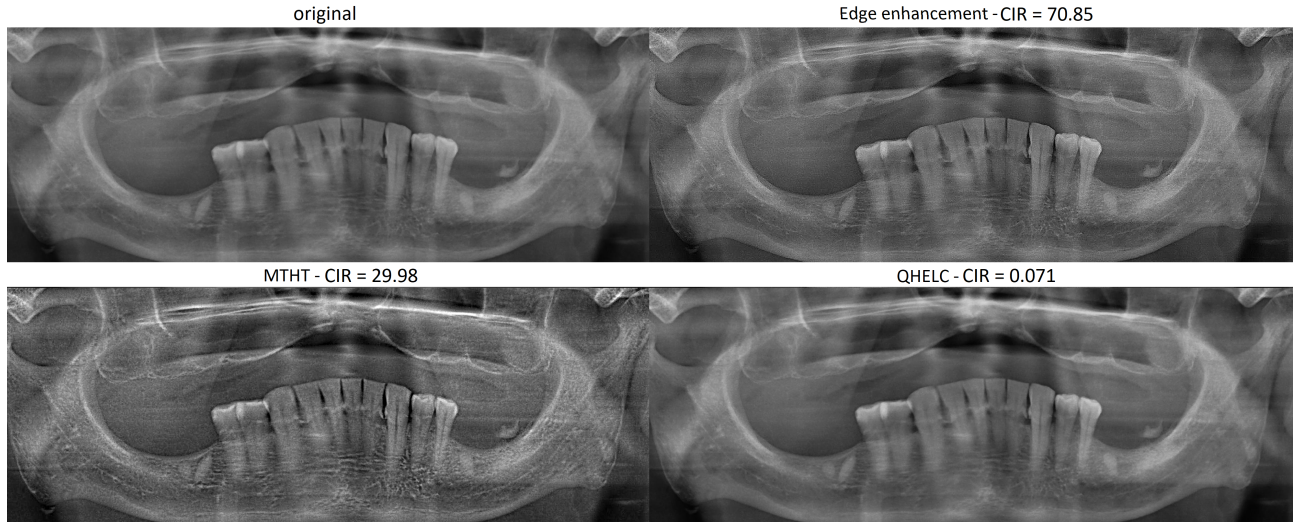


Figure 3.18: An original image of category 10, in the upper left corner, is compared with its enhanced versions with CIR values of 70.85, upper right corner, 29.98, bottom left corner, and 0.071 in the bottom right corner of the figure.

The ten groups of enhanced images were used to train ten models using each chosen architecture, while the IQM were measured in the 60 images of the testing set to later use these values to find the correlation with the detection resulted from the inference performed by the models. The chosen architectures are presented in the following section, whereas the IQM values are furthered discuss in section 4.1.

3.3 Object detection architectures

This section presents the general description of the architectures chosen and used for running the experiments, and the metrics used to evaluate the detection performance of the models.

3.3.1 Selected architectures

Among the object detection deep learning architectures, the two-stages architectures Faster R-CNN by Ren et al. (2015) and Cascade R-CNN by Cai and Vasconcelos (2019), along with the one-stages architectures RetinaNet by Lin et al., 2018, FCOS by Tian et al. (2019), and YOLOv₃ by Redmon and Farhadi (2018) were selected. These architectures are briefly described in this section.

- Faster R-CNN by Ren et al. (2015), is a two-stage architecture that belongs to the Region-based CNN (R-CNN) family along with its predecessor Fast R-CNN. Faster R-CNN was proposed as a faster alternative by being composed of a feature extractor network, a Region Proposal Network (RPN), and an object detector. Figure 3.19 shows an illustration of the Faster R-CNN. The feature extractor network is in charge of extracting features from the input image having as an output a feature map, which is then used by the RPN.

The RPN is constructed by adding convolutional layers to the detector network for predicting object proposals by regressing region boundaries (called anchors) with different scales and aspect ratios. A set number of anchors is regressed for each location in the feature map after a classification score (object/non-object) is assigned to each anchor. This network was proposed to improve the generation of RoI proposals, which was done by the selective search algorithm in R-CNN and Fast R-CNN. The authors unified their proposed RPN with Fast R-CNN as the object detector network, where the RoI Pooling first reshapes all the different sizes of object proposals output by the RPN in a single feature map for the object detector. Then, the object detector classifies the object proposals as objects or background (positives or negatives samples) predicting also the offset values for the bounding boxes. The RPN and detector outputs consist of the 4 coordinates of each anchor and the scores of the probability of object or background for each anchor (Ren et al., 2015).

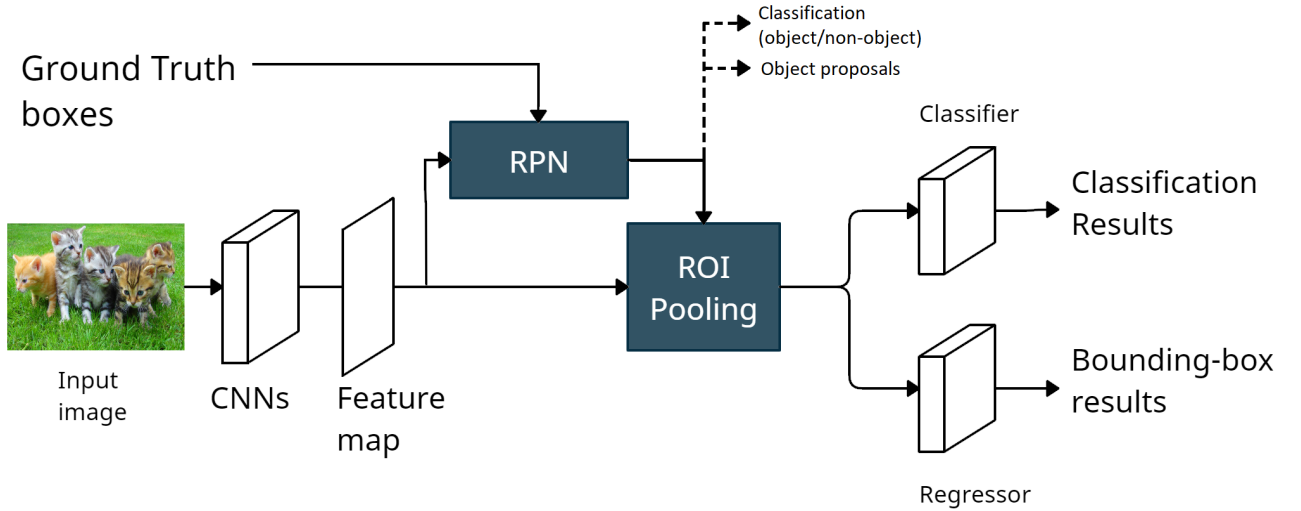


Figure 3.19: Illustration of Faster R-CNN architecture (Ren et al., 2015).

- Cascade R-CNN by Cai and Vasconcelos (2019), is a multi-stage architecture composed of a sequence of detectors, where each detector in the cascade is trained with an increased Intersection over Union (IoU) threshold. The IoU threshold is commonly used to classify positives and negatives samples in anchor-based object detection methods, such as Faster R-CNN. Moreover, the quality of the detector is defined by this threshold, which is usually set to 0.5. Although this value is prone to generate poor quality detections, when larger thresholds are used, the detectors tend to over-fitting, which also diminishes the performance.

In the Faster R-CNN architecture, the object proposals are classified into object (positive) and non-objects (negative), and the positives samples are given a final classification score and a bounding box. In Cascade R-CNN, the bounding box resulted from the first detector is used as an input for the second detector for obtaining a more accurate bounding box

to be the input to the third detector, and so on. When gradually adjusting the boxes, each stage of the cascade becomes more selective against close false positives for training the next stage. Training the detectors with higher IoU thresholds without overfitting is possible because the resampling operation allows the availability of numerous anchors for training all stages. This cascading operation is applied to the bounding box regressor not to the classification head, although it is possible. The inference is also performed using the cascade procedure, which enables higher detection accuracies (Cai & Vasconcelos, 2019).

Cai and Vasconcelos (2019) used the RPN proposed by Ren et al. (2015) for detecting the object proposal in Cascade R-CNN. An illustration of the architecture is shown in Figure 3.20.

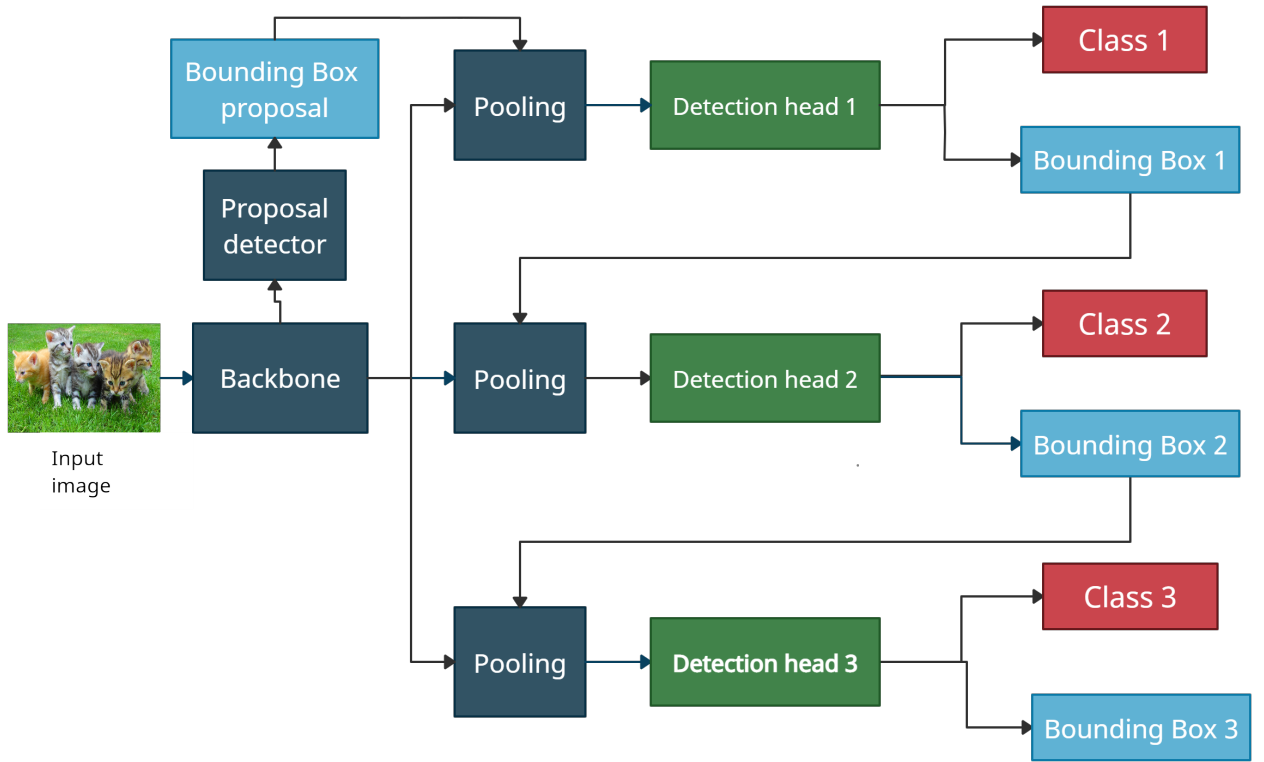


Figure 3.20: Illustration of Cascade R-CNN architecture (Cai & Vasconcelos, 2019).

- RetinaNet by Lin et al. (2018), is a one-stage detector that obtains results comparable to two-stage detectors by addressing the severe foreground-background class imbalance of dense object detection. Unlike two-stages detectors like the R-CNN family, which only evaluates candidates that are previously proposed, one-stage detectors commonly evaluate more than ten thousand candidates during classification. Among these thousands of candidates, the number of negatives samples is much greater than the number of positives samples, thus most locations are easy negatives that are considered but do not contribute useful information. The final contribution of the losses of these easy examples, although being small, exceeds the contribution of the losses of hard examples, affecting the

performance of the detector. RetinaNet handles this class imbalance by using a proposed *focal loss function* instead of the classic cross entropy (CE) loss. The focal loss function is a variation of the CE loss (see equation 3.21), where a modulating factor is added along with a weighting parameter (see equation 3.22), aiming to focus the model's loss on hard examples by down-weighting the contribution of easy examples (Lin et al., 2018).

The CE loss and focal loss function for binary classification is written as follows:

$$CE(p(q), q) = \begin{cases} -\log(p(q)) & \text{if } q = 1 \\ -\log(1 - p(q)) & \text{otherwise} \end{cases} \quad (3.21)$$

and

$$FL(p(q)) = -(1 - p(q))^\gamma \log(p(q)), \quad (3.22)$$

where $p(q) \in [0, 1]$ is the probability of the class with label q , $(1 - p(q))^\gamma$ is the modulating factor, and $\gamma \geq 0$ is the focusing or weighting parameter. While $p(q)$ is defined as (Lin et al., 2018):

$$p(q) = \begin{cases} p(q) & \text{if } q = 1 \\ 1 - p(q) & \text{otherwise.} \end{cases}$$

RetinaNet is composed of a backbone, the Feature Pyramid Network (FPN) (Lin et al., 2017), and two subnetworks. The backbone and FPN are used to extract features at different scales and represent them as anchors, contributing each to a multiscale feature map. The feature maps are then used as inputs for the classification and box regression subnetworks, where the classification subnetwork classifies the anchors as objects or background, predicting a probability for each of them, and the box regression subnetwork regresses the offset between each anchor box and a close ground-truth box (Lin et al., 2018). Figure 3.21 shows an illustration of RetinaNet architecture.

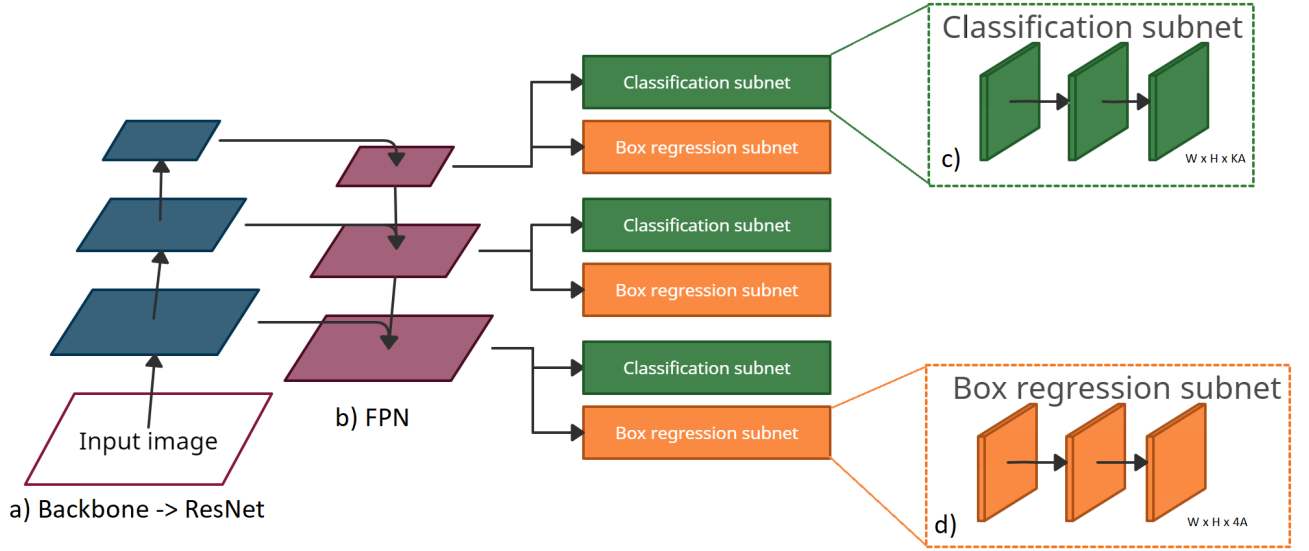


Figure 3.21: RetinaNet illustration, where **a)** and **b)** are the backbone and the FPN, while **c)** and **d)** are the classification and the box regression subnetworks, respectively. The final output of the classification subnetwork has the form $W \times H \times KA$, because of K -length vector of classification targets assigned to each anchor A . Whereas, the box regression subnetwork attached a vector of the coordinates of box regression targets for each anchor, having an output form of $W \times H \times 4A$ (Lin et al., 2018).

- The Fully Convolutional One-Stage Object Detection (FCOS) by Tian et al. (2019) is a one-stage, anchor free object detector that solves object detection using per-pixel prediction. Anchor-based detectors need to tune an important number of hyperparameters related to anchors for achieving good performance, in anchor-free detectors, the number of hyperparameters is considerably less. However, some anchor-free detectors suffer from low recall or need complicated post-processing techniques to improve their results. FCOS introduces a novel *center-ness* branch, which has the effect of suppressing low-quality detected bounding boxes, allowing the detector to outperform its anchor-based counterparts.

FCOS is built on top of FPN, which delivers five feature maps, containing predictions of five feature levels in order to detect objects of different sizes. The locations on the feature map are fed to a three branch detection head composed of the classification head, the center-ness, and the regression. The classification head predicts the per-pixel class probability, while the center-ness branch is in charge of providing the deviation between the location and center of the object that this location falls into. The box regression head outputs a 4D vector (l, t, r, b) for each location, as shown in the left image in Figure 3.22. The coordinates (l, t, r, b) , represent the distant in pixels from the location to the left side (l) of the bounding box, the distant in pixels from the location to the top (t) of the bounding box, the distant in pixels from the location to the right side (r) of the bounding box, and the distant in pixels from the location to the bottom (b) of the bounding box. In the cases where there are overlapping bounding boxes (see Figure 3.22), FCOS addresses

the ambiguity by limiting the range of bounding box regression for each level on the FPN (Tian et al., 2019). An illustration of this architecture is shown in Figure 3.23.

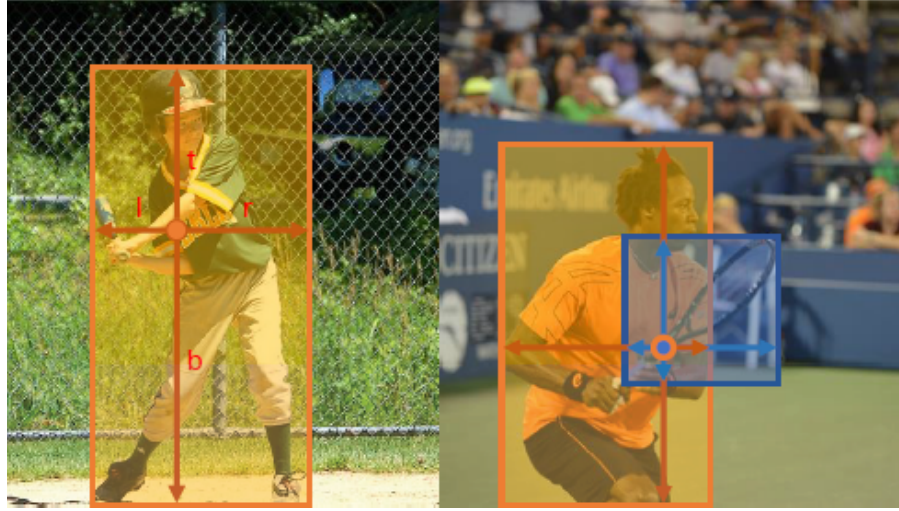


Figure 3.22: Left: Prediction of the 4D vector (l, t, r, b) which encodes the location of a bounding box. Right: an ambiguous sample annotation (Tian et al., 2019).

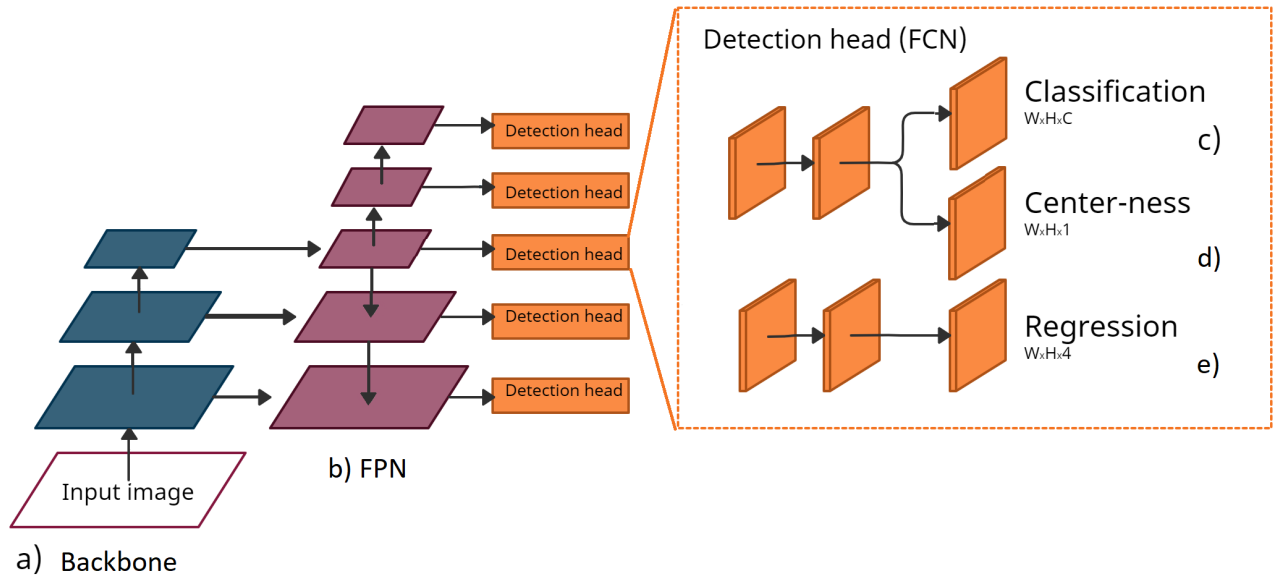


Figure 3.23: FCOS architecture illustration. The backbone **a)** extracts features maps at three levels, while the FPN **b)** predicts features at five levels. The classification head **c)**, with an output of $W \times H \times C$, for an input of width W and height H with C classes. The center-ness head **d)** outputs the normalize distance from the center, whereas the regression head **e)** output is the 4D vector (l, t, r, b) (Tian et al., 2019).

- YOLOv₃ by Redmon and Farhadi (2018), where YOLO stands by You Only Look Once, is a one-stage, very fast, and simple architecture that solves the detection task as a regression problem. The feature extractor of YOLOv₃, called Darknet-53, is inspired by ResNet (He et al., 2016) and FPN (Lin et al., 2017). Several convolutional layers are added to Darknet-53 for extracting features at three different scales, having different

resolution, in order to allow the detection of objects of different sizes (large, medium, and small). The predictions are done by using a grid cell on top of each feature map, and by setting a number of anchors to be predicted for every of these grid cells. The anchors are calculated by computing k-means clustering, and are later given the probability of belonging to a certain class. The final output consists of the predicted bounding box, the objectness score (probability of a cell of being the center of an object), and the label class encoded in a 3-D tensor for each cell in the grid. Bounding box prediction is performed by adjusting the dimensions of the anchor boxes. The objectness score is obtained using logistic regression, while the class predictions are found using independent logistic classifiers, since they allow the classification of multiple labels (Redmon & Farhadi, 2018). In Figure 3.24, a YOLOv₃ illustration can be appreciated.

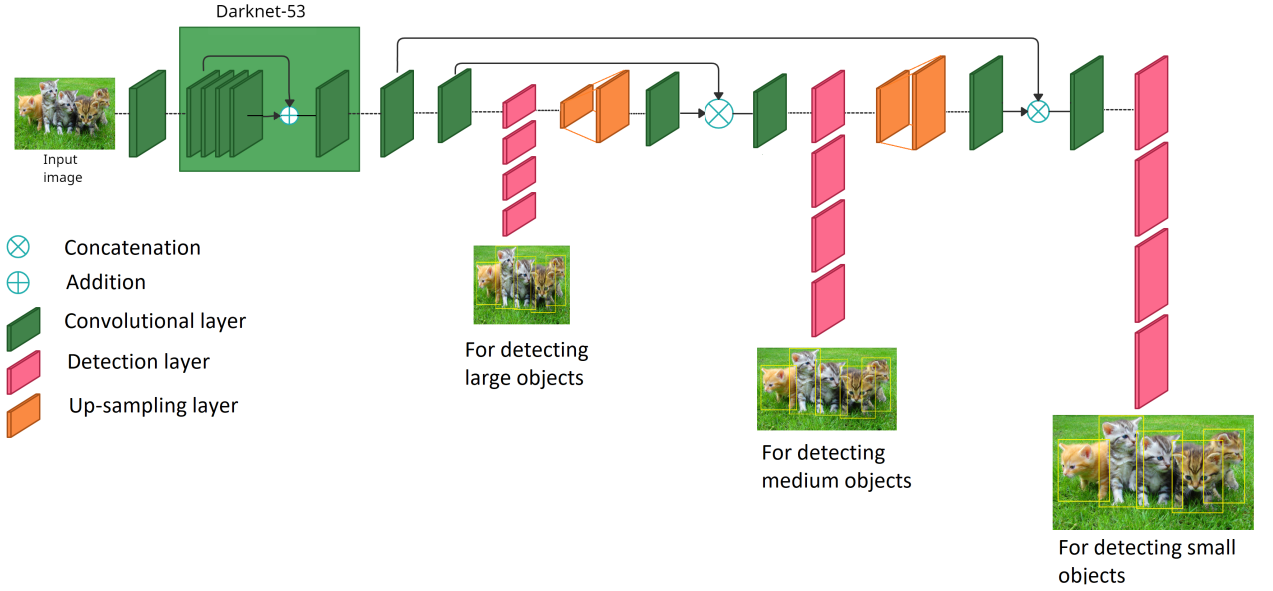


Figure 3.24: YOLOv₃ architecture illustration (Redmon & Farhadi, 2018).

3.3.2 Detection evaluation metrics

The detection evaluation metrics used for characterizing the performance of an object detector are the Average Precision (AP) and Average Recall (AR), which can be computed using various IoU values. IoU is a metric used to measure the overlapping between two regions, in this case bounding boxes, and is defined as:

$$IoU(Area_{DB}, Area_{GTB}) = \frac{Area_{DB} \cap Area_{GTB}}{Area_{DB} \cup Area_{GTB}}, \quad (3.23)$$

where $Area_{DB}$ is the area of detected boxes and $Area_{GTB}$ is the area of ground truth boxes. A graphical representation of IoU can be appreciated in Figure 3.25.

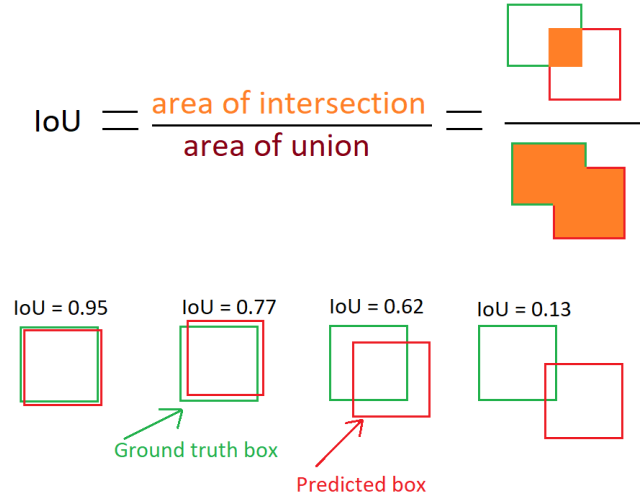


Figure 3.25: IoU quantifies the area of overlap between a predicted box and a ground truth box. Different scenarios are shown in this illustration, where predicted boxes have IoU values between 0.95 and 0.13.

The performance of the models on the testing set was evaluated by measuring the *Accuracy*², *Precision*, and *Recall*. These metrics are calculated with the threshold of *IoU* set to 0.5 and 0.75 using:

$$\text{Accuracy} = \frac{TP}{TP + FP + FN}, \quad (3.24)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ and} \quad (3.25)$$

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (3.26)$$

Where *TP* stands for True Positive and represents the number of teeth correctly detected, *FP* stands for False Positive and represents the number of anatomical structures mistakenly detected as teeth, and the *FN* is the number of teeth that are not detected or are detected with $\text{IoU} \leq 0.5$ or $\text{IoU} \leq 0.75$, which are the False Negatives. In order to compute these metrics, each test image was observed and the number of *TP*, *FP*, and *FN* were written down. Figure 3.26 shows examples of TP, FP, and FN in image with one object.

²True negatives are not taken into consideration for object detection task, since it refers to the parts of the images that a model did not predict as an object.

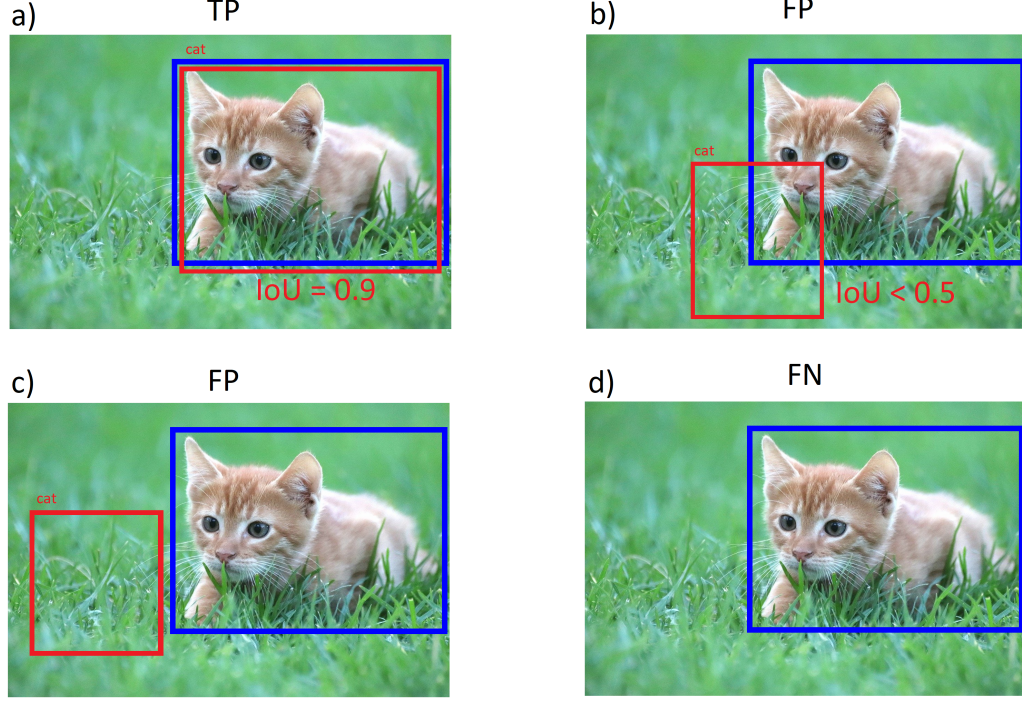


Figure 3.26: **a)** A predicted bounding box with the class label *cat* (red box) overlaps the ground truth box (blue box) with an IoU of 0.9, this is considered a TP. **b)** A predicted bounding box overlaps the ground truth box with an $IoU < 0.5$, which is considered a FP. **c)** a predicted bounding box with the class label *cat* does not overlap the ground truth box, this is a FP. **d)** there are not prediction for the object, this is a FN example.

The AP and AR metrics are calculated taking into consideration the precision and recall obtained by the detector for every class in the dataset:

$$AP = \frac{1}{classes} \sum_q^{classes} P_q, \quad (3.27)$$

$$AR = \frac{1}{classes} \sum_q^{classes} R_q, \quad (3.28)$$

where *classes* is the number of classes in the dataset, P_q is the precision and R_q the recall obtained for the class label q . Since in this work we only have one class label, *tooth*, we have $AP = P_{tooth}$ and $AR = R_{tooth}$. The final value of AP and AR depend on the IoU value selected for classifying the predictions as TP or FP. For $AP_{IoU=0.5}$ and $AR_{IoU=0.5}$, all predictions with an $IoU \geq 0.5$ will be considered as TP, while for $AP_{IoU=0.75}$ and $AR_{IoU=0.5}$, only the predictions with an $IoU \geq 0.75$ will be considered as TP.

3.4 Experimental setup

In order to exclude extra toothless regions and minimize the background in the images, the first step before applying the IET and training the models was to trim the images. The nasal septum, the inferior mandibular border, and the posterior mandibular ramus were used as reference points to draw a unique bounding box in 100 images and train a model using Faster R-CNN. The resulted model was used to detect the area between these points, shown in Figure 3.27 a), and automatically trimmed the remaining images in the dataset. In Figure 3.27 b), the trimmed image resulted from the described process can be appreciated.

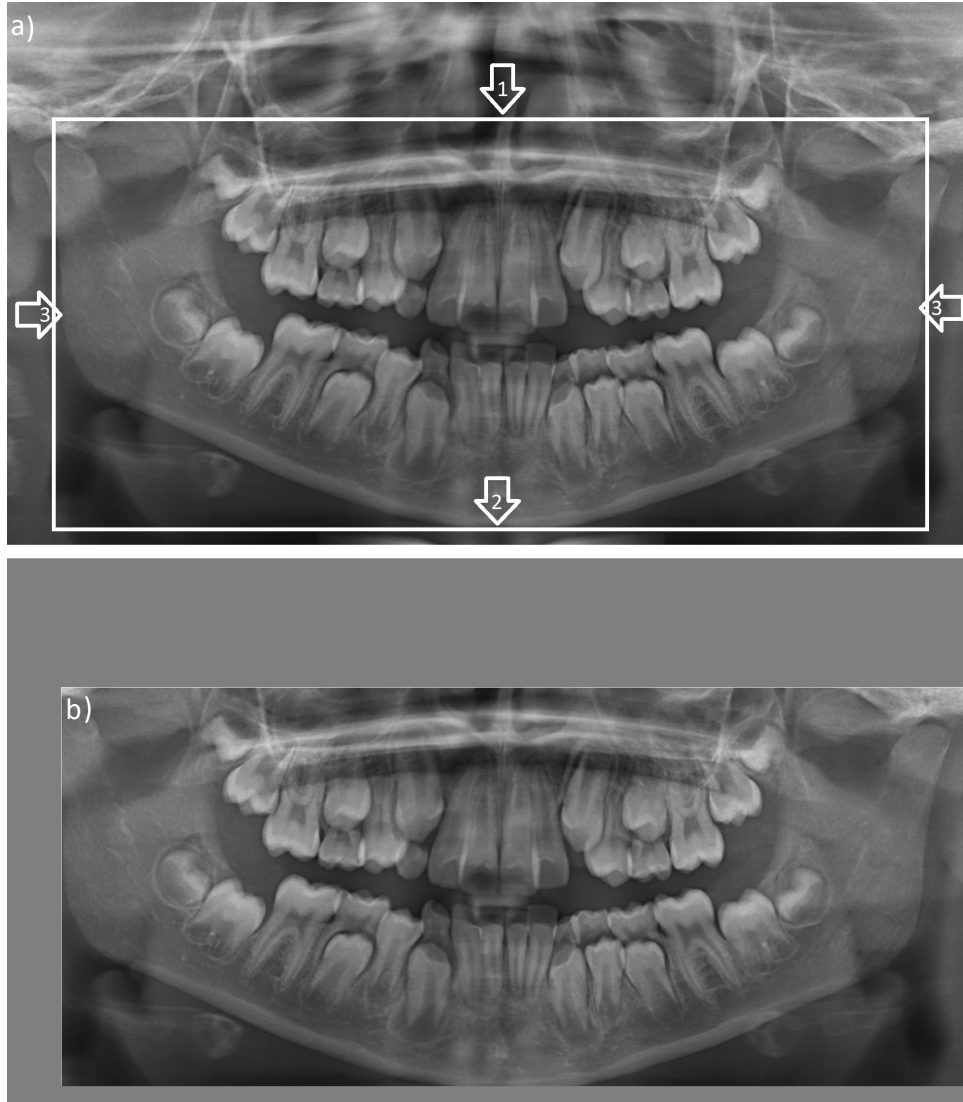


Figure 3.27: The image in **a)** has the original size of 1991×1127 pixels, as the rest of the images in the dataset of Silva et al., 2018. The reference points are shown with the help of the white arrows, where **1** is the nasal septum, **2** is the inferior mandibular border, and **3** are the posterior mandibular ramus. In **b)**, with a dimension of 1886×829 , the trimmed sections of **a)** are highlighted in gray. In this image we can see that the detection of the reference points was successful except for the left posterior mandibular ramus. Because of this, the right part of the image is wider than expected.

The experiments were performed using an open-source object detection toolbox based on PyTorch named MMDetection (K. Chen et al., 2019), which was chosen by its implementations of state-of-the-art segmentation and object detection algorithms. MMDetection provides a set of pre-trained models from different object detection and segmentation methods with training time, box AP³, backbone, learning schedule, etc., as some of the characteristics that differentiate the pre-trained models to select from. These models are available in its Model Zoo, which is a platform where companies and research groups organize and release their code and pre-trained weights.

MS-COCO pre-trained weights were selected to fine tune the models with the dataset of 300 images. The MS-COCO (Lin et al., 2014) is a big dataset of natural images customized for object detection and segmentation purposes. Among the pre-trained object detection baselines available in the MMDetection Model Zoo, one model with its corresponding weight was chosen, taking into consideration the models that achieved the highest $AP_{IoU=0.5:0.05:0.95}$. The metric $mAP = AP_{IoU=0.5:0.05:0.95}$ is one of the detection evaluation metrics used by MS-COCO (Lin et al., 2014) and is computed averaging over 10 IoU threshold, from 0.5 to 0.95 with steps of 0.05, across all categories/classes. In Table 3.3, the architecture, the pre-trained CNN used for generating feature maps (backbone), and the $AP_{IoU=0.5:0.05:0.95}$ of the selected pre-trained models are found.

Table 3.3: Chosen pre-trained models per architecture.

Architecture	Backbone	$AP_{IoU=0.5:0.05:0.95}$
Faster R-CNN	ResNext-101	42.1
Cascade R-CNN	ResNext-101	44.7
RetinaNet	ResNext-101	41.0
FCOS	ResNext-101	42.6
YOLOv ₃	DarkNet-53	33.7

The hyperparameters tuning was performed using the images without any IET apply to them. After several attempts, the training parameters batch size, the learning rate, weight decay, and warmup iterations, were adjusted to achieve a high $AP_{IoU=0.5:0.05:0.95}$ on the validation set.

The final value of the parameters for each method can be found in Table 3.4. While in Table 3.5 are the range of values that were tested for each hyperparameter during tuning.

Table 3.4: Hyperparameters for each method.

Architecture	Batch size	Learning rate	Weight decay	Warmup iters
Faster R-CNN	2	0.01	0.0001	48
Cascade R-CNN	2	0.009	0.001	24
RetinaNet	2	0.008	0.0001	4
FCOS	2	0.02	-	48
YOLOv ₃	8	0.001	0.001	-

³ $AP_{IoU=0.5:0.05:0.95}$ for the COCO dataset.

Table 3.5: Range of values tested during tuning.

Batch size	Learning rate	Weight decay	Warmup iters
2-8	0.001 - 0.1	0.0001 and 0.001	0-92

The IET described in section 3.2.1 were applied to the 300 images after trimming the images and before fine-tuning a total of eleven models using each chosen architecture. Once the hyperparameters tuning was done and the parameters were set, the first model trained for each architecture was the baseline model, which was trained and tested using the images without any IET apply to them. After that, the ten other models were trained with an IET applied to the panoramic images as the pre-processing step. Training these models for 100 epochs took less than an hour for YOLOv₃ models and more than 4 hours for Faster R-CNN, Cascade R-CNN, RetinaNet, and FCOS models. The optimizer used across all architectures was the Stochastic Gradient Descent (SGD), while horizontal image flipping with a ratio of 0.5 was used as the only form of augmentation during training. Panoramic images can be horizontally flipped since we were interested in detecting the teeth without classified them, which makes it irrelevant that the teeth on the right can be on the left side and vice versa. On the other hand, using vertical flipping would change the maxillary teeth for the mandibular teeth in the image, which would give an unreal scenery, since panoramic images cannot be taken with the maxillary teeth down.

The inference on the testing set was performed by selecting the best weight for each method, considering the highest $AP_{IoU=0.5:0.05:0.95}$ on the validation set after training for at most 100 epochs. The number of true positives, false positives, and false negatives resulted from the inference of each model was used to calculate the Accuracy, Precision, and Recall considering two thresholds of $IoU \geq 0.5$ and $IoU \geq 0.75$. These values, with the IQM measured per image in the test set, were used to compute the cross-correlation. The steps followed in this work can be found on Figure 3.28, while the results from these experiments are presented in chapter 4

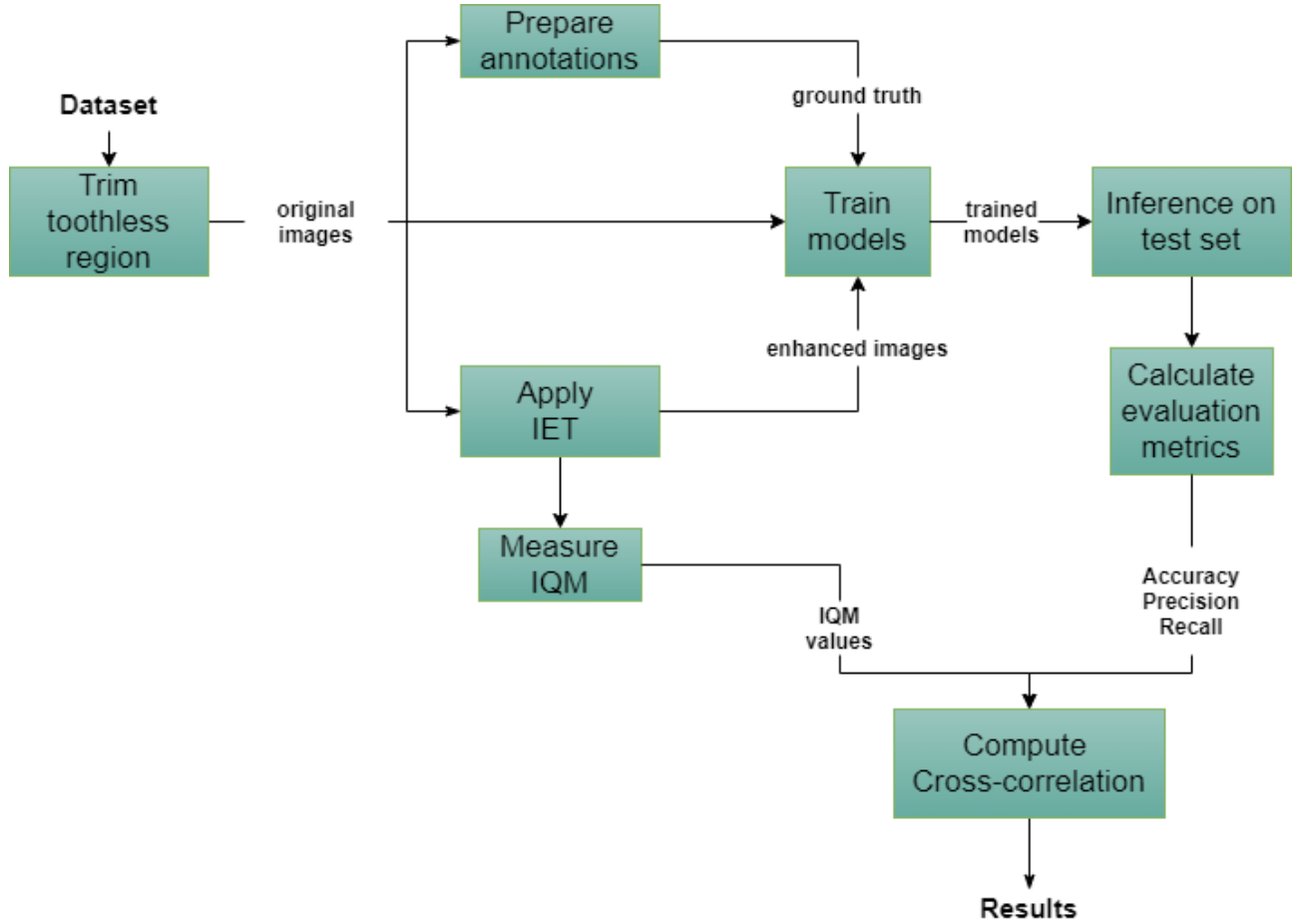


Figure 3.28: Diagram of steps, followed to obtain the results of this work.

The hyperparameter tuning, training of the models, and inference on the test set were all performed in Google Colaboratory (Colab) (Bisong, 2019). Google Colab is a product from Google Research that allows writing and executing Python code in the browser with free access to GPUs, being commonly used for machine learning. The serverless Jupiter Notebook environment provided by Google Colab was run in Google Chrome web browser in an Acer Aspire A513-53 laptop, with a processor Intel Core i5-7200U CPU 2.5 GHz and a RAM of 4 GB.

CHAPTER 4

RESULTS AND DISCUSSION

This section presents the results of the experiments carried out to find the correlation between the IQM measured and the detection results of the five architectures. The results are displayed in the following order:

- In section 4.1, the IQM measured to quantify the performance of each IET applied to the images as a pre-processing step.
- Section 4.2 presents the detection performance of the models measured by computing the *Accuracy*, *Precision*, and *Recall* with the *IoU* threshold set to 0.5 and 0.75.
- The detection evaluation metrics are also used to compare the performance of the models with each other.
- With the IQM values measured on the testing set, and the detection results of the eleven models, the cross-correlation is computed and shown in section 4.3 using heat maps, one for each architecture.

This procedure is repeated in a local dataset of panoramic images, the results of this extra experiment are presented and discussed in section 4.4.

4.1 IQM results

The IQM measured on the pre-processed test set, along with their standard deviation, are displayed in Table 4.1. For the original images, only the entropy and global contrast can be calculated, since the rest of the metrics need two images to be computed. According to the values in the table, the IET that most changes introduced to the images are BBHE, CLAHE, Edge Enhancement, and MTHT, while QHELC changes are the mildest. This can be also appreciated in the outcome images of these IET, a comparison with an original image is shown in Figures 4.1 and 4.2, where each image is present with its IQM values. The image, belonging to category number 3, is first compared, in Figure 4.1, with the Anisotropic Diffusion, Bilateral Filtering, Edge Enhancement, Maximum, and Minimum filters. In Figure 4.2, the same image is compared to the contrast enhancement techniques, where it can be seen that BBHE and CLAHE introduce the most changes to the contrast of the image.

Table 4.1: Average of IQM measured on the testing group.

IET	AMBE	Entropy	SSIM	PSNR	PSNR-HVS	EPI	Contrast	CIR
Original dataset	-	6.998 $\pm$ 0.197	-	-	-	-	33.12 $\pm$ 6.23	-
Anisotropic diffusion	9.4 ⁻⁷ $\pm$ 8.6 ⁻⁷	6.98 $\pm$ 0.11	0.963 $\pm$ 0.007	41.02 $\pm$ 1.4	82.55 $\pm$ 1.45	0.423 $\pm$ 0.078	32.7 $\pm$ 3.45	0.74 $\pm$ 0.06
BBHE	5.7 ⁻² $\pm$ 1.5 ⁻²	6.83 $\pm$ 0.1	0.71 $\pm$ 0.029	14.81 $\pm$ 0.56	57.92 $\pm$ 0.98	0.423 $\pm$ 0.042	74.15 $\pm$ 0.24	39.2 $\pm$ 41.08
Bilateral Filtering	2.2 ⁻⁶ $\pm$ 1.5 ⁻⁶	6.99 $\pm$ 0.11	0.97 $\pm$ 0.006	43.0 $\pm$ 0.99	85.06 $\pm$ 0.99	0.448 $\pm$ 0.09	32.86 $\pm$ 3.5	0.68 $\pm$ 0.04
CLAHE	5.9 ⁻² $\pm$ 1.1 ⁻²	7.65 $\pm$ 0.07	0.866 $\pm$ 0.008	19.04 $\pm$ 0.62	61.9 $\pm$ 0.69	0.951 $\pm$ 0.03	50.48 $\pm$ 2.77	1.7 $\pm$ 1.63
Edge Enhancement	1.4 ⁻⁴ $\pm$ 1.8 ⁻⁴	7.10 $\pm$ 0.1	0.682 $\pm$ 0.035	29.22 $\pm$ 1.07	72.77 $\pm$ 1.28	1.293 $\pm$ 0.071	35.22 $\pm$ 3.66	58.19 $\pm$ 14.64
GRMMCE	5.2 ⁻⁴ $\pm$ 2.0 ⁻⁴	7.06 $\pm$ 0.1	0.959 $\pm$ 0.007	35.63 $\pm$ 1.0	77.58 $\pm$ 0.94	0.672 $\pm$ 0.022	34.58 $\pm$ 3.58	4.36 $\pm$ 4.82
Maximum	2.0 ⁻² $\pm$ 2.4 ⁻³	7.05 $\pm$ 0.11	0.921 $\pm$ 0.012	30.29 $\pm$ 1.7	72.47 $\pm$ 1.94	0.22 $\pm$ 0.021	34.59 $\pm$ 3.88	1.41 $\pm$ 0.64
Minimum	2.0 ⁻² $\pm$ 2.4 ⁻³	6.93 $\pm$ 0.11	0.923 $\pm$ 0.012	30.39 $\pm$ 1.69	72.57 $\pm$ 1.91	0.137 $\pm$ 0.024	31.46 $\pm$ 3.23	1.41 $\pm$ 0.64
MTHT	6.0 ⁻⁴ $\pm$ 5.2 ⁻⁴	7.16 $\pm$ 0.09	0.767 $\pm$ 0.028	27.79 $\pm$ 1.12	69.45 $\pm$ 1.08	0.309 $\pm$ 0.015	37.12 $\pm$ 3.78	35.2 $\pm$ 12.11
QHELC	1.1 ⁻³ $\pm$ 9.4 ⁻⁴	6.98 $\pm$ 0.11	0.993 $\pm$ 0.001	36.91 $\pm$ 0.98	79.48 $\pm$ 1.11	0.913 $\pm$ 0.015	36.33 $\pm$ 3.23	0.19 $\pm$ 0.17

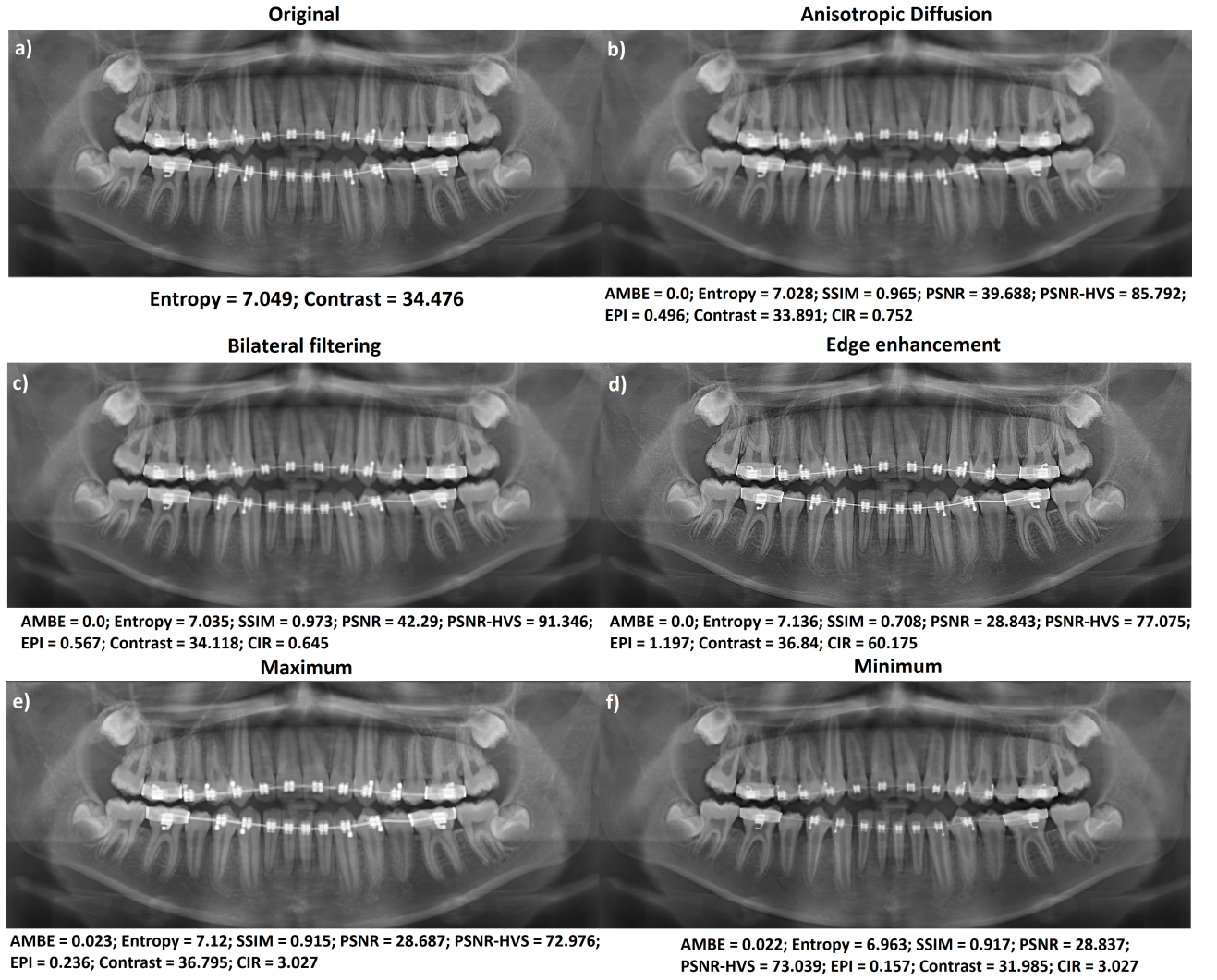


Figure 4.1: a) Original image; b) Anisotropic Diffusion output image; c) Bilateral Filtering output image; d) Edge Enhancement output image; e) Maximum output image; f) Minimum output image. The changes introduced by the b) Anisotropic Diffusion and the c) Bilateral Filtering are almost visually imperceptible, while the d) Edge Enhancement shows an image with highlighted edges. In e), teeth have a thickened appearance, while in f), the effect is the opposite.

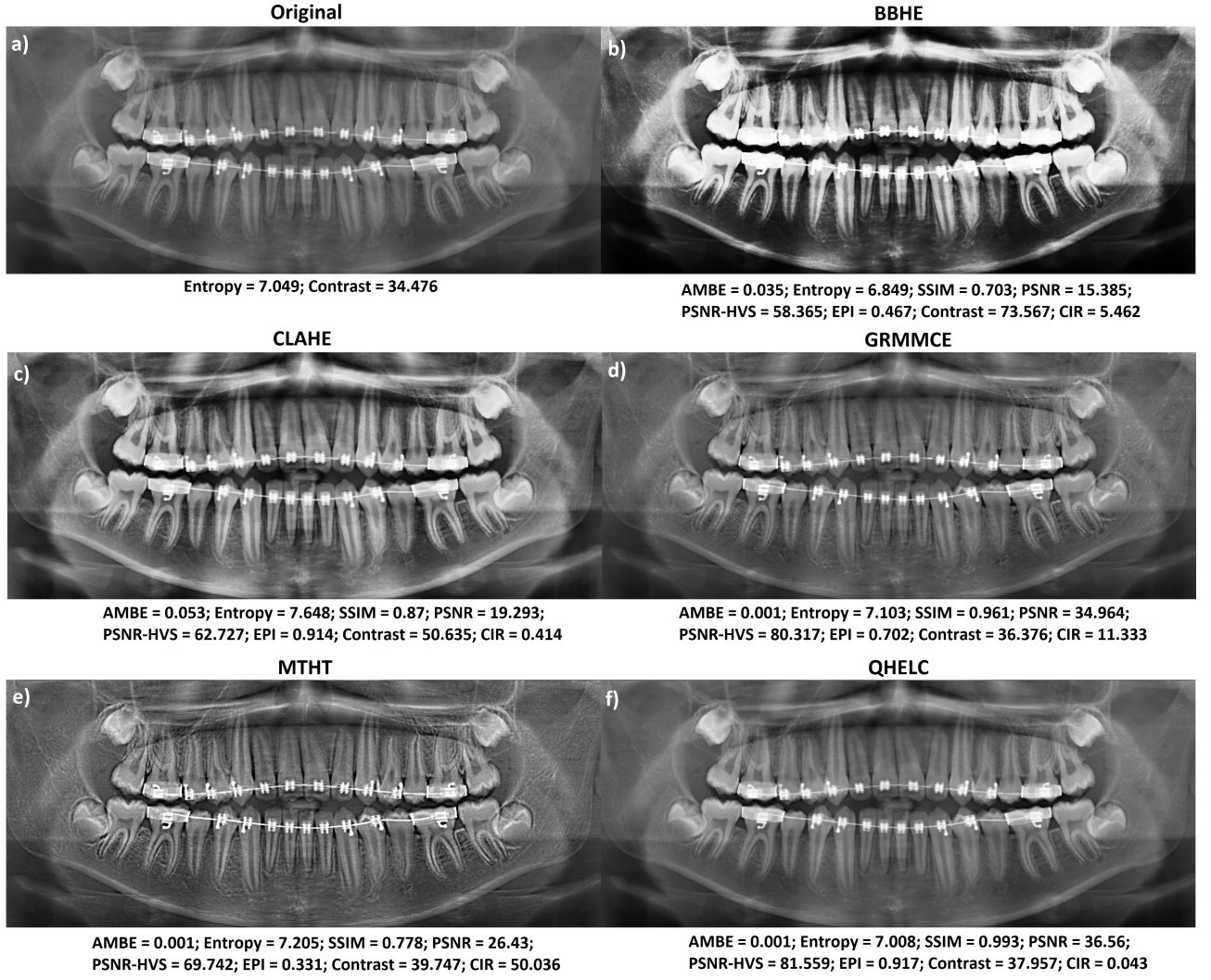


Figure 4.2: a) Original image; b) BBHE output image; c) CLAHE output image; d) GRMMCE output image; e) MTHT output image; f) QHELC output image. The BBHE image in b), presents the greatest change in its contrast, followed by the CLAHE image in c), both also having a visually perceptible change in their brightness. It can also be perceived a change in the e) MTHT image, which presents a high value of CIR in comparison to the rest of the images in this figure. Lastly, the changes introduced by QHELC in f), are practically imperceptible to the eye, which is translated to its IQM values, shown below the image.

4.1.1 Discussion

The Bilateral Filtering and the Anisotropic Diffusion are the two IET that less distortion introduced to the panoramic images, achieving the highest PSNR and PSNR-HVS. These smoothing edge-preservation filters, present higher values of EPI than the contrast enhancement technique MTHT and the Maximum and Minimum filters, with the Minimum filter, as was expected, having the lowest value of EPI in Table 4.1. The majority of the techniques do not introduce a great change in the brightness of the images, with CLAHE and BBHE, the last being a contrast enhancement technique that focuses on brightness preservation, having the two highest values of AMBE, while the Anisotropic Diffusion preserves the most of this feature.

For the contrast enhancement techniques, BBHE and CLAHE are the ones that increase the most the global contrast of the images. BBHE also, unlike CLAHE, introduces an important change in the local contrast, having a high value of CIR, followed by MTHT and exceeded only by the Edge Enhancement technique. The biggest changes in the contrast, either globally or locally, introduce distortion and noise to the images, represented by low values of SSIM, PSNR, and PSNR-HVS. In fact, the three techniques with the highest values of CIR, are also the ones with the lowest values of SSIM. While BBHE and CLAHE present the lowest values of PSNR and PSNR-HVS. Despite these values, CLAHE manages to preserve the edges of the objects, having the closest value of EPI to one, while also raising the most the entropy of the images. The other contrast enhancement technique with an EPI value close to one is QHELC, this technique also has the highest SSIM value in the table, which indicates a low distortion introduced to the images.

4.2 Teeth detection

The detection performance of the models was tested, considering as true detection those teeth that present bounding boxes with $IoU \geq 0.5$ and $IoU \geq 0.75$. Among the evaluation metrics considered for evaluating the performance of the models, Accuracy was taken as a reference, as it is a metric that is computed with both FP and FN (see equation 3.24), to define which model performed better in relation to the others.

Table 4.2 shows the Faster R-CNN models results. When using a threshold of 0.5, seven out of ten models have a better Accuracy than the baseline, being the model trained and tested with the images enhanced with MTHT the one with the best results, while the Bilateral Filtering model has the second-best performance among the models. The best Precision is achieved by the Edge Enhancement model, although its Recall is the lowest in the table. The highest Recall across all the models and methods is achieved by CLAHE model, however, its Precision is the one of the lowest, affecting also the Accuracy of the model. With the 0.75 threshold, the Precision of the model remains as one of the lowest in the table, the reason for this might be the high value of global contrast of the images. The relationship between the Precision and the global contrast will be further discussed in section 4.3. On the other hand, when the threshold was change to 0.75, the Accuracy and Precision of the Bilateral Filtering model increase, making the model even better. The number of false positives for this model goes from 5 to 0, while the number of false negatives increases by one, due to the IoU of these detections being less than 0.75. The MTHT model also improves, as can be seen in Table 4.2, and both models keep having a better Accuracy than the baseline model. On the other hand, the BBHE model achieves the lowest Accuracy and Precision in the table, while the Edge Enhancement model present the lowest Recall among the models of this architecture.

Table 4.2: Teeth detection results of Faster R-CNN models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.

Model	Faster R-CNN					
	Threshold = 0.5			Threshold = 0.75		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	98.92	99.49	99.43	99.31	99.94	99.37
Anisotropic Diffusion	99.09	99.66	99.43	99.31	99.94	99.37
BBHE	98.81	99.32	99.49	99.20	99.71	99.49
Bilateral filtering	99.20	99.71	99.49	99.43	100.00	99.43
CLAHE	99.15	99.49	99.66	99.26	99.83	99.43
Edge Enhancement	99.03	99.83	99.20	99.03	99.94	99.09
GRMMCE	99.03	99.54	99.49	99.26	99.89	99.37
Maximum filter	99.09	99.71	99.37	99.26	99.89	99.37
Minimum filter	98.86	99.60	99.26	98.92	99.71	99.20
MTHT	99.26	99.71	99.54	99.37	99.83	99.54
QHELC	99.09	99.60	99.49	99.26	99.94	99.31

Figure 4.3 shows the detection performed by the baseline model in comparison with the detection of the Bilateral Filtering **b)**, CLAHE **c)**, and MTHT **d)** models on a challenging image, with 42 teeth, belonging to category 6. In this image, the baseline model missed two teeth, both the Bilateral Filtering and MTHT missed one tooth, marked all with the blue boxes on images **a)**, **b)** and **d)**, while in **c)** we can see that all teeth are detected.

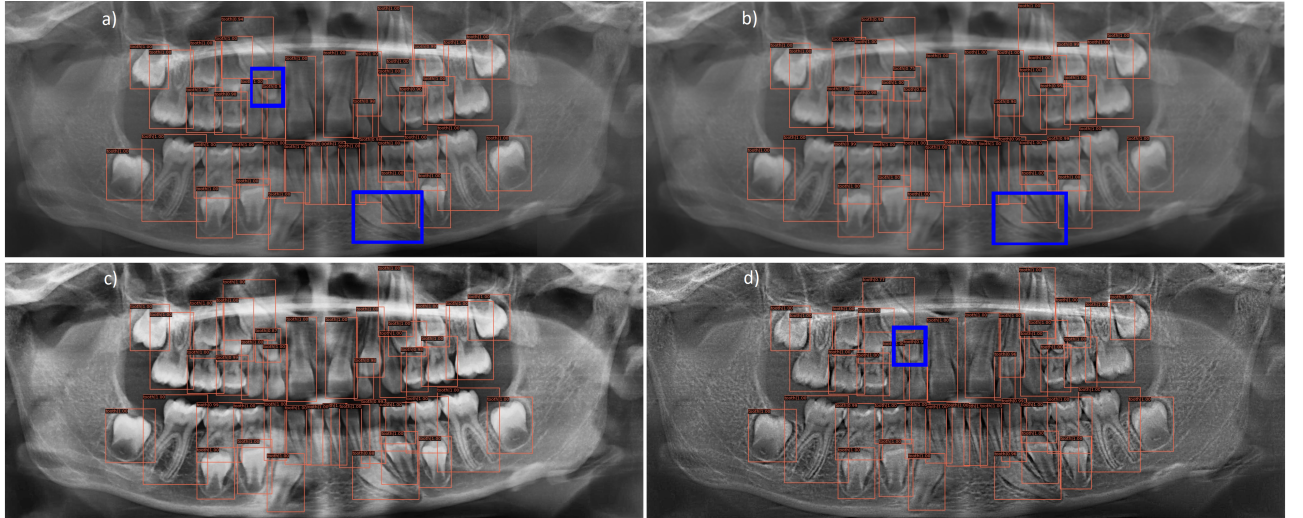


Figure 4.3: Inference of Faster R-CNN: **a)** baseline model, **b)** Bilateral Filtering model, **b)** CLAHE model, **c)** MTHT model. The undetected teeth are marked with the blue boxes, being two in **a)**, and one tooth in **b)** and **d)**.

The performance of the models trained using RetinaNet are shown in Table 4.3. In this table, we can see that GRMMCE, the Bilateral Filtering and QHELC models have a better Accuracy than the baseline model. The Bilateral Filtering, and QHELC models achieve the best performance, with the GRMMCE model having the second-best Accuracy. While the best Recall among the models is achieved by CLAHE model and the Precision of the baseline model is not improved by any model. In this table, the model with the lowest Accuracy and Precision is BBHE, while the Edge Enhancement model presents the lowest Recall among the models.

BBHE is the contrast enhancement technique that causes the greatest increase in the global contrast of the images, leading to a high number of false positives and the subsequent low Precision of this model. By changing the threshold to 0.75, the performance of the models drop due to the IoU scores of the bounding boxes being less than 0.75. This decrease is expected since, according to the experiments carried out by the authors on COCO dataset, $AP_{IoU=0.5}$ is at least 15% greater than $AP_{IoU=0.75}$ (Lin et al., 2018). With this threshold, CLAHE, Maximum, and QHELC models perform better than the baseline model, and the Minimum filter present the worse performance.

Table 4.3: Teeth detection results of RetinaNet models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.

model	RetinaNet					
	Threshold = 0.5			Threshold = 0.75		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	98.81	99.43	99.37	92.00	100.00	92.00
Anisotropic Diffusion	98.69	99.37	99.31	91.19	100.00	91.19
BBHE	97.37	97.92	99.43	91.72	99.81	91.88
Bilateral filtering	98.92	99.43	99.49	91.65	100.00	91.65
CLAHE	98.70	99.09	99.60	92.69	99.94	92.74
Edge Enhancement	98.24	99.25	98.97	90.68	100.00	90.68
GRMMCE	98.86	99.37	99.49	91.83	99.94	91.88
Maximum filter	98.64	99.09	99.54	92.17	100.00	92.17
Minimum filter	98.52	99.43	99.09	89.59	100.00	89.59
MTHT	98.47	99.09	99.37	91.60	99.88	91.71
QHELC	98.92	99.37	99.54	92.22	100.00	92.22

In Figure 4.4, the inference on an image, with 32 teeth belonging to category 1, is performed to compare the **a)** baseline model with **b)** BBHE, **c)** GRMMCE, and **d)** Edge Enhancement models. The image in **a)**, on the top left, shows 30 detected teeth and two undetected teeth marked with blue boxes, while in image **b)**, in addition to miss the detection of one tooth, the BBHE model have two false positives marked with red boxes in the image. In **c)**, the GRMMCE model misses the detection of only one tooth, also marked with a blue box, while we can see in **d)** that three teeth are not detected by the Edge Enhancement model, one less than in the original image in **a)**. In fact, none of the models were able to detect all the teeth on this image, not even the Bilateral Filtering or QHELC.

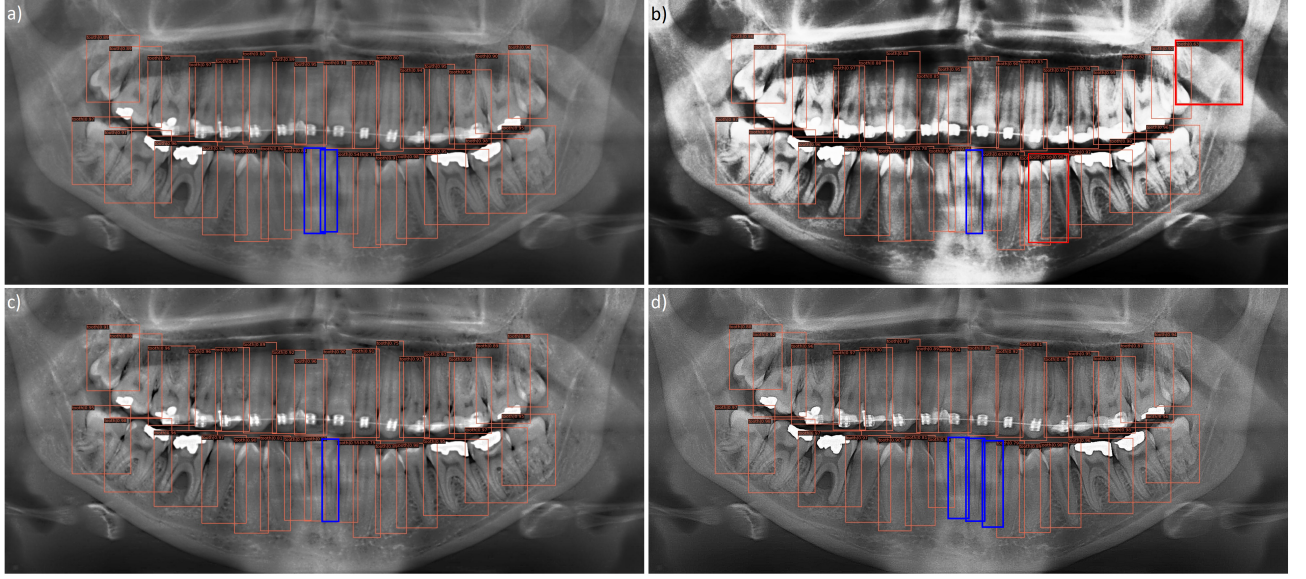


Figure 4.4: Comparison of teeth detection performed by RetinaNet **a)** baseline model, **b)** BBHE, **c)** GRMMCE, and **d)** Edge Enhancement on an image belonging to category 1. The red box in image **b)** are false positives, while the blue boxes indicates false negatives.

The performance of the models trained on Cascade R-CNN is shown in Table 4.4. Here, the QHELC model is the only model that has slightly better values of Accuracy and Precision than the baseline model, with an Accuracy of 99.49%, a Precision of 99.94%, and a Recall of 99.54%. While the model with the worse performance is again the BBHE model, having a Precision 5.37% less than the Precision of the QHELC model. While with 0.75, the GRMMCE and QHELC models have both the same classification results, making the baseline model the second-best of this table.

Table 4.4: Teeth detection results of Cascade R-CNN models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.

Model	Cascade R-CNN					
	Threshold = 0.5			Threshold = 0.75		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	99.43	99.94	99.49	99.31	100.00	99.31
Anisotropic Diffusion	99.37	99.89	99.49	99.03	99.94	99.09
BBHE	95.35	95.76	99.54	97.64	98.25	99.37
Bilateral filtering	99.37	99.94	99.43	99.09	100.00	99.09
CLAHE	98.03	98.47	99.54	98.86	99.43	99.43
Edge Enhancement	99.26	99.94	99.31	98.97	100.00	98.97
GRMMCE	99.37	99.83	99.54	99.37	100.00	99.37
Maximum filter	99.32	99.77	99.54	99.09	99.94	99.14
Minimum filter	99.20	100.00	99.20	98.80	100.00	98.80
MTHT	99.31	99.94	99.37	99.20	100.00	99.20
QHELC	99.49	99.94	99.54	99.37	100.00	99.37

In Figure 4.5, the difference between the detection performed by the baseline model **a)** and QHELC **b)** is compared when using a threshold of 0.5 in an image of category 9. In **a)**, we can see neighboring teeth being detected together in one box by the baseline model, while in **b)**, the QHELC model is able to independently detect the two teeth. In Figure 4.6, the baseline model **a)** is compared with the GRMMCE **b)** and QHELC **c)** models when the inference is done using the 0.75 threshold. Figure 4.6 shows an image belonging to category 5, with implants, restorations and, a crown. The baseline model does not detect the crown in **a)**, marked with a blue box, while the GRMMCE and QHELC models are able to detect the crown and the rest of teeth, restorations and implants.

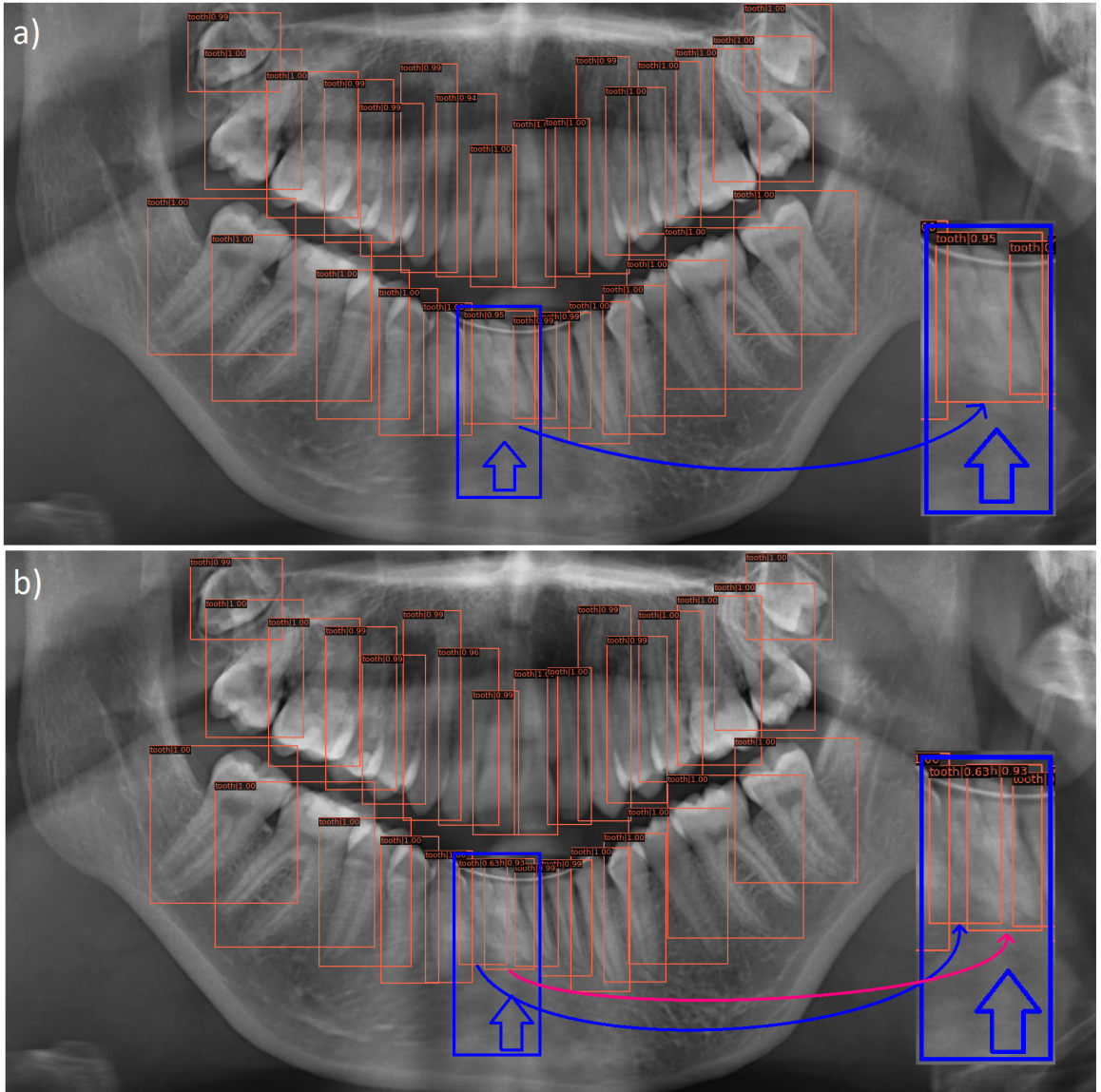


Figure 4.5: The above image shows the detection performed by Cascade R-CNN **a)** baseline and **b)** QHELC models in a challenging image from category 9 with 30 teeth when using 0.5 as a threshold. In **a)**, neighboring teeth were detected together in one box, while in **b)**, QHELC is able to detect these two teeth independently.

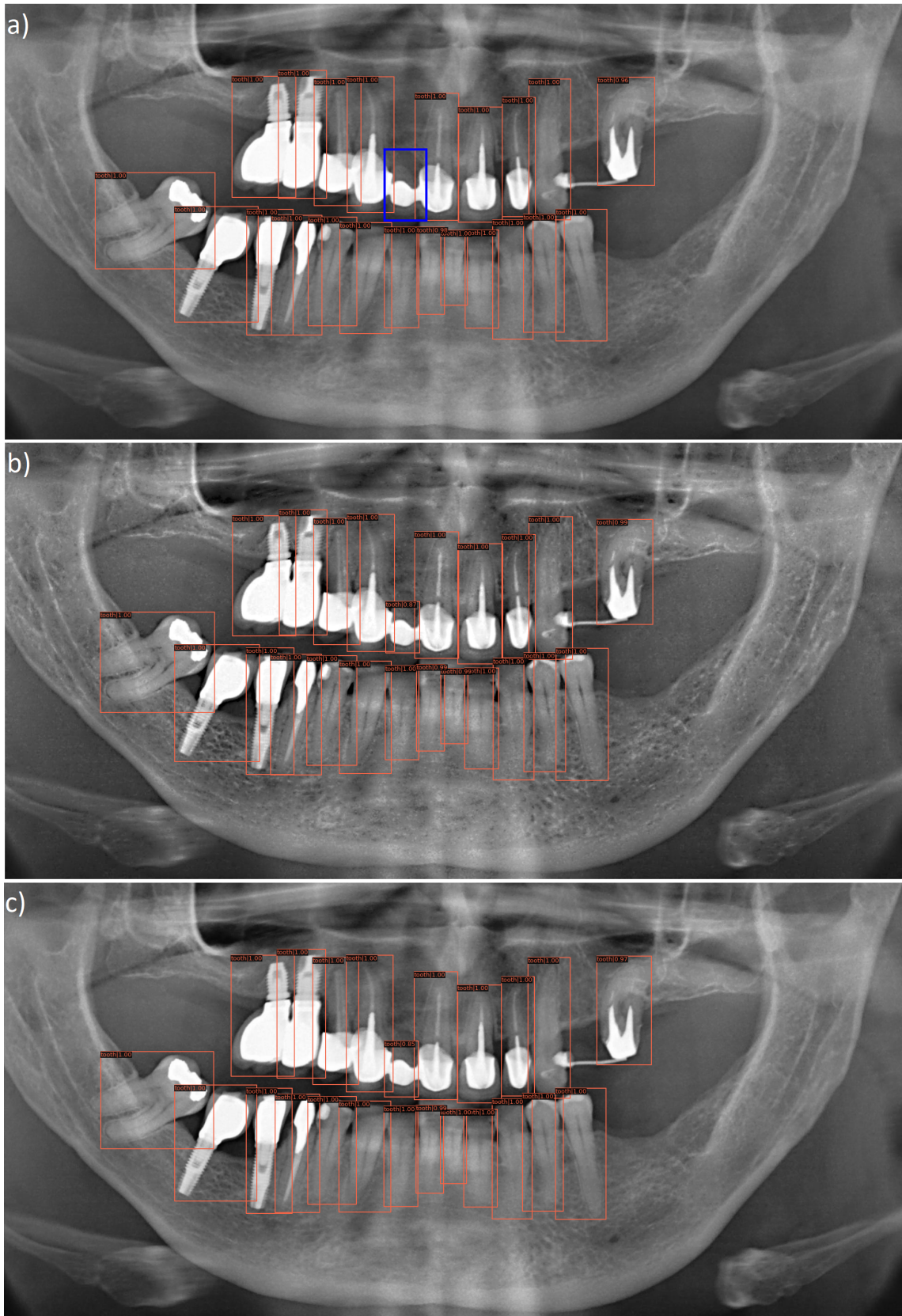


Figure 4.6: In these images, the inference performed by the baseline model **a)** of Cascade R-CNN, is compared with the GRMMCE **b)** and QHELC **c)** models while using an 0.75 of threshold. In **a)**, the model is unable to detect a crown, marked with a blue box, while the GRMMCE and QHELC models are able to detect it.

The results of FCOS' models are presented in Table 4.5. Here, the QHELC model achieves the best results, and it is followed by the baseline model, while the BBHE model presents the lowest Accuracy among the models. As can be seen in this table, except for the Maximum filter, all models present a perfect Precision ($FP = 0$), which is because the models do not mistakenly detect the background or the region between teeth as teeth. This value of Precision is also achieved by some RetinaNet and Cascade R-CNN models (see Tables 4.3 and 4.4), specifically with the 0.75 threshold, and one model of Faster R-CNN and YOLOv₃ (see Tables 4.2 and 4.6). However, the best accuracy is less than that of most models trained with Faster R-CNN, RetinaNet, and Cascade R-CNN. When 0.75 is used, the Accuracy is even lower than the values achieved by the aforementioned architectures, because the teeth are detected with values of IoU less than 0.75. This behavior can be expected since, according to the results presented by the authors, $AP_{IoU=0.5}$ is at least 15% greater than $AP_{IoU=0.75}$ when using FCOS on COCO dataset (Tian et al., 2019). The image in Figure 4.7 is chosen to show the inference performed by the baseline model in **a**) and the QHELC model in **b**). This is a challenging image with 42 teeth and while the baseline model is unable to localize 4 of them, marked with the blue boxes, the QHELC model failed to detect only two teeth.

Table 4.5: Teeth detection results of FCOS models. The results of the models with the best and second-best Accuracy are highlighted in red and blue, respectively.

Model	FCOS					
	Threshold = 0.5			Threshold = 0.75		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	97.83	100.00	97.83	56.72	100.00	56.72
Anisotropic Diffusion	97.43	100.00	97.43	56.72	100.00	56.72
BBHE	97.37	100.00	97.37	57.46	100.00	57.46
Bilateral filtering	97.54	100.00	97.54	58.20	100.00	58.20
CLAHE	97.71	100.00	97.71	58.43	100.00	58.43
Edge Enhancement	97.48	100.00	97.48	55.40	100.00	55.40
GRMMCE	97.54	100.00	97.54	55.35	100.00	55.35
Maximum filter	97.60	99.94	97.66	55.23	100.00	55.23
Minimum filter	97.48	100.00	97.48	56.09	100.00	56.09
MTHT	97.71	100.00	97.71	54.77	100.00	54.77
QHELC	98.11	100.00	98.11	57.06	100.00	57.06

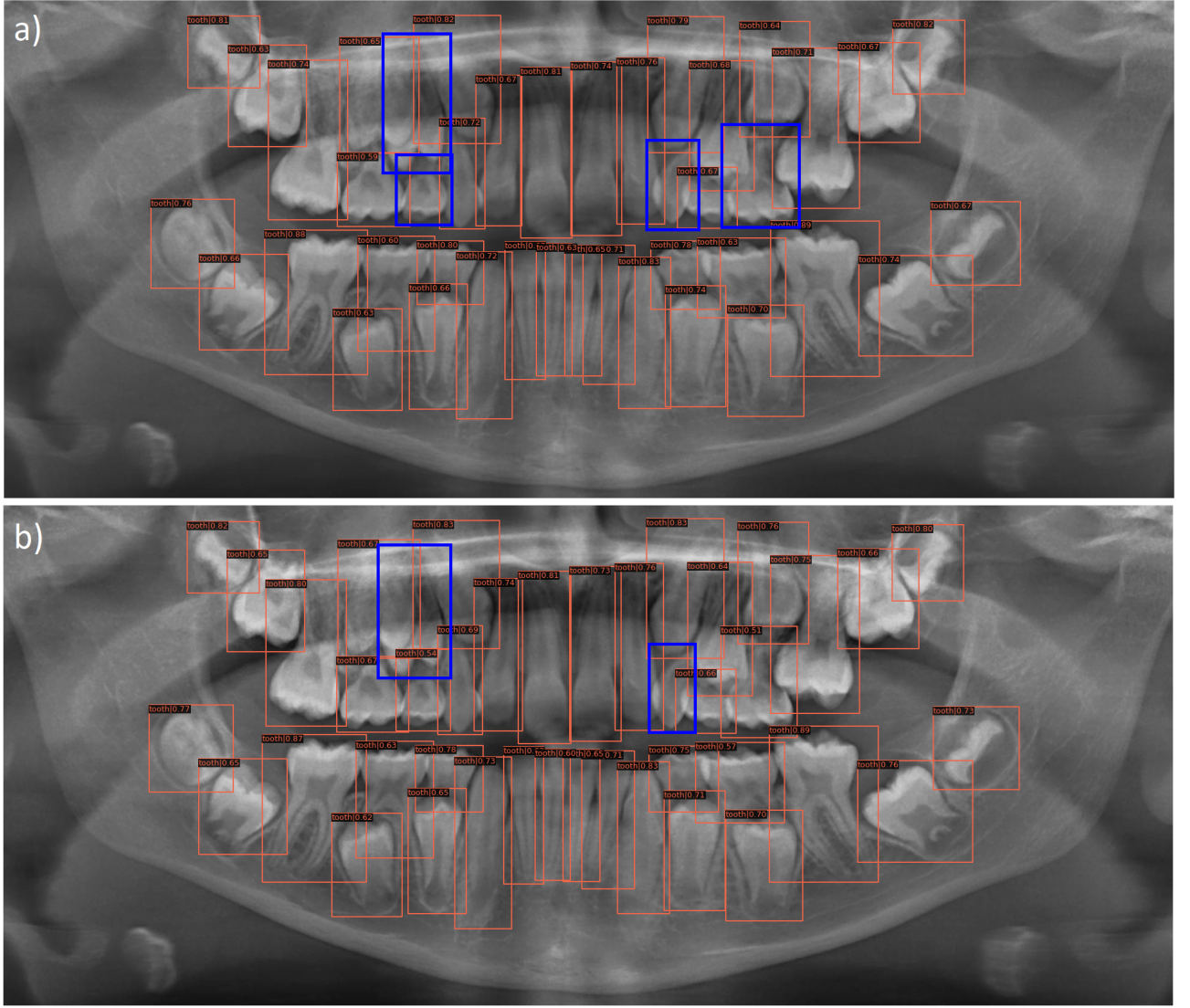


Figure 4.7: Inference performed by FCOS a) baseline model and b) QHELC model in an image with 42 teeth. Both models failed to detect all 42 teeth, blue boxes marked the four and two undetected teeth in a) and b), respectively.

The last chosen method to run the experiments was YOLOv₃, the results of the models trained using this architecture can be found in Table 4.6. As can be seen from the table, there was no IET applied to the images that resulted in an improvement of the results achieved by the baseline model. In this table, GRMMCE has the second-best results after the baseline model, and the Edge Enhancement model presents the lowest Accuracy. Also, unlike the results obtained in the other architectures, the BBHE model presents the third-best Accuracy and Recall of this table. When switching to the 0.75 threshold, the performance of the models worsen by more than 10%. This decrease in the performance is expected since, according to the results presented by the authors, $AP_{IoU=0.5}$ is 23.5% greater than $AP_{IoU=0.75}$ when using the method on COCO dataset (Redmon & Farhadi, 2018). Figure 4.8 shows some undetected teeth, marked with a blue box, when performing inference with the baseline model of YOLOv₃ in images of categories 1, 2, 3, 4, 5, and 9.

Table 4.6: Teeth detection results of YOLOv₃ models.

Model	YOLOv ₃					
	Threshold = 0.5			Threshold = 0.75		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	96.29	99.94	96.34	87.14	100.00	87.14
Anisotropic Diffusion	94.18	99.82	94.34	81.89	99.93	81.93
BBHE	95.84	99.76	96.05	86.69	99.87	86.79
Bilateral filtering	95.03	99.88	95.14	83.10	99.86	83.19
CLAHE	95.66	99.82	95.83	86.92	99.87	87.02
Edge Enhancement	92.76	99.69	93.02	80.53	99.86	80.62
GRMMCE	95.95	99.82	96.11	86.17	99.93	86.22
Maximum filter	95.26	99.82	95.43	85.83	99.93	85.88
Minimum filter	95.49	99.94	95.54	85.83	99.93	85.88
MTHT	95.04	99.64	95.37	82.57	99.93	82.62
QHELC	95.83	99.94	95.88	86.57	99.93	86.62

4.2.1 Discussion

The results in Tables 4.2-4.6 show that using an IET as a pre-processing step before training in different object detection architectures will not necessarily result in an improvement in the detection. There are IET that do improve the detection results, among them, the two models that stand out are Bilateral Filtering and QHELC, being the ones that are more frequently better than the baseline model. The improvements of these two IET do not introduce noise or distortion to the original images, preserving their brightness, local contrast, and entropy. On the other hand, the Edge Enhancement and BBHE, which are the IET that introduce a high level of noise to the images, are the models with the worse classification results among the models and architectures.

When comparing the results among the architectures, we notice that if we would like to choose only one method to perform teeth detection in panoramic images, the best choice would be Cascade R-CNN when using a threshold of 0.5 for inference, while if the selected threshold was 0.75, Faster R-CNN would be the method to be chosen. This is because Faster R-CNN models improve their performance when using the 0.75 threshold, the reason behind this is the overlapping IoU scores of the detected teeth. Most of the true positives are detected with a score higher than 0.75, while the false positives have an IoU score of less than 0.75, which results in an improvement, especially in the Precision of the models. Except for Faster R-CNN, the methods' performance worsen when the threshold used was 0.75. For Cascade R-CNN, the Accuracy of the models decays by less than 1%, with two models, BBHE and CLAHE, improving their performance because of the fewer false positives that these models present. Also, the Bilateral Filtering results, the best model in Table 4.2 with a threshold of 0.75, are only 0.06% better than the two best among Cascade R-CNN models in Table 4.4 with a threshold of 0.75, making Cascade R-CNN the second-best option for teeth detection on panoramic images. The 0.75 IoU threshold requires the models to have more precise detections. Thus, the decrease in the performance of RetinaNet, FCOS, and YOLOv₃ is to be expected when using this threshold, because of the IoU scores of the detections being less than 0.75, specially in FCOS and YOLOv₃.

The inference results in challenging images and their impact in the detection results is further discussed in section A.2 of the Appendix.

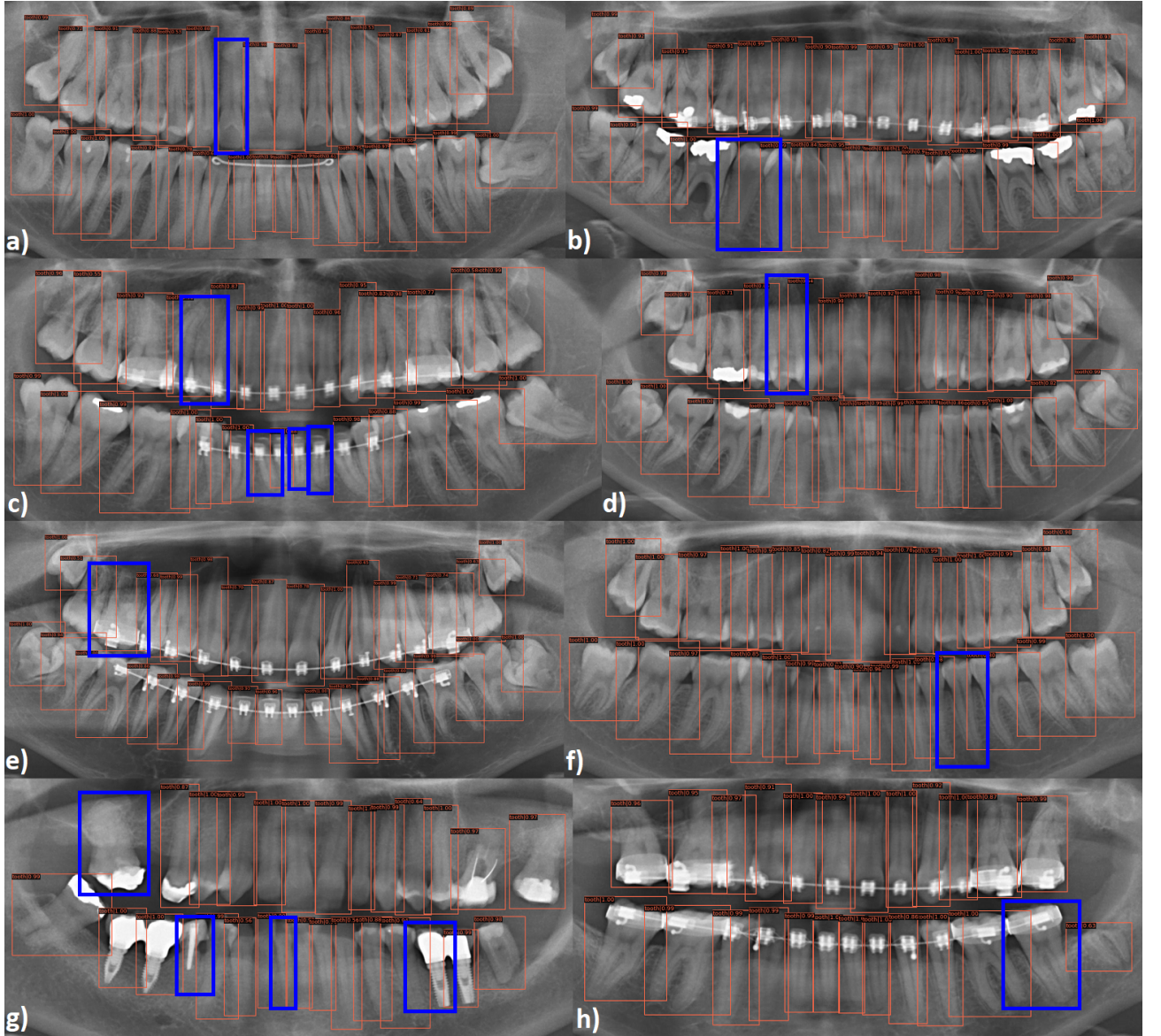


Figure 4.8: Some undetected teeth, marked with a blue box, when using the baseline model of YOLOv₃ with an 0.5 of threshold. The images **a)**, **b)**, and **c)** belong to category 1. In **d)**, an image of category 2 is shown, while **e)** and **d)** present images of categories 3 and 4, respectively. At the bottom, images **g)** and **h)** correspond to images of categories 5 and 9, respectively.

4.3 Cross-correlation

In order to find which features of the image are related to improvements on the detection performance of the teeth, the cross-correlation between these values was computed, and heat maps are used to appreciate these results. The correlation heat maps were computed using the detection results of the models and the IQM measured per image in the testing set.

The correlation heat map of the Faster R-CNN models with the 0.5 threshold is shown in Figure 4.9 **A**, while Figure 4.9 **B** shows the heat map formed with the detection results

corresponding to the 0.75 threshold. The correlation between the detection results: Accuracy, Precision, and Recall with the IQM: AMBE, entropy, SSIM, PSNR, PSNR-HVS, EPI, global contrast and CIR, is highlighted with a green box in the heat map **A** of Figure 4.9. The relationships between the IQM are found in the yellow box, and since they are computed among the IQM values measured per image in the testing set, the correlations values do not change in the next heat maps. In the green area of Figure 4.9 **A**, we can see some very weak negative relationships: the Accuracy with AMBE and CIR, the Precision with AMBE, entropy and global contrast, and Recall with PSNR, PSNR-HVS, and CIR. The heat map in Figure 4.9 **B** shows the same scenario as in **A**. Both heat maps also shown a very weak positive relationship between Recall and entropy, and the Precision with PSNR/PSNR-HVS. The small variations between the detection results in Table 4.2, where the detection results of the best model are less than 1% more than the detection results of the model with the worst performance, could be the reason behind the nearly neutral relationships found in these heat maps.

The relationships found in the section highlighted with the yellow box establish that the SSIM, PSNR, and PSNR-HVS metrics are negative affected by changes in brightness, entropy, global contrast, local contrast, and edge distortion. Which is due to the negative relationships between these three metrics and AMBE, entropy, global contrast, CIR, and EPI, respectively. The metric that most affects the SSIM metric is CIR, while PSNR/PSNR-HVS are more affected by the increment in the global contrast. This is clearly verified by the values present in Table 4.1, as have been already discussed in subsection 4.1.1, where the techniques Edge Enhancement, BBHE, and MTHT present the highest CIR and the lowest SSIM, while BBHE and CLAHE present the highest global contrast and the lowest PSNR/PSNR-HVS.

Regarding the positive relationships among the IQM, there are very strong positive relationships between SSIM, PSNR, and PSNR-HVS, in fact, PSNR and PSNR-HVS have an almost 1 to 1 relationship. While there are also the ones between CIR, EPI and the global contrast, the global contrast and AMBE, EPI and the entropy, and AMBE with the entropy. The EPI having a moderate and weak relationship with the entropy and CIR, respectively, means that enhancing the edges will result in an image with reacher details and a higher local contrast. Finally, incrementing the global contrast of the images will result in the increment of the mean brightness, resulting in a higher value of AMBE, which is verified by the strong relationship between these two IQM metrics.

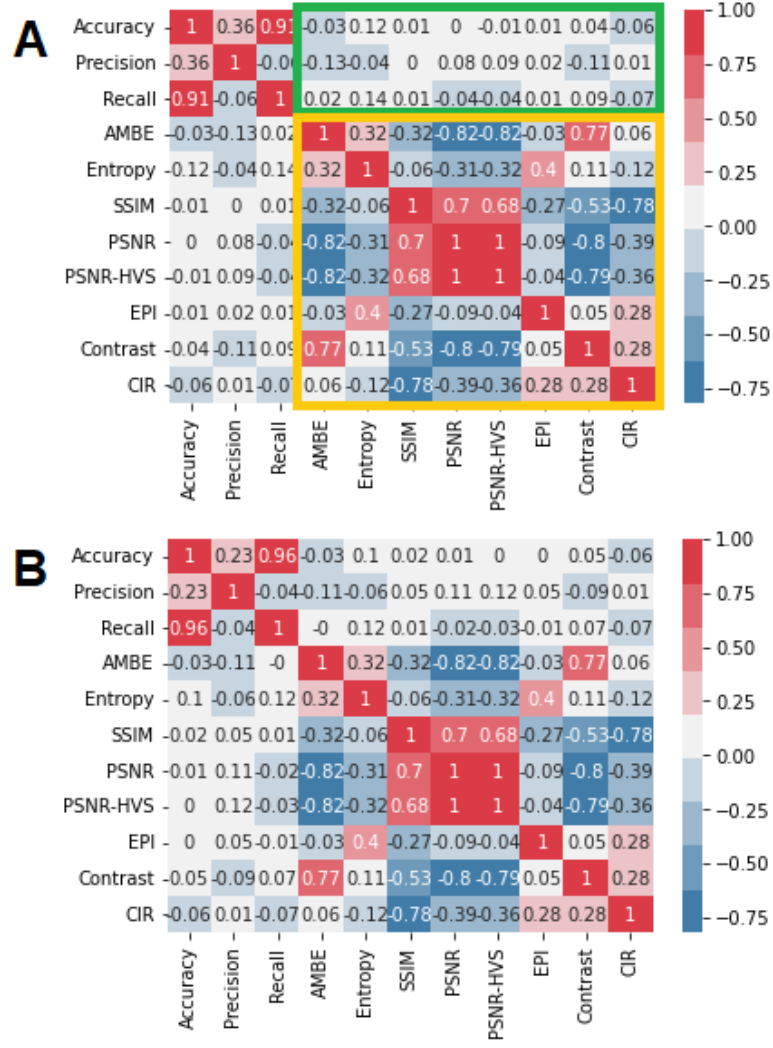


Figure 4.9: Correlation heat maps of Faster R-CNN models, where **A** was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and **B** is the heat map formed with the 0.75 detection results.

The correlation heat maps of RetinaNet models with the threshold of 0.5 and 0.75 are shown in Figure 4.10 **A** and **B**, respectively. In Figure 4.10 **A**, we can see that the Precision holds a weak negative relationship of -0.22 with the global contrast, this negative relationship is again present in the heat map **B** with a value of -0.11 . This means that the Precision will be negatively affected by an increment of the global contrast of the images, which is the case of BBHE model. As can be seen in Table 4.1, BBHE is the enhancement technique that increases the most the global contrast of the images, which results in the BBHE model having the lowest Precision among the models in Table 4.3.

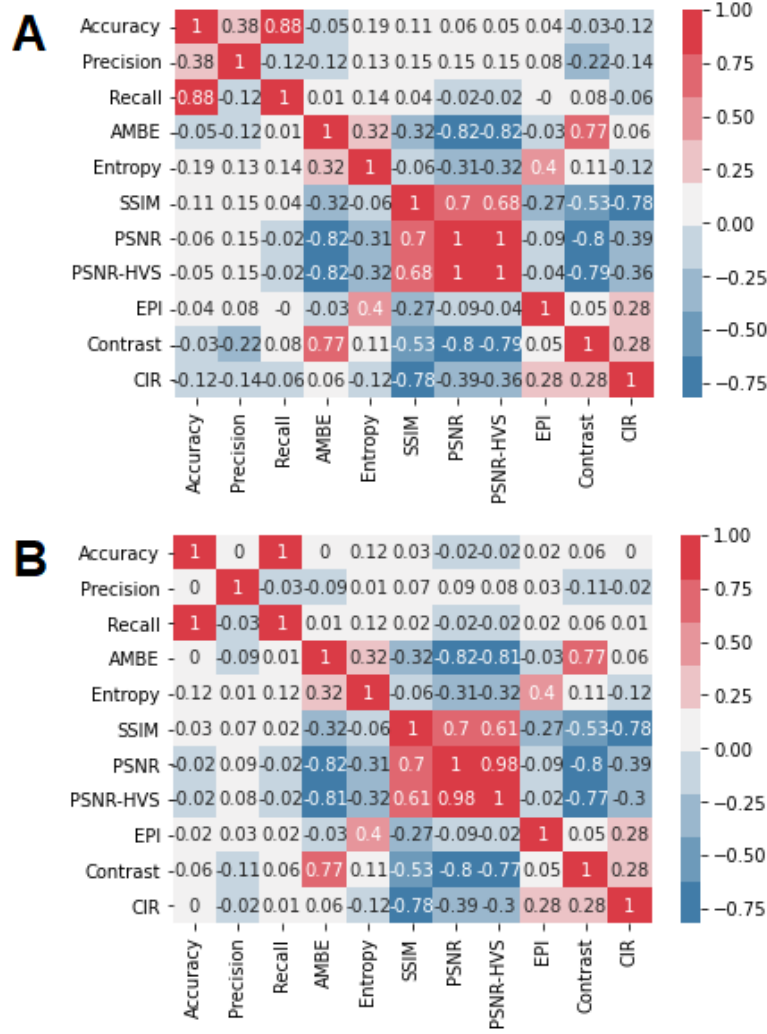


Figure 4.10: Correlation heat map of RetinaNet models, where **A** was formed using the detection results of the model when using the threshold of 0.5 and the IQM measured per image in the testing set, and **B** is the heat map formed with the 0.75 detection results.

In the correlation heat map of Cascade R-CNN models in Figure 4.11 **A**, the Accuracy and the Precision have a negative relation with AMBE and the global contrast, while they have a positive relationship with SSIM, PSNR, and PSNR-HVS. The negative relationship between AMBE and the global contrast with the Accuracy and the Precision means that a change in the brightness and global contrast of the original images would negatively affect the Accuracy and Precision of the teeth detection. Whereas the positive relationship between SSIM, PSNR, and PSNR-HVS with the Accuracy and Precision, is telling us that an IET that does not introduce much noise or distortion to the images, is more likely to achieve higher values of Accuracy and Precision. The images enhance with the Bilateral Filtering present the highest values of PSNR and PSNR-HVS, the model trained with these images achieves the second-best Precision 99.94% and the third-best Accuracy on Table 4.4. The heat map shown in Figure 4.11 **B** also shows that the Precision has a negative relationship with AMBE and the global contrast, and a positive one with SSIM, PSNR, and PSNR-HVS.

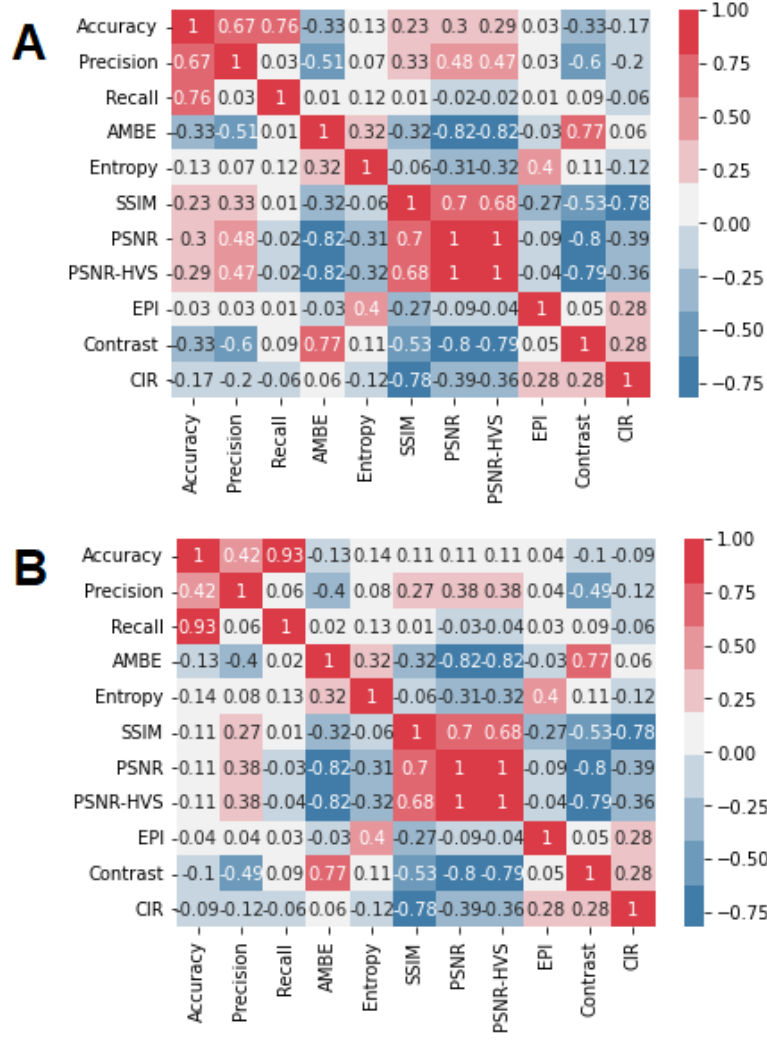


Figure 4.11: Correlation heat map of Cascade R-CNN models, where **A** was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and **B** is the heat map formed with the 0.75 detection results.

Finally, the correlation heat maps of the FCOS and YOLOv₃ models are presented in Figure 4.12 and Figure 4.13. According to these heat maps, the teeth detection results of FCOS and YOLOv₃ have only very weak relationships with the considered IQM. The closeness between the accuracies of the models in Table 4.5 may explain the almost neutrality present in the FCOS heat maps. The heat map in Figure 4.12 **B** does not have the Precision, since the value of this metric is 100% for all models in Table 4.5 when inference is performed using 0.75 as a threshold.

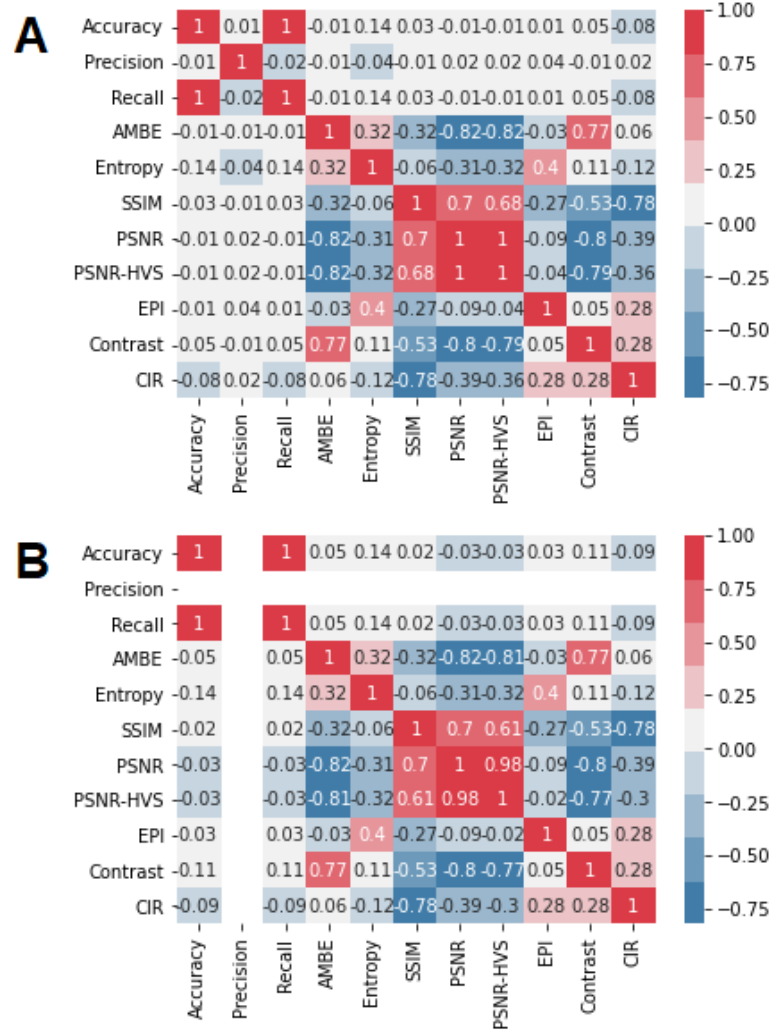


Figure 4.12: Correlation heat map of FCOS models, where **A** was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and **B** is the heat map formed with the 0.75 detection results.

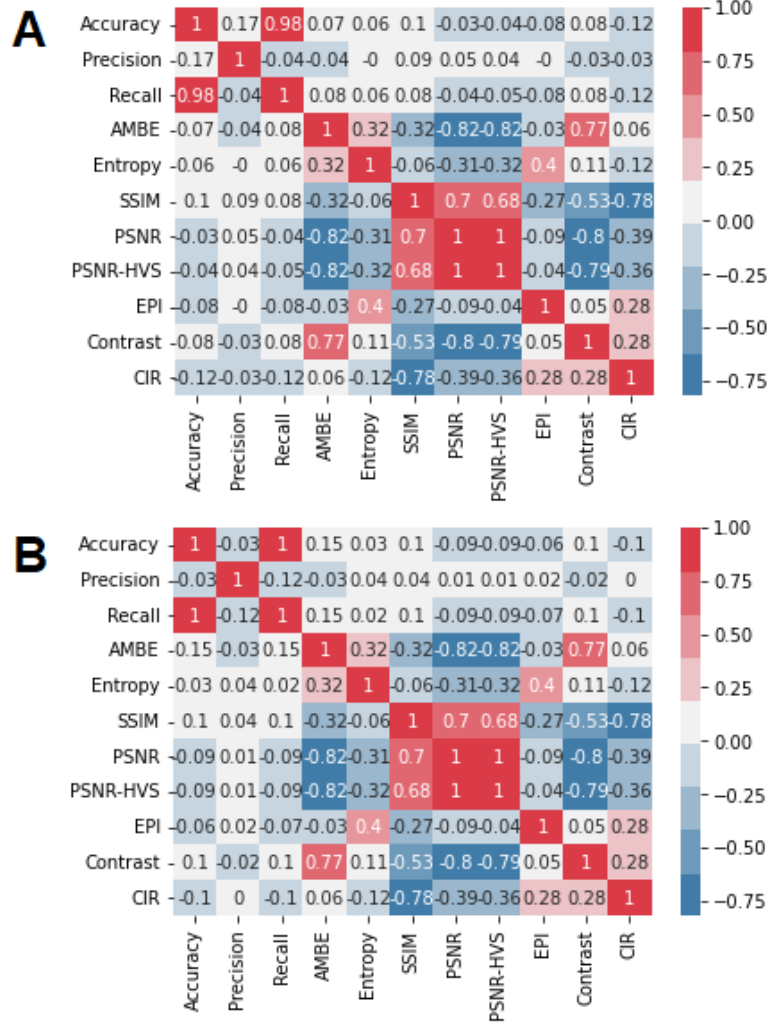


Figure 4.13: Correlation heat map of YOLOv3 models, where **A** was formed using the detection results of the models when using the threshold of 0.5 and the IQM measured per image in the testing set, and **B** is the heat map formed with the 0.75 detection results.

4.3.1 Discussion

Faster R-CNN correlation heat maps show very weak negative relations between the Accuracy with AMBE and CIR, Precision with AMBE, entropy, and global contrast, and the Recall with PSNR, PSNR-HVS, and CIR. According to the models results in Table 4.2, the best and second-best Accuracy and Precision are achieved by the MTHT and Bilateral Filtering models, while CLAHE has the best Recall, followed by MTHT. However, the Precision of CLAHE is affected by the higher number of false positives, which is also the case of the BBHE model. This can be understood by observing that CLAHE and BBHE images present the highest change in the brightness and the global contrast, two of the IQM that negatively affect the Precision. While the Bilateral Filtering achieves one of the best detection results due to the bilateral filtered images presenting almost insignificant changes in the brightness, entropy, and global contrast. Continuing with what was found in these heat maps, Recall having a negative relation, although it is very weak, with PSNR, PSNR-HVS, and CIR is contradictory when observing

the IQM measured in the Bilateral Filtering and MTHT images. According to this relationship, images with high PSNR, PNSR-HVS, and CIR will tend to have low Recall. Although this is not the case with the Bilateral Filtering images, which have the highest PSNR and PSNR-HVS among the IET. While MTHT introduced a significant change to the local contrast of the images, presenting the third-highest value of CIR (see Table 4.1). The very weak negative relationship between Recall and CIR makes sense when looking at the Recall achieved by the Edge Enhancement model and the CIR value resulting from applying this technique to the images. It could be assumed that if the models are trained and tested with a greater number of images, this correlation could be clearer. Although at the moment it can only be conjectured about it.

The heat maps of Faster R-CNN, RetinaNet, Cascade R-CNN, and YOLOv₃ show that the Precision has a negative relationship with AMBE and the global contrast, this means that the Precision will be negatively affected by a variation in the brightness and an increment in the global contrast of the images. The Recall shows a very weak negative relationship with CIR, PSNR, and PSNR-HVS and a very weak positive relationship with the entropy and global contrast in every heat map in Figures 4.9-4.13, which means that the Recall is negatively affected by the increment of the local contrast, but it benefits from lower values of the PSNR and PSNR-HVS and of an increment in the entropy and global contrast of the images. In Table 4.1, we can see that the Edge Enhancement presents the highest CIR, whereas the CLAHE technique presents low values of CIR, PSNR, and PSNR-HVS, the second-highest value of global contrast, and the highest value of entropy. The Recall achieved by the Edge Enhancement model achieves the lowest Recall in four architectures, while the CLAHE model achieves one of the best Recall in four architectures. The Accuracy and Precision of the BBHE model is one of the lowest in four of the architectures, Faster R-CNN, RetinaNet, Cascade R-CNN, and FCOS. This is due to the fact that the BBHE technique introduces the biggest change in the brightness and global contrast of the images. While the Edge Enhancement model is the lowest or one of the lowest in the five architectures because of the Recall achieved by the model. This could be seen as an unexpected result that could change if the Edge Enhancement is applied with a mask that enhances the edges in a softer manner than the mask used in the experiments (see 3.4), although this can only be taken as an assumption that should be tested.

Cascade R-CNN detections results show the strongest relationships with the IQM among the architectures. In Figure 4.11 **A** and **B**, we can see a moderate positive relationship between the Precision and SSIM, PSNR, and PSNR-HVS; in **A**, while the Accuracy has a weak positive relationship with these three IQM. The Anisotropic Diffusion and the Bilateral Filtering are the two IET with the highest PSNR and PSNR-HVS, with the third and second-best values of SSIM, respectively, while QHELC has the highest SSIM value. Among the models' detection results of Faster R-CNN, RetinaNet, Cascade R-CNN, and YOLOv₃, the Precision of the models Anisotropic Diffusion, Bilateral Filtering, and QHELC is in almost every occasion, among the three best values on the Tables 4.2, 4.3, 4.4, and 4.6.

4.4 Inference on local panoramic images

The next experiment was performed by first selecting 20 images belonging to the *Periodontist Institute - Periodontal Diagnostic Center, Asunción, Paraguay*. This experiment was done to find if the correlation results presented above would reproduce in a different set of images. In contrast to the dataset of Silva et al. (2018), these images are sometimes blurry or have poor contrast. Before applying the IET, the images were cut to exclude toothless regions, as with the previous dataset. Then, the inference was performed using the eleven models of each Faster and Cascade R-CNN with a threshold of 0.5. The average IQM values measured on the 20 images are presented in Table 4.7, while Table 4.8 shows the detection results of the models of both architectures.

In the Table 4.7, we can see that the Anisotropic Diffusion and the Bilateral filtering are again the IET that less noise and distortion introduced to the images, with QHELC technique being the one with the best SSIM and EPI, and the third-best PSNR/PSNR-HVS. While the BBHE technique introduces the greatest increase in the global contrast of the images, whereas the Edge Enhancement changes the most the local contrast. The global contrast of these images presents a greater variability than the 60 images of the testing set, which is also present on the final global contrast of the enhanced images, except for the BBHE technique, which presents the lowest standard deviation for this metric.

Table 4.7: Average of IQM measured on the 20 images.

IET	AMBE	Entropy	SSIM	PSNR	PSNR-HVS	EPI	Contrast	CIR
Original dataset	-	6.69 $\pm$ 0.35	-	-	-	-	51.42 $\pm$ 10.84	-
Anisotropic Diffusion	1.3 ⁻⁶ $\pm$ 1.1 ⁻⁶	7.35 $\pm$ 0.33	0.968 $\pm$ 0.016	43.71 $\pm$ 3.05	86.08 $\pm$ 3.57	0.621 $\pm$ 0.053	51.24 $\pm$ 10.80	0.64 $\pm$ 0.10
BBHE	6.1 ⁻² $\pm$ 3.6 ⁻²	7.81 $\pm$ 0.15	0.864 $\pm$ 0.057	18.83 $\pm$ 3.46	61.75 $\pm$ 3.48	1.686 $\pm$ 0.541	73.69 $\pm$ 0.97	4.49 $\pm$ 13.36
Bilateral Filtering	3.1 ⁻⁶ $\pm$ 2.9 ⁻⁶	7.35 $\pm$ 0.33	0.975 $\pm$ 0.012	45.25 $\pm$ 2.92	88.12 $\pm$ 3.36	0.672 $\pm$ 0.045	51.31 $\pm$ 10.82	0.58 $\pm$ 0.09
CLAHE	6.8 ⁻² $\pm$ 3.8 ⁻²	7.61 $\pm$ 0.18	0.897 $\pm$ 0.046	19.86 $\pm$ 2.21	62.96 $\pm$ 1.90	1.583 $\pm$ 0.24	56.76 $\pm$ 8.74	0.70 $\pm$ 0.27
Edge Enhancement	1.2 ⁻⁴ $\pm$ 2.1 ⁻⁴	7.37 $\pm$ 0.30	0.693 $\pm$ 0.075	30.74 $\pm$ 2.51	76.79 $\pm$ 3.15	6.978 $\pm$ 1.607	52.38 $\pm$ 10.96	13.9 $\pm$ 17.70
GRMMCE	4.1 ⁻⁴ $\pm$ 3.4 ⁻⁵	7.38 $\pm$ 0.32	0.945 $\pm$ 0.022	37.40 $\pm$ 2.79	79.50 $\pm$ 3.00	1.533 $\pm$ 0.143	52.00 $\pm$ 10.89	2.14 $\pm$ 2.27
Maximum	1.7 ⁻² $\pm$ 5.3 ⁻³	6.87 $\pm$ 0.32	0.936 $\pm$ 0.030	32.81 $\pm$ 3.07	75.63 $\pm$ 3.58	0.716 $\pm$ 0.041	52.33 $\pm$ 11.13	1.45 $\pm$ 0.23
Minimum	1.7 ⁻² $\pm$ 5.3 ⁻³	6.84 $\pm$ 0.30	0.937 $\pm$ 0.029	32.91 $\pm$ 3.05	75.71 $\pm$ 3.55	0.692 $\pm$ 0.044	50.34 $\pm$ 10.57	1.70 $\pm$ 0.26
MTHT	5.0 ⁻⁴ $\pm$ 4.6 ⁻⁴	7.41 $\pm$ 0.32	0.780 $\pm$ 0.072	30.27 $\pm$ 3.08	71.95 $\pm$ 3.13	3.209 $\pm$ 0.412	53.05 $\pm$ 11.04	5.88 $\pm$ 7.21
QHELC	3.0 ⁻³ $\pm$ 1.5 ⁻³	7.38 $\pm$ 0.31	0.978 $\pm$ 0.005	42.01 $\pm$ 1.72	85.03 $\pm$ 2.10	1.034 $\pm$ 0.026	52.07 $\pm$ 10.37	1.08 $\pm$ 0.19

According to the Faster R-CNN models' results, the GRMMCE model outperformed the baseline model in Accuracy and Precision, whereas the Maximum model has the best Recall among the models. Also, the MTHT model presents the lowest Accuracy, and BBHE and Maximum the lowest Precision. The inference performed by the Maximum and Edge Enhancement models is compared with the inference done by the baseline model in Figure 4.14. The image **a)** in this figure is showing the inference performed by the baseline model, where two teeth go undetected, whereas in **b)** only one tooth is not detected by the Maximum model. On the other hand, the baseline model can detect all teeth in **c)**, unlike the Edge Enhancement model in **d)**. The three undetected teeth are marked with the blue boxes in this figure. For the Cascade R-CNN models, QHELC is the only model with a better Accuracy than the baseline, while CLAHE shows the best Recall among the models. As for the lowest results of the metrics, the Edge Enhancement presents the lowest Accuracy and Recall, while BBHE achieves the lowest

Precision. The image in Figure 4.15 is chosen to compare the performance of the baseline and QHELC models. The image in **a)** shows that the baseline model cannot detect one tooth, while in **b)** we can see that all teeth are detected by QHELC.

Table 4.8: Teeth detection results of Faster and Cascade R-CNN models trained using the 20 images belonging to a local clinic.

Model	Faster R-CNN			Cascade R-CNN		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Baseline	99.00	99.50	99.50	99.00	99.83	99.16
Anisotropic Diffusion	98.84	99.50	99.33	99.00	99.83	99.16
BBHE	98.51	99.17	99.33	97.86	98.51	99.33
Bilateral Filtering	98.67	99.33	99.33	98.84	99.50	99.33
CLAHE	98.84	99.33	99.50	99.00	99.33	99.67
Edge Enhancement	98.33	99.83	98.49	97.83	99.66	98.16
GRMMCE	99.17	99.83	99.33	99.00	99.66	99.33
Maximum	98.84	99.17	99.67	98.84	99.50	99.33
Minimum	99.00	99.66	99.33	99.00	99.83	99.16
MTHT	98.01	99.33	98.66	98.50	99.50	99.00
QHELC	98.67	99.33	99.33	99.33	99.83	99.50

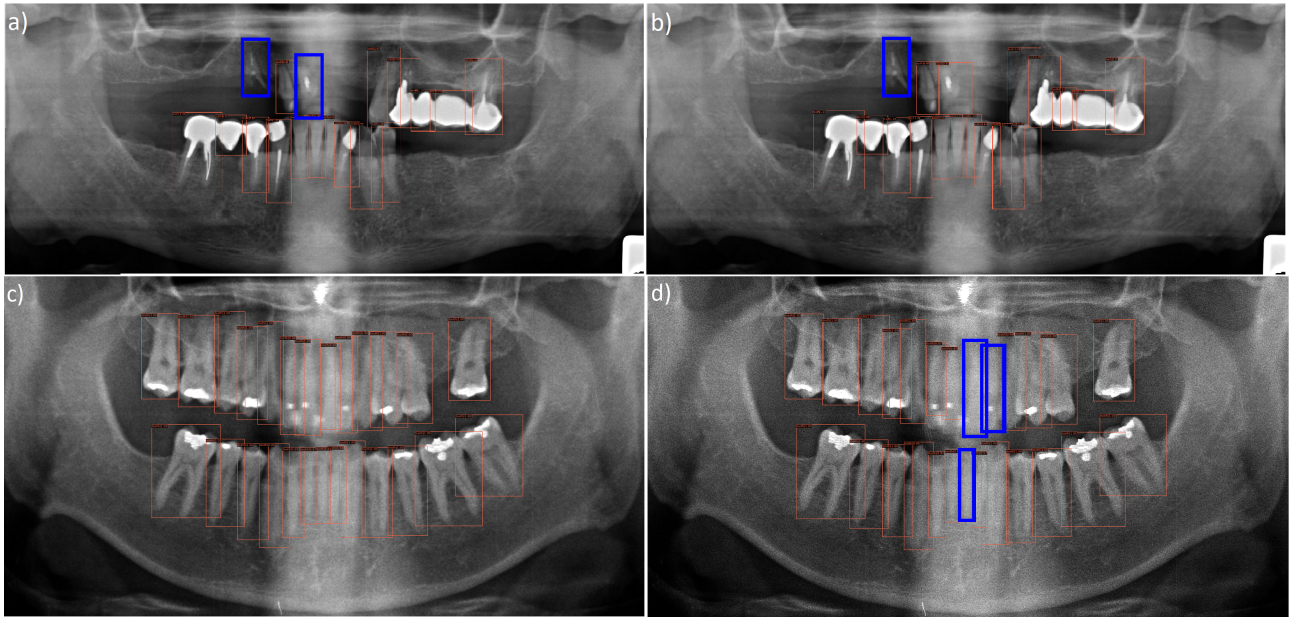


Figure 4.14: Above images: inference performed by the Faster R-CNN baseline model in **a)** is compared with the Maximum model in **b)**. Below images: inference performed by the baseline model **c)** is compared against the Edge Enhancement model **d)**.

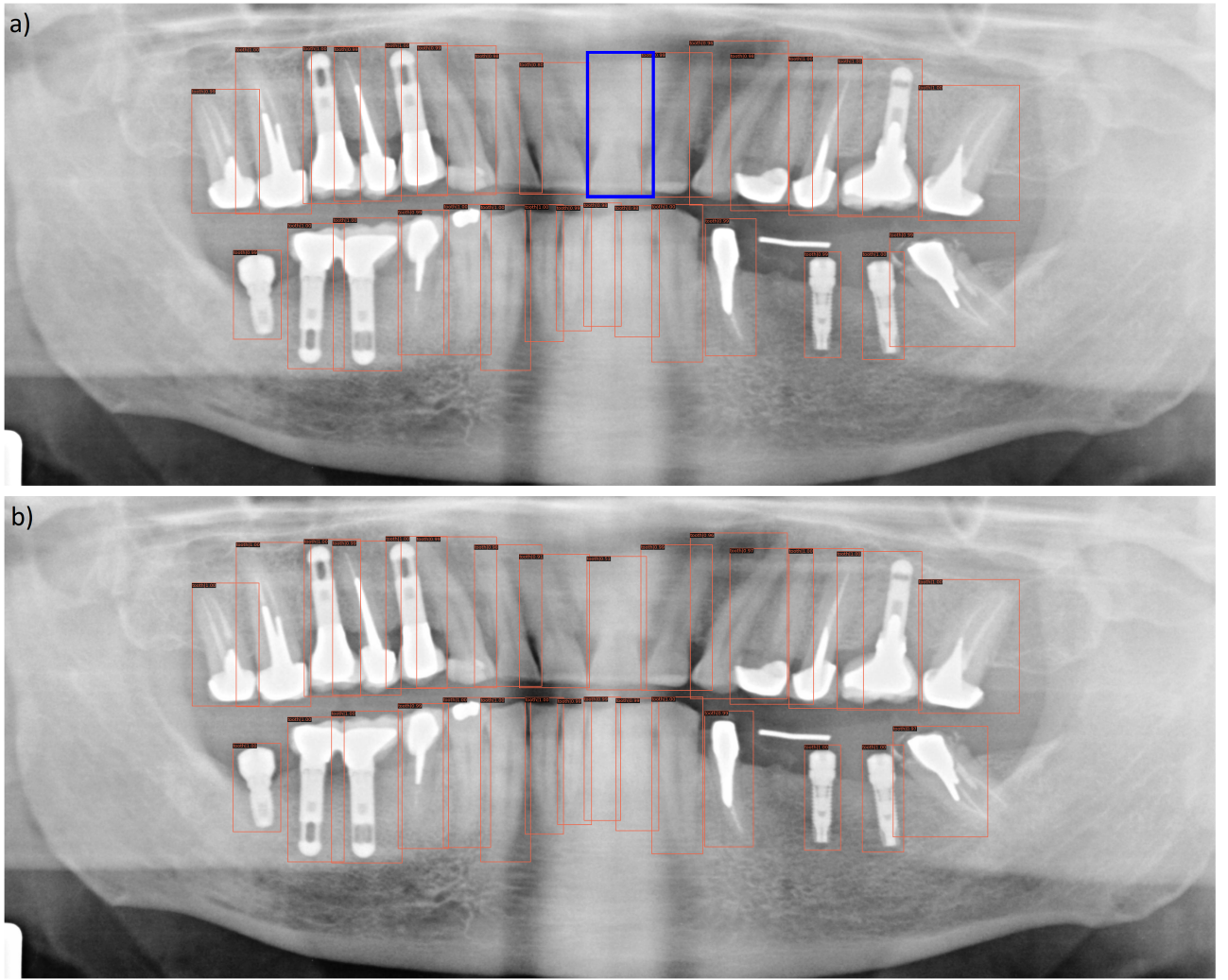


Figure 4.15: Inference performed by Cascade R-CNN baseline model in **a)** is compared against QHELC model in **b)**. Here we can see that while the QHELC model can detect all teeth in the image, the baseline model missed one of the teeth.

The correlation heat maps of Faster and Cascade R-CNN are shown in Figure 4.16, where **A** corresponds to Faster R-CNN and **B** to Cascade R-CNN. A close look at the heat map **A** shows a very weak negative relationship between the detection evaluation metrics Accuracy and Recall with the IQM entropy, EPI, and CIR. There is also present a very weak positive relationship between these two detection evaluation metrics and the SSIM and the global contrast. This means that the Accuracy and Recall of a model will tend to decrease if there is an increment in the entropy and local contrast of the images, and an enhancement in the edges. Although, an increment in the global contrast of the images would be beneficial for the values achieved by these metrics. The BBHE is the technique with the highest mean entropy and global contrast, and the second-highest value of CIR in Table 4.7, which affects the Accuracy and Recall achieved by the BBHE model. Also, the Edge Enhancement presents the lowest Recall among the models while CLAHE achieves the second-best Recall, these techniques present the highest value of EPI and the second-highest global contrast, respectively. In addition, the heat map shows that the Precision is very weakly negatively affected by SSIM, which means that

the Precision would improve if the difference between the contrast and luminance, and the correlation coefficient of the enhanced images and the original images increase. It should be noted that this negative relationship between Precision and the SSIM metric is only present in this heat map.

The relationships among the IQM for these 20 images are also shown in the heat map **A**. The correlation coefficients show that the entropy has a strong positive one relationship with the global contrast, and CIR has a very weak negative relationship with the global contrast, not a positive as with the images of the testing set. This means that when the global contrast of the images is increased, the entropy will also increase, although the local contrast will tend to mildly decrease.

In the heat map belonging to Cascade R-CNN models, the very weak negative relation of -0.28 between AMBE and the Precision stands out, along with a very weak positive relation of 0.18 between the Recall and the global contrast. The first relation, also present in the heat maps of Figure 4.11, means that changes in the brightness of the images would have a negative impact on the Precision results. CLAHE and BBHE are the two techniques that introduce the most change to the images' brightness, and the corresponding models present the lowest values of Precision among Cascade R-CNN detection results. The very weak positive relationship between the Recall and the global contrast, along with the very weak negative relationship between the Recall and CIR, means that the Recall would benefit from an increment in the global contrast but not in the local contrast of these images. This can be verified by observing that CLAHE, as mentioned before, achieves the highest Recall among Cascade R-CNN models.

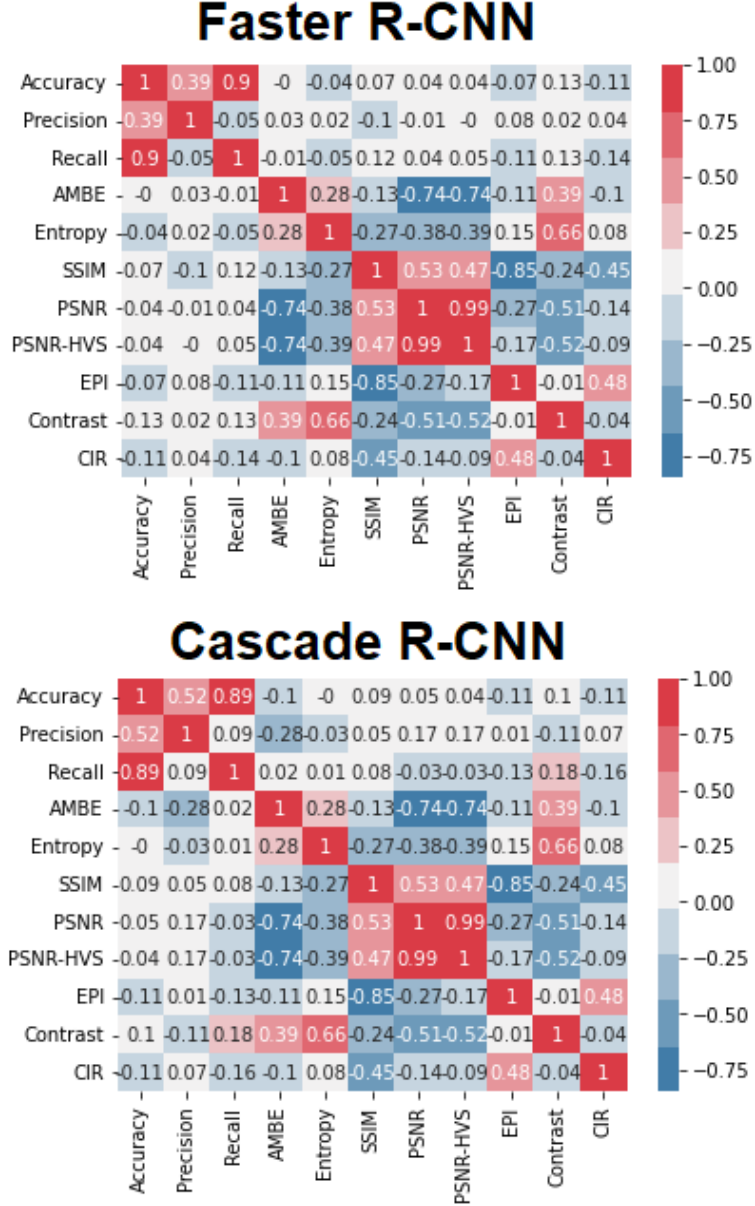


Figure 4.16: Correlation heat maps of Faster and Cascade R-CNN models, formed using the detection results of Table 4.8 and the IQM of each of the 20 images.

4.4.1 Discussion

The detection results show that the behavior of the models tested in images outside the original dataset is similar to the results shown in section 4.2, which means that the models are robust enough to generalize on images with very different features to those of the panoramic images of Silva et al. (2018). On the other hand, the heat maps do present some minor differences, which can be due to the variability between the features in these images, resulting in some different correlations with and between the IQM values. For increasing the relevance of these results, a larger number of annotated images would be necessary for fine-tuning the pre-trained models instead of using the models trained with the dataset of 300 images to make the inference on the local images. Using images taken with different equipment would also enrich the results.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This work presented the correlation results between teeth detection results and IQM, where the detection was performed by different object detection architecture, and ten IET were used as a pre-processing step. The dataset used was composed of 300 panoramic images belonging to a public dataset by Silva et al. (2018). After trimming the images to exclude the extra toothless region, the IET were applied to the images. The selected IET consist of five denoising or edge sharpening filters, and five contrast enhancement techniques. For quantifying the changes introduced to the images by the IET, eight IQM were selected and measured in the testing set. The set of images without any IET applied to them and the ten sets of enhanced images were used to train eleven models using Faster R-CNN, RetinaNet, Cascade R-CNN, FCOS, and YOLOv₃ architectures. The Accuracy, Precision, and Recall of each model were computed to evaluate the detection performance on the panoramic images. These evaluation metrics were measured using IoU thresholds of 0.5 and 0.75.

In order to find the relationship between the IQM measured in the enhanced panoramic images and the detection evaluation results, the cross-correlation between these values was calculated for each architecture. The relationships found are shown using two heat maps per architecture, one per threshold, with Cascade R-CNN showing the strongest relationships among the architectures. According to its heat maps, the Precision has a moderate positive relationship with PSNR, and PSNR-HVS, and a moderate and strong negative relationship with AMBE and the global contrast, respectively. These relationships are in fact present in the heat maps of the four other architectures, with the exception that they are very weak. Whereas for the Recall, this metric shows a very weak negative relationship with CIR, PSNR, and PSNR-HVS, and very weak positive relationships with the entropy and global contrast, also

present in the rest of the heat maps. These relationships mean that the Precision is negatively affected by an IET that introduces changes in the brightness and global contrast of the images, which increases the model possibility of having more false positives. While if an IET preserves the structural similarities without introducing much noise or distortion to the images, the model trained with this IET will be more likely to have fewer false positives, thus achieving a better Precision score. On the other hand, when a significant change in the local contrast is introduced into the images, the number of false negatives is more likely to increase, affecting the Recall achieved by a model. Although, a moderate increment in the global contrast and an improvement of the entropy of the images will result in fewer false negatives, allowing to achieve better Recall values.

QHELC introduces low, sometimes insignificant, changes to the brightness, entropy, global, and local contrast of the images, having high values of SSIM, PSNR, PSNR-HVS, and EPI. The model trained with the QHELC images achieves one of the best Accuracy, Precision, and Recall in Cascade R-CNN. In fact, the QHELC model has better detection results than the baseline model in four of the five chosen architectures. Cascade R-CNN QHELC model achieved the best performance among the models and architectures with an Accuracy of 99.49%, a Precision of 99.94%, and a Recall of 99.54%. The models trained with the Bilateral Filtering applied to the images also perform better than the baseline model in three out of five architectures. Regarding the models with the lowest values of the detection metrics, the model that presents the lowest Recall in four of the five architectures is the model trained with the Edge Enhancement images, while BBHE and CLAHE present the lowest Precision values in three of the architectures. The images enhanced with the Edge Enhancement technique present the highest change in the local contrast, hence presenting a high value of CIR. Whereas, the images enhanced with BBHE and CLAHE present the greatest changes in their brightness and global contrast. The low values of Precision achieved by BBHE and CLAHE models is understood not only because this metric is negatively affected by the increment in the AMBE and global contrast values, but also because incrementing the brightness, global contrast, and local contrast of the images decreases the values of SSIM, PSNR, and PSNR-HVS measured.

Faster R-CNN heat maps present a contradictory negative relation between the Recall and the IQM: PSNR, PSNR-HVS, and CIR. It is a contradictory relationship because it does not correlate to the fact that the MTHT and Bilateral Filtering models obtain one of the best Recall in this architecture. With the Bilateral Filtering images presenting the highest PSNR and PSNR-HVS values, and the MTHT images with one of the highest CIR values. This negative relation is also present in YOLOv₃ heat maps, but in this architecture, the Edge Enhancement, Anisotropic Diffusion, Bilateral Filtering, and MTHT models present the lowest values of Recall. What it does correlate to the aforementioned about images enhanced with these IET, images with high PSNR and PSNR-HVS, Anisotropic Diffusion and Bilateral Filtering, and images with a large change in local contrast, Edge Enhancement and MTHT. The heat maps of Faster R-CNN also show very weak negative relations of the Accuracy with AMBE and CIR and the Precision with AMBE and global contrast. Relationships that are related to

the fact that BBHE and CLAHE, as was mentioned in the paragraph above, are models with low Accuracy and Precision values, but high Recall. As a matter of fact, the Faster R-CNN CLAHE model achieves the highest Recall, with a value of 99.66%, among all models and architectures. CLAHE models of Cascade R-CNN and RetinaNet also achieve the best recall among the models in each of these architectures, meaning that this model achieves the best Recall value in three of the chosen architectures. This is related to the positive relationship between the Recall and the entropy and global contrast, since CLAHE introduces an increment in the global contrast, increasing the entropy of the images, but not the local contrast.

In addition to the experiments on the dataset of 300 images, an extra experiment was carried out by performing inference in 20 images belonging to a local clinic. The detection results of this inference are similar to the results achieved by the models of Faster R-CNN and Cascade R-CNN architectures, which were used to perform teeth detection in these 20 images. On the other hand, the correlation results show some different relationships. Differences that might be related to the fact that there is high variability among these images' features. One of them is the negative relationship between the Precision and SSIM metric, present in the Faster R-CNN heat map. This implies that the Precision will tend to improve when using an IET that increases the difference between the luminance, global contrast, and the structural similarity of the original and the enhanced images. The relationships among the IQM values also present some differences with the correlations found among the IQM measured on the 60 images of the testing set. For the 20 images, the entropy has a strong positive relationship with the global contrast, and CIR has a very weak negative relationship with the global contrast. It is important to mention that the relevance of the results obtained in these 20 images would increase if the number of images is at least equal to the dataset used in the first experiments.

5.2 Future work

Future works, identified based on the results obtained, are listed below:

1. To make the experiments described in this work with a larger number of images.

Deep learning algorithms need a large set of data in order to learn from and generalize. In addition to the number of images in the dataset, the variability in it will provide better results for less frequent scenarios, such as tooth fragments and teeth in unexpected anatomical locations and positions. In addition, if the images are captured by different equipment, the results will become more relevant, since they will more closely resemble what happens in reality.

2. To optimize the parameters of the IET that improve the detection results by testing with different values.

Since the Bilateral Filtering model outperformed the baseline model in more than one architecture, taking into consideration the parameters to apply this IET to the images

could have an impact in the detection results. For the Bilateral Filtering, performing tests with different size of the filter, as well as the smoothing intensities of the domain and space filters, could lead to values that improve the detection results even more.

3. To use the IET that improve the detection results as a pre-processing step for the detection of pathologies or for a different task as segmentation.

Without tests and experiments similar to those developed in this work, it cannot be concluded that the IET that improved the detection results, will also improve the detection of pathologies and the segmentation of teeth in panoramic images, however, they could be a starting point to be considered.

4. To determine which aspects of an object detection architecture correlate with the characteristics of images enhanced with the IET, seeking to find a better understanding of the object detection process when using neural networks.
5. To perform a test of significance of the correlation results obtained, in order to determine the relevance of these results, and discard whether they are a product of chance derived from the images, the chosen IET, and architectures.
6. To study the impact of the IET applied to images belonging to different categories in the detection results.

Applying a contrast enhancement technique, for example, to images with metal appliances, implant, and restorations will have a resulted image with different features than an image with only natural teeth. This could be impacting the detection results in the images, which should be study in order to further understand the impact of applying IET to images as pre-processing.

CHAPTER 6

SUMMARY IN SPANISH

Las imágenes radiológicas dentales son utilizadas como complemento del análisis clínico llevado a cabo por profesionales, apoyando el proceso de diagnóstico y la planificación eficaz de un tratamiento (Jader et al., 2018). Las distintas radiografías dentales permiten la visualización de dientes y estructuras circundantes desde diferentes ángulos dentro de la boca. Entre las mismas, las imágenes panorámicas son las más solicitadas por profesionales debido a que permiten la visualización de todo el conjunto de dientes y de la estructura ósea circundante (Jader et al., 2018; Tuzoff et al., 2019). Sin embargo, la interpretación y lectura de las mismas es compleja, requiriendo un tiempo considerable de parte del profesional de salud bucal (H. Chen et al., 2019).

De manera a reducir la carga laboral de odontólogos, se ha venido trabajando en sistemas computacionales de análisis automático de imágenes, con una mayoría de trabajos basados en redes neuronales enfocados a la detección y segmentación de dientes (Schwendicke et al., 2019). Si bien los resultados obtenidos por algoritmos de aprendizaje profundo han ido en aumento desde la introducción de las redes neuronales convolucionales (CNN por sus siglas en inglés) (Betul, 2017; Chung et al., 2021), un sistema para la localización de estructuras anatómicas debe alcanzar una exactitud comparable a la de profesionales para su uso como herramienta auxiliar en el diagnóstico. Alcanzar este alto nivel de Exactitud sigue siendo un desafío aún para algoritmos de detección basados en redes neuronales profundas (Chung et al., 2021). Sin embargo, el resultado obtenido por un algoritmo basado en redes neuronales también está relacionado con técnicas de preprocesamiento, aumento de datos para el entrenamiento y el afinamiento de los parámetros (Litjens et al., 2017). De hecho, algunos métodos para la detección de dientes publicados recientemente utilizan técnicas de post-procesamiento y enfoques relacionados con la red neuronal para mejorar los resultados obtenidos por sus métodos (ver Betul, 2017; H. Chen et al., 2019; Chung et al., 2021; Muramatsu et al., 2020; Tuzoff et al.,

2019; Zhang et al., 2018).

La implementación de técnicas de preprocesamiento y aumento de datos ha probado ser un factor diferencial entre algoritmos de redes neuronales profundas con la misma arquitectura (Litjens et al., 2017). Sin embargo, seleccionar una técnica de preprocesamiento con el objetivo de obtener mejoras en los resultados de una arquitectura no es una tarea trivial. Es por ello que en este trabajo se ha estudiado el impacto de diferentes técnicas de realce de imagen en la detección de dientes en imágenes panorámicas realizada por 5 arquitecturas de aprendizaje profundo. Para tal fin, un total de 10 técnicas de realce (ver sección 3.2.1) fueron aplicadas como preprocesamiento a las imágenes panorámicas del conjunto de datos. Este conjunto de datos está compuesto de 300 imágenes seleccionadas de un conjunto de imágenes panorámicas públicamente disponible (ver sección 3.1). La salida de las técnicas de realce fue medida con diferentes métricas de calidad de imagen (ver sección 3.2.2), mientras que la detección de dientes fue evaluada hallando la Exactitud, Precisión y Sensibilidad. Finalmente, la correlación entre estos valores fue hallada a manera de determinar la relación existente entre las características de las imágenes y los resultados en la detección obtenidos.

El primer paso para llevar a cabo los experimentos consistió en la selección de imágenes panorámicas del conjunto de imágenes preparado por Silva et al. (2018). Este banco de imágenes consta de 1500 imágenes divididas en 10 categorías, las cuales se diferencian por contener imágenes con o sin 32 dientes, con o sin restauraciones, con o sin implantes, con o sin aparatos dentales (ver Tabla 3.1). La selección consistió en elegir un conjunto de imágenes de cada una de estas categorías, buscando mantener la variabilidad de este banco de imágenes (ver Tabla 3.2). Seguidamente se procedió a preparar las anotaciones para la detección de dientes, lo que consiste en dibujar un rectángulo de tamaño mínimo, llamado *bounding box* en inglés, alrededor de cada diente en las imágenes. Las anotaciones fueron realizadas usando la herramienta para anotación *VGG Image Annotator (VIA)* (Dutta & Zisserman, 2019), y posteriormente corregidas y validadas por dos experimentados odontólogos.

El análisis automático de las imágenes panorámicas es especialmente compleja por tratarse de una imagen bidimensional de una estructura tridimensional, lo que resulta en estructuras anatómicas superpuestas, y por la similitud entre estas estructuras y tejidos (Ehsani Rad et al., 2013; Jader et al., 2018). Para un algoritmo de detección basado en redes neuronales profundas, que requiere de tantas imágenes como sea posible proveerle para generalizar el problema con una exactitud comparable a la de profesionales, la variación entre pacientes constituye además otro desafío, sobre todo si el conjunto de imágenes es pequeño. Adicionalmente, las imágenes radiográficas dentales se ven afectadas por artefactos permanentes y temporales (ver sección 2.2) que representan un desafío extra para un algoritmo de detección. Con el objetivo de explorar los efectos en la detección de dientes usando diferentes métodos de aprendizaje profundo, 10 técnicas de realce fueron aplicadas como preprocesamiento a las imágenes panorámicas.

Los filtros para eliminación de ruido con preservación de bordes *Difusión Anisotrópica* (Perona & Malik, 1990; Sapiro & Ringach, 1996) y *Filtro Bilateral* (Tomasi & Manduchi, 1998) fueron dos de las técnicas de realce seleccionadas. A continuación se aplicaron los filtros de *di-*

latación (Máximo) y *erosión* (Mínimo), aplicados para expandir (dilatación) y reducir (erosión) el tamaño de las formas en las imágenes. Además, el operador Laplaciano (ver ecuación 3.2) fue aplicado de manera a obtener el *Realce de los Bordes* usando la máscara de 3×3 mostrada en la matriz 3.4. Las siguientes 5 técnicas seleccionadas representan algoritmos de realce de contraste: el *algoritmo de ecualización del histograma dividido en dos para la preservación de brillo* (BBHE por sus siglas en inglés) (Kim, 1997), *la ecualización de histograma adaptativo de contraste limitado* (CLAHE por sus siglas en inglés) (Sasi & Jayasree, 2013), *la reconstrucción geodésica multi-escala y morfológica para el realce de contraste* (GRMMCE por sus siglas en inglés) (Mello et al., 2020), *la transformación multi-escala Top-Hat* (MTHHT por sus siglas en inglés) (Mello et al., 2017), y *la ecualización del histograma partido en cuatro con limitación de contraste* (QHELC por sus siglas en inglés) (Brizuela et al., 2019). Las figuras 3.2 al 3.11 muestran imágenes panorámicas realzadas con cada una de las técnicas mencionadas.

Las imágenes realzadas fueron evaluadas con diferentes métricas de calidad de imagen (IQM por sus siglas en inglés).

- El *error de brillo medio absoluto* (AMBE por sus siglas en inglés) fue utilizada para medir el cambio en el brillo de las imágenes al calcular la diferencia absoluta entre el brillo medio de imagen original y realzada (ver ecuación 3.9). Esta métrica es tal que cuanto mayor es la diferencia entre el brillo de las imágenes, mayor es el valor de AMBE.
- La *entropía* fue calculada como medida de la información en las imágenes (ver ecuación 3.10), representada por la cantidad de bits necesarios para codificar a las mismas (Roman et al., 2019). Imágenes con mejor nivel de detalle requieren un mayor número de bits para ser codificadas, lo que significa que mayor es su valor de entropía.
- La similitud estructural entre las imágenes de entrada y salida fue medida con la métrica *medida del índice estructural de similitud* (SSIM por sus siglas en inglés). La misma provee una medida de la distorsión introducida a una imagen al comparar el brillo, el contraste local y el coeficiente de correlación usando la ecuación 3.11. La similitud de las imágenes será igual a 1 si es que las imágenes son idénticas, y cuanto más se aproxime este valor a cero, más diferencia existirá entre las imágenes comparadas.
- La métrica *relación señal pico a ruido* (PSNR por sus siglas en inglés), fue medida para conocer el nivel de ruido introducido a las imágenes por las técnicas de realce (ver ecuación 3.13). La métrica PSNR-HVS fue utilizada con la misma finalidad, con la salvedad de que esta última tiene en cuenta el sistema visual humano (HVS por sus siglas en inglés), al calcular el nivel de ruido con la ecuación ?? en 3.13. Un valor elevado de ambas métricas indica un bajo nivel de ruido, lo que significa que calidad de la imagen fue preservada por la técnica aplicada.
- La preservación de bordes fue medida con la métrica propuesta por Sattar et al. (1997) usando la ecuación 3.16. El valor dado por esta ecuación se acerca a 1 cuanto mayor es

la similitud entre las imágenes, lo que implica una mejor conservación de bordes de la imagen original.

- El contraste global de las imágenes fue calculado al medir el contraste promedio con la ecuación 3.18. Cuanto mayor es el valor medido por esta ecuación, mayor es el contraste de la imagen.
- Para medir la diferencia del contraste local entre dos imágenes, se utilizó la métrica *relación de mejora de contraste* (CIR por sus siglas en inglés). El valor de esta métrica, dado por la ecuación 3.20, será mayor cuanto mayor sea la diferencia entre el contraste local de las imágenes.

Las arquitecturas seleccionadas para realizar los experimentos fueron las arquitecturas de dos etapas Faster R-CNN (Ren et al., 2015) y Cascade R-CNN (Cai & Vasconcelos, 2019), junto con las arquitecturas de etapa única RetinaNet (Lin et al., 2018), FCOS (Tian et al., 2019) y YOLOv₃ (Redmon & Farhadi, 2018). Las arquitecturas de dos etapas realizan la detección de objetos al pasar primeramente la imagen por una red basada en CNN. Esta red provee como salida un mapa de características de la imagen de entrada, que es seguidamente utilizada como entrada de una red neuronal que se encarga de proponer regiones de interés y clasificarlas como objetos o no objetos. En la siguiente etapa, estas regiones son clasificadas y ajustadas por el detector de objetos al final de la red neuronal profunda (ver figuras 3.19 y 3.20). Por su parte, las arquitecturas de etapa única realizan la detección de una manera similar a las de dos etapas, con la diferencia principal que no cuentan con una red neuronal que proponga regiones de interés (ver figuras 3.21, 3.23 y 3.24). Esto hace que estas arquitecturas sean más veloces a la hora de entrenarlas, pero obtienen resultados por lo general inferiores a las arquitecturas de dos etapas.

La detección de dientes, realizada por los diferentes modelos entrenados con las arquitecturas mencionadas, fue evaluada al calcular las métricas Exactitud, Precisión y Sensibilidad. Mientras que la clasificación de bounding boxes como diente o fondo fue realizada teniendo en cuenta dos valores de superposición de la métrica intersección sobre unión (IoU por sus siglas en inglés), la cual está definida según la ecuación 3.23. Los valores de superposición usados fueron la $IoU \geq 0.5$, comúnmente elegido para medir el rendimiento de algoritmos de detección, y la $IoU \geq 0.75$, valor que requiere mayor exactitud en la predicción de objetos (ver figura 3.25). Las métricas Exactitud, Precisión y Sensibilidad se calculan al hallar primeramente el número de verdaderos positivos (TP por sus siglas en inglés), falsos positivos (FP) y falsos negativos (FN) de cada modelo teniendo en cuenta los valores de IoU elegidos. Donde un verdadero positivo es un bounding box correctamente clasificado como diente, un falso positivo es un bounding box clasificado como diente, pero que realmente no contiene un diente, y un falso negativo es un diente no detectado por el modelo (ver figura 3.26). La Exactitud mide la habilidad de un modelo en realizar una predicción correcta al expresar el porcentaje de TP del total de verdaderos positivos, falsos positivos y falsos negativos ($TP + FP + FN$) (ver

ecuación 3.24). La Precisión mide la habilidad de un modelo en diferenciar entre dientes y fondo, al expresar el porcentaje de TP entre todas las detecciones (TP + FP) (ver ecuación 3.25). La Sensibilidad mide la habilidad de un modelo en detectar todos los dientes presentes en el conjunto de imágenes, expresando el porcentaje de TP del total de dientes (TP + FN) (ver ecuación 3.26).

Los experimentos fueron llevados a cabo luego de recortar las 300 imágenes del conjunto de datos. El séptum nasal, el borde inferior de la mandíbula y las ramas posterior izquierda y derecha de la mandíbula fueron los puntos usados como referencia para recortar las imágenes con el objetivo de excluir zonas sin dientes. Para tal fin, 100 imágenes, elegidas al azar de las categorías, fueron anotadas con un único bounding box con sus lados en los puntos anatómicos mencionados. Seguidamente, la arquitectura Faster R-CNN fue usada para entrenar un modelo, hacer la detección de la zona con dientes, y recortar automáticamente las imágenes (ver figura 3.27). Con las imágenes recortadas, se aplicaron las técnicas de realce a las 300 imágenes para luego entrenar un modelo por técnica con 192 imágenes en el conjunto de entrenamiento, 48 en el conjunto de validación y 60 en el conjunto de prueba. Esto resulta en un total de 11 modelos entrenados por arquitectura, el primero entrenado con las imágenes originales, el modelo base, y 10 entrenados con una técnica de realce aplicada como preprocesamiento.

El entrenamiento de los modelos se realizó seguidamente a la selección de modelos pre-entrenados de cada arquitectura, seleccionados de la variedad de modelos disponibles de MMDetection (K. Chen et al., 2019), y del afinamiento de hiperparámetros. Los modelos pre-entrenados con el conjunto de datos MS-COCO (Lin et al., 2014) fueron seleccionados con sus respectivos pesos al considerar los modelos que alcanzaron los valores más altos de la métrica $mAP = AP_{IoU=0.5:0.05:0.95}$ (ver tabla 3.3). La métrica mAP es una de las métricas de evaluación de detección utilizadas por MS-COCO y se calcula promediando 10 valores de IoU, de 0.5 a 0.95 con pasos de 0.05.

El afinamiento de hiperparámetros fue realizado usando solo las imágenes originales. Los hiperparámetros considerados para el afinamiento fueron: tamaño del lote (batch size en inglés), tasa de aprendizaje (learning rate en inglés), disminución del peso (weight decay en inglés), iteraciones de calentamiento (warmup iterations en inglés) y los pasos para disminuir la tasa de aprendizaje (steps en inglés). Estos fueron ajustados considerando el mejor valor de la métrica $AP_{IoU=0.5:0.05:0.95}$ en el conjunto de validación (ver tabla 3.4).

El entrenamiento de los modelos fue realizado por 100 épocas, con tiempos de entrenamiento de menos de una hora para YOLOv3, y más de 4 horas para los modelos entrenados en Faster R-CNN, Cascade R-CNN, RetinaNet y FCOS. El optimizador usado en cada arquitectura fue el descenso de gradiente estocástico (SGD por sus siglas en inglés), mientras para el aumento de datos se utilizó la técnica de volteo horizontal con una probabilidad del 0.5. El volteo vertical no se utilizó debido a que las imágenes panorámicas no pueden ser tomadas con el maxilar hacia abajo. Sin embargo, que los dientes del lado derecho estén del lado izquierdo o viceversa, es irrelevante para la detección de dientes, ya que no realizamos la clasificación individual de dientes según alguna nomenclatura dental.

La inferencia en el conjunto de prueba fue realizada al seleccionar los mejores pesos de los modelos entrenados en cada arquitectura. Para ello se consideró el valor más elevado de $AP_{IoU=0.5:0.05:0.95}$ en el conjunto de validación. Seguidamente, el número de verdaderos positivos, falsos positivos y falsos negativos de cada modelo fue usado para el cálculo de la Exactitud, Precisión y Sensibilidad considerando los umbrales de IoU mencionados más arriba. Los valores de estas métricas, junto con los valores de IQM medidos en las imágenes realzadas, fueron utilizados para calcular los valores de correlación cruzada. El conjunto de pasos llevado a cabo para desarrollar los experimentos puede verse en la figura 3.28.

El afinamiento de los modelos, el entrenamiento y la inferencia fueron todos realizados en el navegador usando Google Colaboratory (Bisong, 2019). Esta plataforma de uso de GPUs gratuita fue corrida en el navegador Google Chrome en una computadora de marca Acer modelo Aspire A513-53, con un procesador Intel Core i5-7200U CPU 2.5 GHz y una RAM de 4 GB.

Los valores de IQM medidos en las imágenes del conjunto de prueba (ver tabla 4.1) indican que las técnicas BBHE, CLAHE, Realce de Bordes y MTHT son las que un mayor cambio introducen a las imágenes. BBHE, seguido de CLAHE, es la técnica que más eleva al contraste global de las imágenes, presentando además los valores más bajos de SSIM, PSNR y PSNR-HVS. Por otra parte, QHELC es la técnica que presenta el valor más elevado de SSIM, también presentando valores elevados de PSNR, PSNR-HVS y EPI. Siendo una técnica que eleva levemente el contraste global de las imágenes. Además de QHELC, el Filtro Bilateral y la Difusión Anisotrópica son las dos técnicas que menor distorsión y ruido introducen a las imágenes, obteniendo los valores más elevados de PSNR y PSNR-HVS. Los valores de EPI de estas técnicas de eliminación de ruido son mayores a los alcanzados por MTHT, la dilatación y erosión, con esta última presentando el valor más bajo de EPI. Respecto al brillo medio de las imágenes, ninguna técnica introdujo un cambio considerable a esta característica, con CLAHE y BBHE presentando los valores más elevados de AMBE y la Difusión Anisotrópica el valor más bajo de esta métrica. En lo que respecta al contraste local de las imágenes, el Realce de Bordes, seguido de BBHE y MTHT, son las técnicas que presentan los valores más elevados de CIR en la tabla 4.1. La técnica CLAHE presenta los valores más elevados de la entropía y EPI, mientras que obtiene el segundo valor más bajo de PSNR y PSNR-HVS.

Los resultados de detección de dientes llevada a cabo por los diferentes modelos en cada arquitectura se presentan en las tablas 4.2 al 4.6. Mientras que con el objetivo de encontrar qué características de las imágenes se relacionan con mejoras en la detección de dientes, el cálculo de la correlación cruzada fue realizado usando los valores de IQM y los resultados de la detección de cada arquitectura. La correlación correspondiente a cada arquitectura es mostrada a través de mapas de calor, uno por cada umbral IoU seleccionado para realizar la inferencia en las imágenes de prueba.

Para la arquitectura Faster R-CNN, un total de 7 modelos tienen una mejor Exactitud que el modelo base cuando el valor de IoU es de 0.5. Presentando la mejor Exactitud el modelo entrenado con las imágenes realzadas con MTHT, seguido del modelo entrenado por el Filtro Bilateral. Mientras que los valores más bajos de esta métrica fueron obtenidos por

el modelo BBHE y el modelo entrenado con las imágenes a las cuales se les aplicó la erosión (filtro Mínimo). El valor más elevado de Sensibilidad fue alcanzado por el modelo CLAHE seguido por QHELC, mientras que los modelos Realce de Bordes y filtro Mínimo presentan los valores más bajos de Sensibilidad. A pesar de ello, el modelo Realce de Bordes alcanza el valor más elevado de Precisión, seguido por el Filtro Bilateral y MTHT. Cuando el valor de IoU fue cambiado a 0.75, la Exactitud y Precisión del modelo Filtro Bilateral aumentaron debido a que el número de FP se redujo a 0. El modelo MTHT también presenta mejores valores de Exactitud y Precisión con este umbral. De hecho, estos dos modelos son los únicos con mejores valores de Exactitud y Sensibilidad que el modelo base, cuando el umbral IoU es de 0.75. Respecto a los mapas de calor correspondientes a esta arquitectura (ver Figura 4.9), los mismos muestran algunas relaciones negativas muy débiles. Según ambos mapas, la Exactitud se verá negativamente afectada por cambios en el brillo y el contraste local de las imágenes, al mostrarse una relación negativa entre esta métrica y AMBE y CIR. Mientras que la Precisión también presenta una relación negativa con AMBE, además de la entropía y el contraste global. Por último, la Sensibilidad posee una relación negativa muy débil con PSNR, PSNR-HVS y CIR, indicando que esta métrica se ve afectada positivamente por el ruido introducido pero negativamente por cambios en el contraste local.

En la tabla de resultados de la arquitectura RetinaNet, tabla 4.3, los modelos GRMMCE, Filtro Bilateral y QHELC presentan una mejor Exactitud que el modelo base. Mientras que el mejor valor de Sensibilidad es alcanzado por el modelo CLAHE y la Precisión del modelo base no es superada por ninguno de los modelos. Por otra parte, el modelo con la Exactitud y Precisión más baja es BBHE, y el modelo con la Sensibilidad más baja es el de Realce de Bordes. Cuando el valor de IoU es cambiado a 0.75, la Sensibilidad de los modelos decrece, afectando también a la Exactitud, debido a que los bounding boxes poseen IoU menores a 0.75. Con este valor de IoU, los modelos CLAHE, filtro Máximo y QHELC poseen una mejor Exactitud y Precisión que el modelo base, con el filtro Mínimo presentando los peores valores de estas métricas. En los mapas de calor de RetinaNet en la figura 4.10, se puede ver que la Precisión está negativamente relacionada con el contraste global. Lo que nos indica que esta métrica se verá negativamente afectada frente a cambios en el contraste global de las imágenes, ejemplo de ello es el modelo entrenado con las imágenes realzadas con BBHE.

Entre los modelos entrenados en Cascade R-CNN, tabla 4.4, el modelo QHELC es el único que tiene una Exactitud y Sensibilidad ligeramente mejor que el modelo base. Por otra parte, el modelo con el peor rendimiento es nuevamente el modelo BBHE, teniendo una Precisión 5.37% menos que la Precisión del modelo QHELC. Mientras que cuando IoU toma el valor 0.75, los modelos GRMMCE y QHELC tienen los mismos resultados de clasificación y son seguidos por el modelo base, el cual presenta el segundo mejor rendimiento. En lo que respecta a los mapas de calor de esta arquitectura (ver figura 4.11), la relación positiva de la Exactitud y Precisión con las métricas SSIM, PSNR y PSNR-HVS es una relación que destaca. Esta relación positiva moderada, presente con esta intensidad solo en los mapas de calor de Cascade R-CNN, indica que principalmente la Precisión de un modelo tenderá a ser mejor cuando el cambio en el brillo

medio y el contraste global de las imágenes, además del ruido introducido, se mantienen al mínimo posible. Por otra parte, estos mapas también muestran una relación negativa entre la Exactitud y Precisión con AMBE y el contraste global. Lo que está en resonancia con el hecho de que el modelo BBHE alcanza el valor más bajo de Precisión entre los modelos de esta arquitectura.

Los resultados de los modelos entrenados usando FCOS se presentan en la tabla 4.5. De acuerdo a los mismos, el modelo QHELC logra los mejores resultados y es seguido por el modelo base, mientras que el modelo BBHE presenta la Exactitud más baja entre los modelos. Además, con excepción del filtro Máximo, todos los modelos presentan una Precisión perfecta ($FP = 0$), lo cual se debe a que los modelos no detectan por error el fondo o la región entre dientes como dientes. Finalmente, según los resultados de la detección realizada por los modelos en YOLOv₃ (Tabla 4.6), los valores alcanzados por el modelo base no fueron superados por ningún otro modelo. Cuando el umbral IoU es 0.75, los valores de detección de los modelos decrecen en más del 10%. Tanto los mapas de calor de FCOS (ver figura 4.12) como los mapas de calor de YOLOv₃ (ver figura 4.13), muestran relaciones muy débiles entre las IQM consideradas y los resultados en la detección de dientes. La cercanía entre los valores de Exactitud presentes en la tabla 4.5 podrían ser la razón por la cual los mapas de calor de esta arquitectura muestran relaciones tan leves.

En un análisis generalizado de los mapas de calor, se puede ver que los mapas de Faster R-CNN, RetinaNet, Cascade R-CNN y YOLOv₃ muestran que la Precisión tiene una relación negativa con AMBE y el contraste global. Esto implica la Precisión se verá afectada negativamente por una variación en el brillo y un incremento en el contraste global de las imágenes. Mientras que la Sensibilidad muestra una relación negativa muy débil con CIR, PSNR y PSNR-HVS, y una relación positiva muy débil con la entropía y el contraste global. Esto significa que la Sensibilidad se ve afectada negativamente por el incremento del contraste local, pero se beneficia si los valores de PSNR y PSNR-HVS son bajos, del mejoramiento de detalles de la imagen y una mejora moderada del contraste global. En la tabla 4.1, podemos ver que el Realce de Bordes es la técnica que más cambio introdujo al contraste local de las imágenes, presentando el valor más elevado de CIR. Esto repercute en el número de falsos negativos que obtiene el modelo entrenado con imágenes realizadas con dicha técnica, y es la razón por la que presenta el menor valor de Sensibilidad en cuatro de las arquitecturas. Por otra parte, CLAHE presenta valores bajos de CIR, PSNR y PSNR-HVS, y el mejor valor de entropía junto con el segundo valor más elevado de contraste global, el modelo correspondiente alcanza el mejor o uno de los mejores valores de Sensibilidad también en cuatro arquitecturas. La Exactitud y Precisión del modelo BBHE están entre las más bajas en cuatro de las arquitecturas, Faster R-CNN, RetinaNet, Cascade R-CNN y YOLOv₃. Esto se debe a que BBHE es la técnica que un mayor cambio introduce al contraste global y brillo de las imágenes. Mientras que el modelo Realce de Bordes presenta también un bajo valor de Exactitud en las cinco arquitecturas, a consecuencia de la Sensibilidad alcanzada por el modelo. Esto podría verse como un resultado inesperado que podría cambiar si el realce de bordes se realiza con una máscara diferente a la

utilizada en este trabajo (ver matriz 3.4), aunque esto solo puede tomarse como una suposición que debe probarse.

Respecto a las relaciones entre las mismas métricas de calidad de imagen, presentadas en el mapa de calor **A** de la figura 4.9 en la sección amarilla, estas establecen que las métricas SSIM, PSNR y PSNR-HVS se ven afectadas negativamente por cambios en el brillo, la entropía, el contraste global, el contraste local y la distorsión de los bordes. Lo cual se debe a las relaciones negativas entre estas tres métricas y AMBE, la entropía, contraste global, CIR y EPI, respectivamente. La métrica que más afecta a la métrica SSIM es CIR, mientras que PSNR/PSNR-HVS se ven más afectados por el incremento en el contraste global. Esto se verifica claramente con los valores presentes en la tabla 4.1, donde las técnicas Realce de Bordes, BBHE y MTHT presentan el CIR más alto y el SSIM más bajo, mientras que BBHE y CLAHE presentan el mayor contraste global y el menor PSNR/PSNR-HVS.

En cuanto a las relaciones positivas entre las IQM, la existente entre SSIM, PSNR y PSNR-HVS es muy fuerte, de hecho, PSNR y PSNR-HVS tienen una relación casi 1 a 1. Además, se encuentran las relaciones entre CIR, EPI y el contraste global, el contraste global y AMBE, EPI y la entropía, y AMBE con la entropía. El hecho de que EPI tenga una relación positiva con CIR y el contraste global significa que el valor de EPI tenderá a aumentar siempre que se modifique el contraste local y aumente el contraste global de las imágenes. Aumentar el contraste global de las imágenes también aumentará el brillo medio de las imágenes, lo que dará como resultado un valor más alto de AMBE, y este incremento en el brillo, junto con la preservación/realce de los bordes, dará como resultado una leve mejora del valor de entropía de una imagen.

Al término de las pruebas con las cinco arquitecturas seleccionadas, un experimento adicional fue realizado usando 20 imágenes pertenecientes al Instituto Periodontal - Centro de Diagnóstico Periodontal, Asunción, Paraguay. El mismo fue realizado a fin de determinar si los resultados de correlación encontrados podrían reproducirse en un conjunto de imágenes diferente. Este conjunto de imágenes, a diferencia del conjunto de 300 imágenes, contiene imágenes en ocasiones borrosas o con alto/bajo contraste y con gran variabilidad de características de imagen. Al igual que se procedió con las 300 imágenes, las 20 imágenes fueron inicialmente recortadas para excluir la zona sin dientes. Seguidamente se aplicaron las técnicas de realce y se utilizaron los modelos entrenados en Faster R-CNN y Cascade R-CNN para realizar la inferencia.

Los valores medidos de las métricas de calidad de imagen en las 20 imágenes, junto con la desviación estándar de los mismos, es mostrada en la tabla 4.7. En la tabla podemos ver que la Difusión Anisotrópica y el Filtro Bilateral vuelven a ser los IET que menos ruido y distorsión introducen en las imágenes, siendo la técnica QHELC la que tiene mejores valores de SSIM y EPI, y el tercer mejor PSNR/PSNR-HVS. Mientras que la técnica BBHE introduce el mayor aumento en el contraste global de las imágenes, y el Realce de Bordes en el contraste local. El contraste global de estas imágenes presenta una mayor variabilidad que las 60 imágenes del conjunto de prueba, que también está presente en el contraste global final de las imágenes

mejoradas, a excepción de la técnica BBHE, que presenta la desviación estándar más baja para esta métrica.

Entre los resultados de los modelos Faster R-CNN, presentes en la tabla 4.8, el modelo GR-MMCE obtiene mejores resultados de Exactitud y Precisión que el modelo base, mientras que el filtro Máximo obtiene la mejor Sensibilidad. Además, el modelo MTHHT presenta el valor más bajo de Exactitud y el de Realce de Bordes la Sensibilidad más baja. Respecto a la inferencia realizada por los modelos de Cascade R-CNN, QHELC es el único modelo con una Exactitud mayor a la del modelo base, con CLAHE obteniendo la Sensibilidad más elevada. El mapa de calor de Faster R-CNN (ver figura 4.16 **A**) muestra una relación negativa muy débil entre la Exactitud y la Sensibilidad con la entropía, EPI y CIR. Estas métricas también presentan una relación positiva muy débil con SSIM y el contraste global. Esto significa que mejorar los detalles de las imágenes, realzar los bordes e introducir un cambio en el contraste local de las imágenes no sería beneficioso para estas métricas, pero sí un leve incremento del contraste global. Además, el mapa de calor muestra que la Precisión se ve afectada negativamente por SSIM, lo que significa que la Precisión mejoraría si aumentara la diferencia entre el contraste global y la luminancia, y el coeficiente de correlación de las imágenes realzadas y las imágenes originales. Cabe señalar que esta relación entre Precisión y la métrica SSIM solo está presente en este mapa de calor.

Las relaciones entre las IQM medidas en las 20 imágenes también se muestran en el mapa de calor **A**. Donde se encuentran algunas diferencias con las relaciones encontradas en el mapa de calor 4.9 **A**. Los coeficientes de correlación muestran que la entropía tiene una fuerte relación positiva con el contraste global, y CIR tiene una relación negativa muy débil con el contraste global, no positiva como con las imágenes del conjunto de prueba. Esto significa que cuando aumenta el contraste global de las imágenes, la entropía también aumentará, aunque el contraste local tenderá a disminuir ligeramente.

En el mapa de calor perteneciente a los modelos Cascade R-CNN (ver figura 4.16 **B**), destaca la débil relación negativa de -0,28 entre AMBE y la Precisión, junto con una muy débil relación positiva de 0.18 entre la Sensibilidad y el contraste global. La primera relación, también presente en los mapas de calor de la Figura 4.11, significa que los cambios en el brillo de las imágenes tendrían un impacto negativo en los resultados de Precisión. La relación positiva entre la Sensibilidad y el contraste global, junto con la relación negativa entre dicha métrica y CIR, significa que la Sensibilidad se beneficiaría de un aumento en el contraste global pero no en el contraste local de estas imágenes. Esto se puede verificar observando que CLAHE, como se mencionó anteriormente, logra la mayor Sensibilidad entre los modelos Cascade R-CNN.

Con base en los resultados hallados en el trabajo, se pueden hacer las siguientes afirmaciones:

- El uso de una técnica de realce como preprocesamiento para el entrenamiento de algoritmos de detección no necesariamente resultará en mejores resultados en la detección de dientes en imágenes panorámicas.
- La Precisión en la detección de dientes se ve negativamente afectada por cambios en el

brillo y contraste global de las imágenes, debido a que aumenta la posibilidad de que un modelo presente un mayor número de falsos positivos. Prueba de ello son las técnicas BBHE y CLAHE, las cuales introducen los mayores cambios al brillo y contraste global de las imágenes, presentando valores bajos de Precisión en 4 de las arquitecturas.

- Un modelo entrenado con imágenes realzadas con una técnica que preserve las similitudes estructurales de las imágenes originales, a la vez de introducir el menor ruido posible a las mismas, tenderá a presentar un menor número de falsos positivos, es decir, podrá alcanzar una mejor Precisión. En este aspecto, las técnicas de realce QHELC y Filtro Bilateral presentan elevados valores de SSIM, PSNR, PSNR-HVS y EPI, los modelos correspondientes obtienen mejores resultados de detección que el modelo base en la mayoría de arquitecturas.
- Cambios en el contraste local de las imágenes tienen un impacto en el número de falsos negativos de un modelo, lo que afecta a la Sensibilidad del mismo. Ejemplo de esto es la técnica Realce de Bordes, cuyo modelo presenta la menor Sensibilidad en 4 de las arquitecturas. Esta técnica posee valores bajos de SSIM, PSNR y PSNR-HVS, a la vez del mayor valor de CIR entre las técnicas aplicadas.
- Un incremento moderado del contraste global y la entropía de las imágenes resulta en un menor número menor de falsos negativos, lo que está en resonancia con la relación positiva entre estas métricas de calidad de imagen y la métrica Sensibilidad. Respecto a esta relación, el modelo CLAHE obtiene los mejores valores de Sensibilidad en tres de las arquitecturas. Las imágenes realzadas con CLAHE presentan un moderado incremento del contraste global, un valor más elevado de entropía, pero no así de contraste local.
- Los resultados de correlación no pueden expandirse a otros conjuntos de imágenes panorámicas, lo que se evidencia por los resultados obtenidos en los mapas de calor correspondientes al experimento con las 20 imágenes pertenecientes a una clínica local. Aunque es importante recalcar que la relevancia de los resultados obtenidos con estas 20 imágenes incrementaría si el número de imágenes es al menos igual al utilizado para el resto de los experimentos.

En lo que respecta a los trabajos futuros, identificados con base en los resultados obtenidos, uno de los más importantes es realizar los experimentos con un número mayor de imágenes panorámicas. Además, que las imágenes sean adquiridas con diferentes equipos y presenten la mayor variabilidad posible es un punto igualmente relevante para acercarse a un escenario lo más real posible. Otro interesante experimento sería optimizar los parámetros de las técnicas que han obtenido mejores resultados de detección, lo cual podría ayudar a obtener aún mejores resultados. Asimismo, usar las técnicas de realce que mejoraron los resultados de detección de dientes en la detección de patologías dentales o en una tarea diferente como segmentación de dientes. Es decir, realizar experimentos con las mismas a fin de corroborar si poseen el mismo

impacto que en la detección de dientes. Finalmente, determinar qué aspectos de una arquitectura para la detección de objetos se correlacionan con características de imágenes realzadas por diferentes técnicas. A fin de encontrar una mejor comprensión del proceso de detección de objetos utilizando redes neuronales.

BIBLIOGRAPHY

- Amer, Y., & Aqel, M. (2015). An efficient segmentation algorithm for panoramic dental images. *Procedia Computer Science*, 65, 718–725. <https://doi.org/10.1016/j.procs.2015.09.016>
- Bai, X. (2013). Top-hat by reconstruction operators based multi-scale multi-structuring element method for multiple linear feature detection with simple post-processing. *Optik*, 124(20), 4246–4251. <https://doi.org/https://doi.org/10.1016/j.ijleo.2012.12.073>
- Bai, X., Zhou, F., & Xue, B. (2012). Image enhancement using multi scale image features extracted by top-hat transform. *Optics and Laser Technology*, 44, 328–336. <https://doi.org/10.1016/j.optlastec.2011.07.009>
- Barker, J., Sarathy, S., & Tao, A. (2016). *Detectnet: Deep neural network for object detection in DIGITS*. <https://developer.nvidia.com/blog/detectnet-deep-neural-network-object-detection-digits/>
- Bengio, Y. (2009). *Learning deep architectures for AI* (Vol. 2).
- Betul, A. (2017). Tooth detection with convolutional neural networks. *2017 Medical Technologies National Congress (TIPTEKNO)*, 1–4. <https://doi.org/10.1109/TIPTEKNO.2017.8238075>
- Bisong, E. (2019). Google colab. https://doi.org/10.1007/978-1-4842-4470-8_7
- Brizuela, I., Medina Caballero, R. D., Silva, J., Román, J., & Vázquez Noguera, J. (2019). Quadri-histogram equalization using cutoff limits based on the size of each histogram with preservation of average brightness. *Signal, Image and Video Processing*, 13. <https://doi.org/10.1007/s11760-019-01420-9>
- Brownlee, J. (2021). *A gentle introduction to object recognition with deep learning*. [Accessed: 2021-06-24]. <https://machinelearningmastery.com/object-recognition-with-deep-learning/>
- Cai, Z., & Vasconcelos, N. (2019). Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 1483–1498. <https://doi.org/10.1109/TPAMI.2019.2956516>
- Chen, H., Zhang, K., Lyu, P., Li, H., Zhang, L., Wu, J., & Lee, C.-H. (2019). A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports*, 9. <https://doi.org/10.1038/s41598-019-40414-y>
- Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., ...

- Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.
- Chen, S.-D., & Ramli, A. (2003). Minimum mean brightness error bi-histogram equalization in contrast enhancement. *IEEE Transactions on Consumer Electronics*, 49(4), 1310–1319. <https://doi.org/10.1109/TCE.2003.1261234>
- Chung, M., Lee, J., Park, S., Lee, M., Lee, C. E., Lee, J., & Shin, Y.-G. (2021). Individual tooth detection and identification from dental panoramic x-ray images via point-wise localization and distance regularization. *Artificial Intelligence in Medicine*, 111, 101996. <https://doi.org/10.1016/j.artmed.2020.101996>
- Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2016/file/577ef1154f3240ad5b9b413aa7346a1e-Paper.pdf>
- Dutta, A., & Zisserman, A. (2019). The VGG Image Annotator (VIA). *ArXiv*, abs/1904.10699.
- Egiazarian, K., Astola, J., Lukin, V., Battisti, F., & Carli, M. (2006). A new full-reference quality metrics based on HVS. *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, 22–24.
- Ehsani Rad, A., Rahim, M., Rehman, A., Altameem, A., & Saba, T. (2013). Evaluation of current dental radiographs segmentation approaches in computer-aided applications. *IETE Technical Review*, 30, 210–222. <https://doi.org/10.4103/0256-4602.113498>
- Facon, J. (2002). *Processamento e análise de imagens* [Available online at: <https://web.tecgraf.puc-rio.br/~scuri/inf1378/pub/CursoProcImagem.pdf>].
- Fei-Fei, L., Justin, J., & Serena, Y. (2018). *Lecture 11: Detection and segmentation*. [Available online at: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf].
- Gonzalez, R. C., & Woods, R. E. (2001). *Digital image processing* (2nd). Addison-Wesley Longman Publishing Co., Inc.
- Gordon, R., & Rangayyan, R. (1984). Feature enhancement of film mammograms using fixed and adaptive neighborhoods. *Applied optics*, 23, 560. <https://doi.org/10.1364/AO.23.000560>
- Greenewald, K., Melnyk, I., & Kingsbury, B. (2019). *Estimating information flow in deep neural networks*. [Accessed: 2021-06-28]. <https://www.ibm.com/blogs/research/2019/06/deep-neural-networks/>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Horé, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM, 2366–2369. <https://doi.org/10.1109/ICPR.2010.579>
- Jader, G., Fontineli, J., Ruiz, M., Lima, K., Pithon, M., & Oliveira, L. (2018). Deep Instance Segmentation of Teeth in Panoramic X-ray Images. *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 400–407. <https://doi.org/10.1109/SIBGRAPI.2018.00058>

- Kamarudin, J., Abdullah, A., & Sallehuddin, R. (2017). A review of deep learning architectures and their application. https://doi.org/10.1007/978-981-10-6502-6_7
- Kim, Y.-T. (1997). Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE Transactions on Consumer Electronics*, 43(1), 1–8. <https://doi.org/10.1109/30.580378>
- Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25. <https://doi.org/10.1145/3065386>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–44. <https://doi.org/10.1038/nature14539>
- Lim, S., Mat Isa, N. A., Ooi, C., & Toh, K. (2013). A new histogram equalization method for digital image enhancement and brightness preservation. *Signal, Image and Video Processing*, 9. <https://doi.org/10.1007/s11760-013-0500-z>
- Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48
- Lin, T.-Y., Pattersor, G., Ronchi, M. R., Cui, Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, L., & Dollár, P. (n.d.). *Data format*. <https://cocodataset.org/#format-data>
- Lira, P., Giralidi, G., & Neves, L. A. (2010). Segmentation and feature extraction of panoramic dental x-ray images. *International Journal of Natural Computing Research (IJNCR)*, 1, 1–15. <https://doi.org/10.4018/jncr.2010100101>
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
- Mello, J., Escobar, R., Martínez, F., Vázquez Noguera, J., Legal-Ayala, H., & Pinto-Roa, D. (2020). Medical image enhancement with brightness and detail preserving using multiscale top-hat transform by reconstruction. *Electronic Notes in Theoretical Computer Science*, 349, 69–80. <https://doi.org/10.1016/j.entcs.2020.02.013>

- Mello, J., Legal-Ayala, H., & Vázquez Noguera, J. (2017). Image color contrast enhancement using multiscale morphology. *Computational Interdisciplinary Sciences*, 8. <https://doi.org/10.6062/jcis.2017.08.03.0129>
- Menotti, D., Najman, L., Facon, J., & Araújo, A. (2007). Multi-histogram equalization methods for contrast enhancement and brightness preserving. *Consumer Electronics, IEEE Transactions on*, 53, 1186–1194. <https://doi.org/10.1109/TCE.2007.4341603>
- Miki, Y., Muramatsu, C., Hayashi, T., Zhou, X., Hara, T., Katsumata, A., & Fujita, H. (2017). Classification of teeth in cone-beam ct using deep convolutional neural network. *Computers in Biology and Medicine*, 80, 24–29. <https://doi.org/https://doi.org/10.1016/j.compbiomed.2016.11.003>
- Muramatsu, C., Morishita, T., Takahashi, R., Hayashi, T., Nishiyama, W., Aiji, Y., Zhou, X., Hara, T., Katsumata, A., Aiji, E., & Fujita, H. (2020). Tooth detection and classification on panoramic radiographs for automatic dental chart filing: Improved classification by multi-sized input data. *Oral Radiology*, 37, 1–7. <https://doi.org/10.1007/s11282-019-00418-w>
- Nil, N. (1985). A visual model weighted cosine transform for image compression and quality assessment. *IEEE Transactions on Communications*, 33(6), 551–557. <https://doi.org/10.1109/TCOM.1985.1096337>
- Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7), 629–639. <https://doi.org/10.1109/34.56205>
- Pierre, S. (2004). *Morphological image analysis*. Springer.
- Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K. E., Braeuning, M. P., & Pizer, S. M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, 11(4), 193–200. <https://doi.org/10.1007/BF03178082>
- Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S. B., Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., & Chowdhury, M. E. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132, 104319. <https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104319>
- Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Roman, J., Vázquez Noguera, J., Legal-Ayala, H., Pinto-Roa, D., Gomez, S., & Garcia Torres, M. (2019). Entropy and contrast enhancement of infrared thermal images using the multiscale top-hat transform. *Entropy*, 21. <https://doi.org/10.3390/e21030244>
- Sapiro, G., & Ringach, D. (1996). Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE Transactions on Image Processing*, 5, 1582–1586.

- Sasi, N., & Jayasree, V. (2013). Contrast limited adaptive histogram equalization for qualitative enhancement of myocardial perfusion images. *Engineering*, 05, 326–331. <https://doi.org/10.4236/eng.2013.510B066>
- Sattar, F., Floreby, L., Salomonsson, G., & Lovstrom, B. (1997). Image enhancement based on a nonlinear multiscale method. *IEEE Transactions on Image Processing*, 6(6), 888–895. <https://doi.org/10.1109/83.585239>
- Schwendicke, F., Golla, T., Dreher, M., & Krois, J. (2019). Convolutional neural networks for dental image diagnostics: A scoping review. *Journal of Dentistry*, 91, 103226. <https://doi.org/10.1016/j.jdent.2019.103226>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 39, 221–248. <https://doi.org/10.1146/annurev-bioeng-071516-044442>
- Shome, S. K. K., & Vadali, S. R. K. (2011). Enhancement of diabetic retinopathy imagery using contrast limited adaptive histogram equalization. *International Journal of Computer Science and Information Technologies*, 2, 2694–2699.
- Silva, G., Oliveira, L., & Pithon, M. (2018). Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications*, 107, 15–31. <https://doi.org/10.1016/j.eswa.2018.04.001>
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.
- Tian, Z., Shen, C., Chen, H., & Tong, H. (2019). FCOS: Fully convolutional one-stage object detection, 9626–9635. <https://doi.org/10.1109/ICCV.2019.00972>
- Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 839–846. <https://doi.org/10.1109/ICCV.1998.710815>
- Tuzoff, D., Tuzova, L., Bornstein, M., Krasnov, A., Kharchenko, M., Nikolenko, S., Sveshnikov, M., & Bednenko, G. (2019). Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, 48, 20180051. <https://doi.org/10.1259/dmfr.20180051>
- American Dental Association. (2021). *Glossary of dental clinical and administrative terms*. <https://www.ada.org/en/publications/cdt/glossary-of-dental-clinical-and-administrative-terms>
- Wainschenker, R., Massa, J. M., & Tristan, P. (2011). *Procesamiento digital de imágenes*. [Accessed: 2021-07-14]. <https://users.exa.unicen.edu.ar/catedras/pdi/FILES/TE/CP1.pdf>
- Wang, C.-W., Huang, C.-T., Lee, J.-H., Li, C.-H., Chang, S.-W., Siao, M.-J., Lai, T.-M., Ibragimov, B., Vrtovec, T., Ronneberger, O., Fischer, P., Cootes, T., & Lindner, C. (2016).

- A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, 31. <https://doi.org/10.1016/j.media.2016.02.004>
- Wang, Y.-P., Wu, Q., Castleman, K., & Xiong, Z. (2003). Chromosome image enhancement using multiscale differential operators. *IEEE Transactions on Medical Imaging*, 22, 685–693. <https://doi.org/10.1109/TMI.2003.812255>
- Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 600–612. <https://doi.org/10.1109/TIP.2003.819861>
- Xiaoran, C. (2019). Image enhancement effect on the performance of convolutional neural networks. [Master’s Thesis in Computer Science. Blekinge Institute of Technology, Faculty of Computing, Department of Computer Science. Available online at: <http://urn.kb.se/resolve?urn=urn:nbn:se:bth-18523>.].
- Yirka, B. (2019). *Solving the three-body problem faster using a deep neural network*. [Accessed: 2021-06-28]. <https://techxplore.com/news/2019-11-three-body-problem-faster-deep-neural.html>
- Zhang, K., Wu, J., Chen, H., & Lyu, P. (2018). An effective teeth recognition method using label tree with cascade network structure. *Computerized Medical Imaging and Graphics*, 68, 61–70. <https://doi.org/10.1016/j.compmedimag.2018.07.001>

APPENDIX

A.1 VGG Image Annotator

The definition and labeling of the teeth in the images of the dataset were performed using VGG Image Annotator (VIA). This image annotation tool can be used in any browser using the link <https://www.robots.ox.ac.uk/~vgg/software/via/via.html> for annotating images, audio, and video (Dutta & Zisserman, 2019). Once the tool is open in the browser, the first step is to load the images to be annotated, Figure A.1 shows the basic explanation provided by the tool.

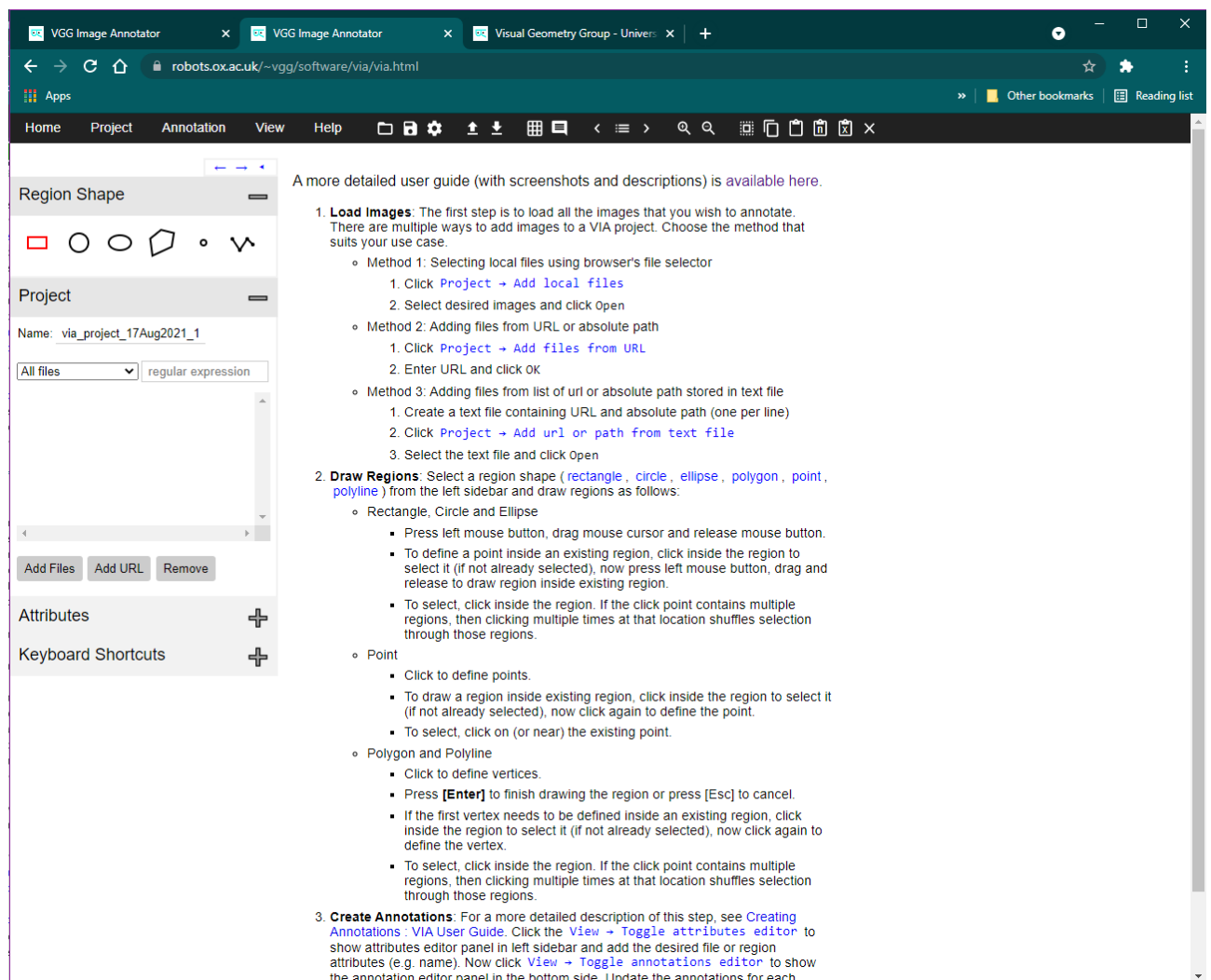


Figure A.1: VIA getting started explanation.

The next step is to proceed to select the shape that will contain the region of interest, in this case, the teeth. VIA allows you to select a rectangle, circle, ellipse, point, polygon, and polyline as shapes to draw the regions on the images depending on the image recognition task. The rectangle shape was selected as we were interested in the object detection task, after which, the tooth label was created in the *Attributes*' section of the tool. Figure A.2 shows the workplace with an annotated image of 32 teeth and the tool sections highlighted. The rectangles were drawn and labelled with the label *tooth*, for each tooth, implant, crown, bridge, tooth fragment, restored, impacted, and embedded tooth present in the set of images.

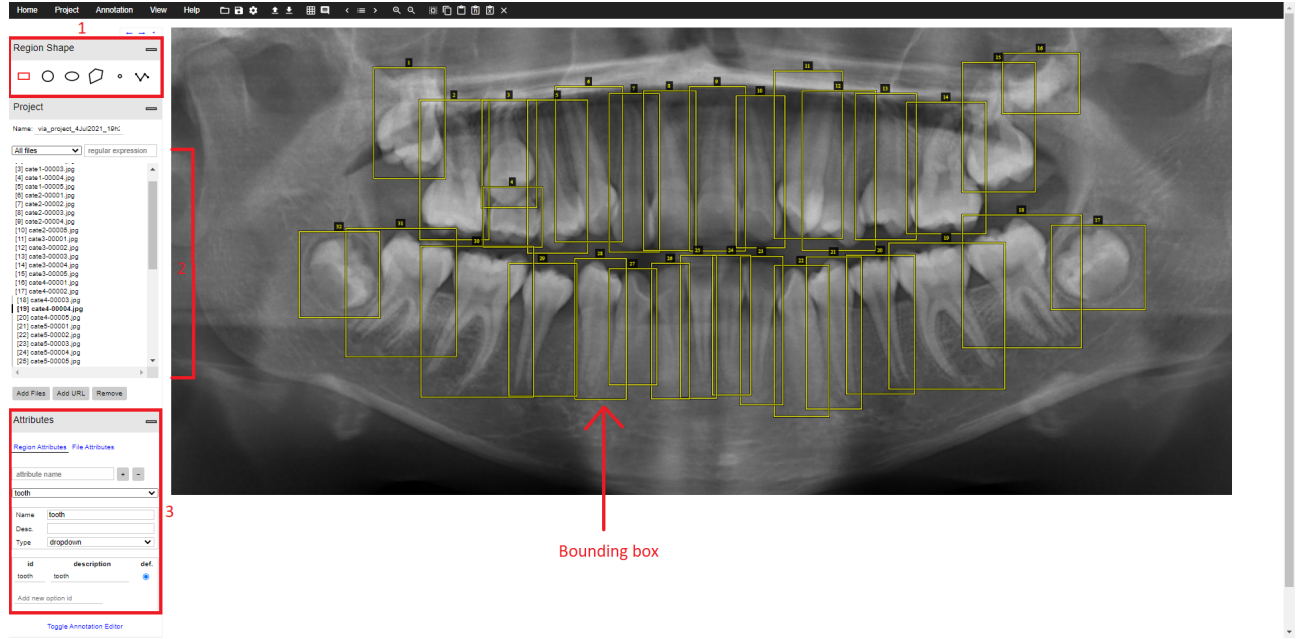


Figure A.2: The selection of the type of region for drawing in the images is chosen in the *Region Shape* area, marked with the red box number 1. In the *Project* area, marked with the number 2, the images added to the project can be seen and selected for the annotation. This area also enables removing and adding images to the project. The area in 3, *Attributes*, is where the name of the regions are added and edited, if necessary.

The set of images (training, validation, and testing) were annotated separately, uploading the set of images of each set independently and exporting them in one of the formats provided by the tool. Figure A.3 shows some annotated images of the training set. The annotation format of the images, exported initially as JSON, was converted to COCO format (Lin et al., n.d.) since the pre-trained models selected were trained with annotations in this format. The annotations of the training and testing set are partially shown in Figure A.4.

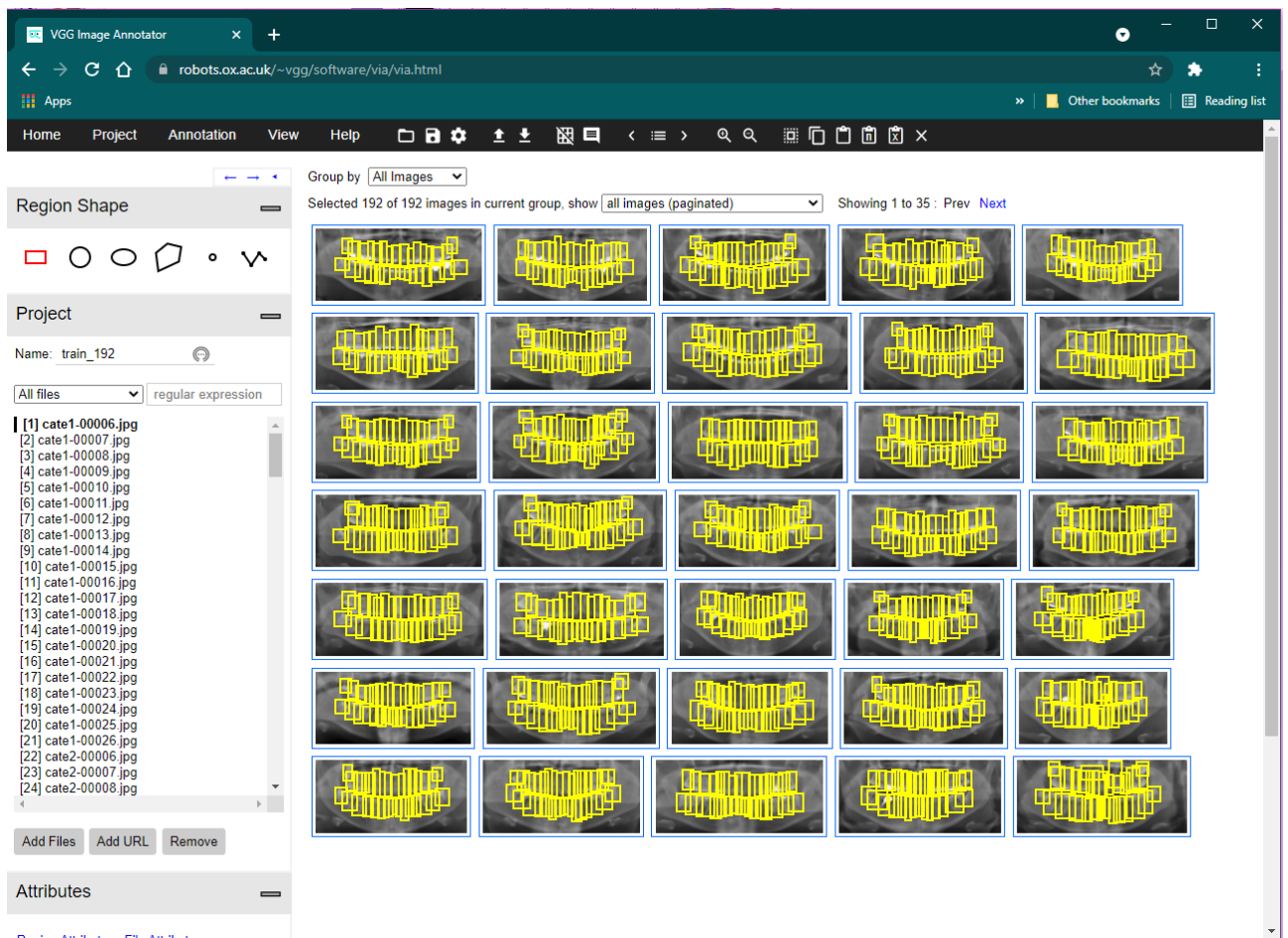


Figure A.3: Set of annotated images of the training set.

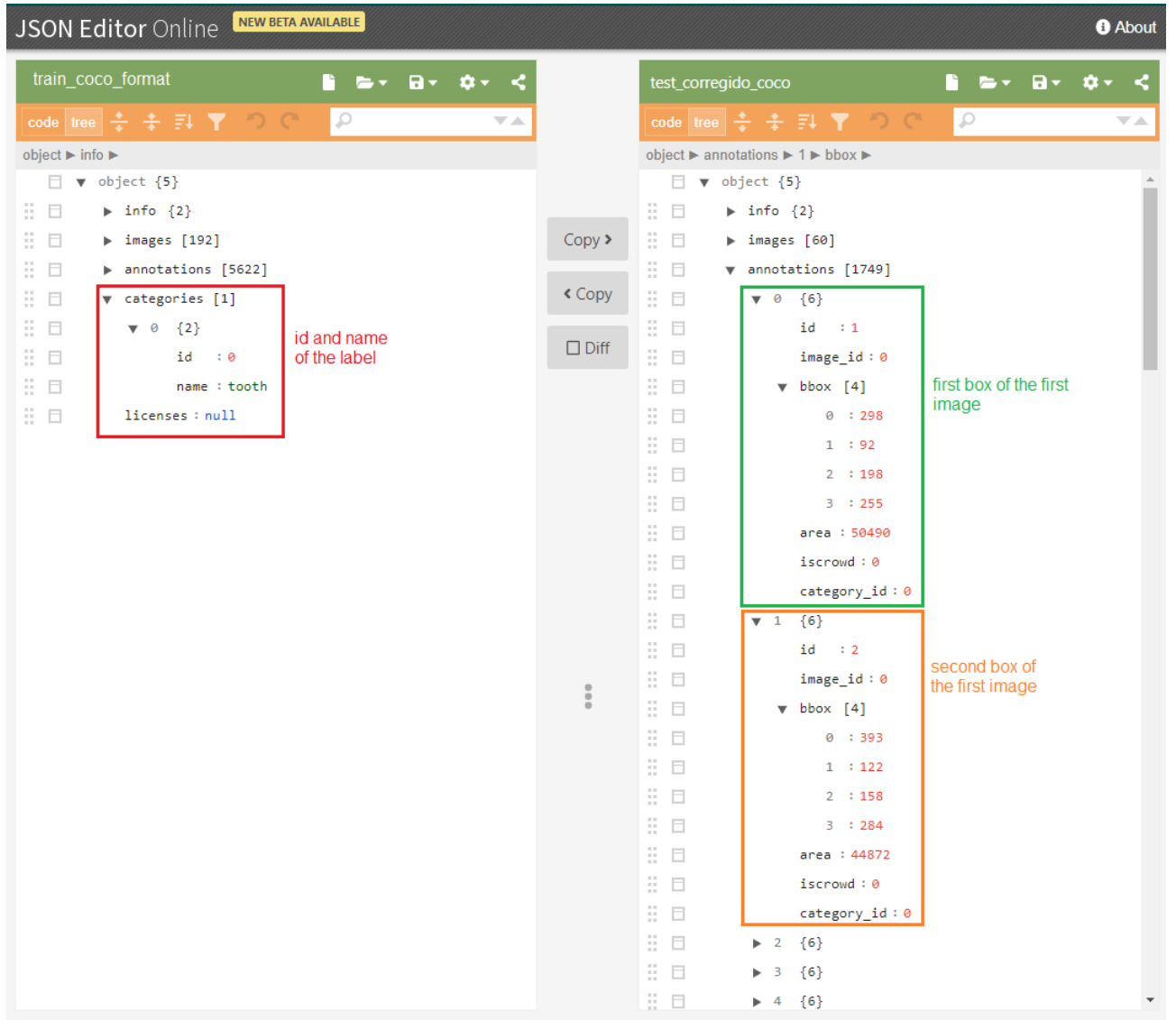


Figure A.4: View of the COCO format JSON files of the training and testing set. The code on the left shows the training set annotations files with 5622 teeth annotated in the 192 images of the set. The red box shows that the only attribute is the label tooth, which is given the ID number 0. In the code on the right, the annotations of the testing set with 1749 of teeth in the 60 images of the set, are shown with the green and yellow boxes highlighting the first and second boxes of the first image. These boxes show the ID, coordinates, are, and category ID of each bounding box. The category ID of both is 0, corresponding to the label tooth.

A.2 Inference on challenging images

The inference performed by the best models on challenging images is presented next. The presence of impacted teeth, roots, or tooth fragments and teeth in unexpected anatomical positions represent a challenge for the models of each architecture. In fact, these images are the reason some models do not achieve better performance. On some occasions, the IET applied to the images help in detecting the mentioned teeth, however, there are teeth that are not detected for any model of the five architectures, not even in Faster R-CNN or Cascade R-

CNN, the two with the best performance. The reason for this, we believe, is that there are not enough teeth with these characteristics within the dataset, making it difficult for the methods to learn these variations and be able to detect them.

The image in Figure A.5 belongs to category number 6, in which there are images with more than 32 teeth and most of them belong to children, hence the presence of non-erupted permanent teeth. The reason the image in this figure represents a challenge is the presence of three non-erupted forming teeth. There is no model, in any architecture, that was able of detecting these three non-erupted teeth, which can be appreciated in images **a)** to **f)** of the figure. The images in **a)** and **b)** correspond to the inference performed by Faster R-CNN baseline and QHELC models. Not even CLAHE, the model with the lowest number of false negatives among all architectures, was able of detecting these three teeth. However, the QHELC model detected the pair of non-erupted teeth with a single bounding box, this can be seen in image **b)**, where the section was enlarged and is indicated with the blue arrow. The image in **c)** and **d)** show the inference performed by the baseline and CLAHE models of RetinaNet. In both of them, the pair of non-erupted teeth are detected with a single bounding box that does not contain one of the non-erupted teeth entirely. These bounding boxes are shown in the enlarged section, indicated with the blue arrows. Lastly, we show the inference performed by the baseline and QHELC models of FCOS in **e)** and **f)**, respectively. In addition to not detecting the three non-erupted teeth, the baseline model is not able to detect a crowded tooth, all marked with blue boxes in image **e)**. On the other hand, QHELC, the model with the least number of false negatives in FCOS, is able of detecting the crowded tooth but not the non-erupted teeth.

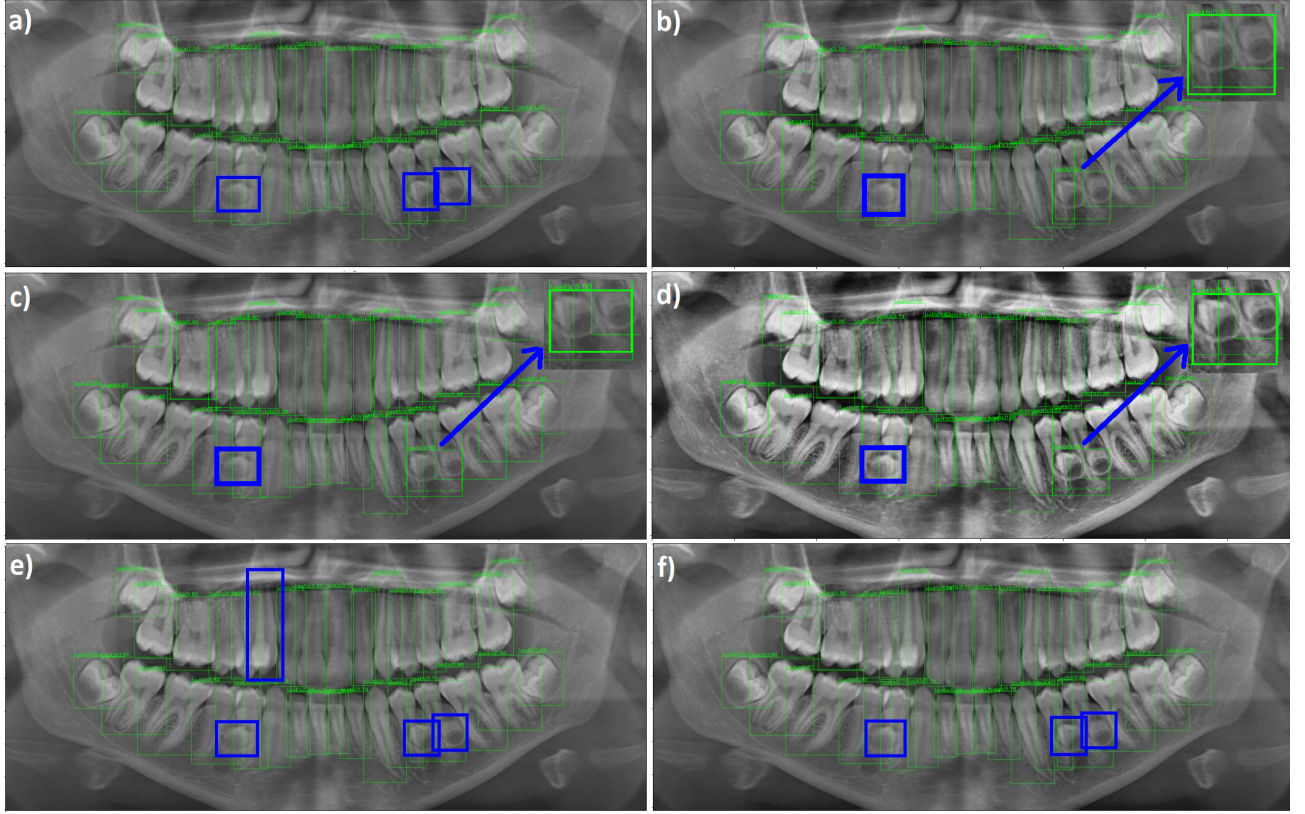


Figure A.5: Inference performed by four models on an image belonging to category 6. **a)** shows the inference performed by the baseline model and **b)** the QHELC model of Faster R-CNN. While **c)** and **d)** shows the inference performed by the baseline and CLAHE models of RetinaNet, respectively. At the bottom, **e)** and **f)** are the FCOS baseline and QHELC models, respectively.

The following image, presented in Figure A.6, belongs to category number 9. The particularity of this image is the presence of an impacted tooth, a canine, which makes it difficult to detect the tooth below, an incisor. The models of Faster R-CNN and RetinaNet are all able of detecting these teeth, whereas in Cascade R-CNN and FCOS, only 6 and 2 models of each architecture, respectively, are able to detect the 28 teeth of this panoramic image. Images **a)** and **b)** shows the inference performed by the baseline and MTHT models of Faster R-CNN, respectively. Both models predict the 28 teeth, what is interesting is the size of the boxes detecting the impacted tooth and the incisor below it. In **a)**, these bounding boxes, marked with the orange box for the impacted tooth and with the purple box for the incisor, are larger than boxes in **b)**, what also happens with the bounding boxes of these teeth in other models of this architecture. In images **c)** and **d)**, where the inference performed by Cascade R-CNN baseline and QHELC models is shown, it can be seen that the bounding boxes are not entirely containing these teeth either. Although the bounding box of the incisor in image **d)**, marked with the purple box, is bigger than the bounding boxes of this tooth in images **a)**, **b)**, **c)**, **e)**, and **f)**, almost containing the hole root. At the bottom, images in **e)** and **f)** represent the inference performed by the baseline and CLAHE models of RetinaNet, respectively. In these images we can see that we have the same scenario as with the images above regarding the

size of the boxes, only in this case, the detection made by CLAHE is not able to improve the detection made by the baseline model.

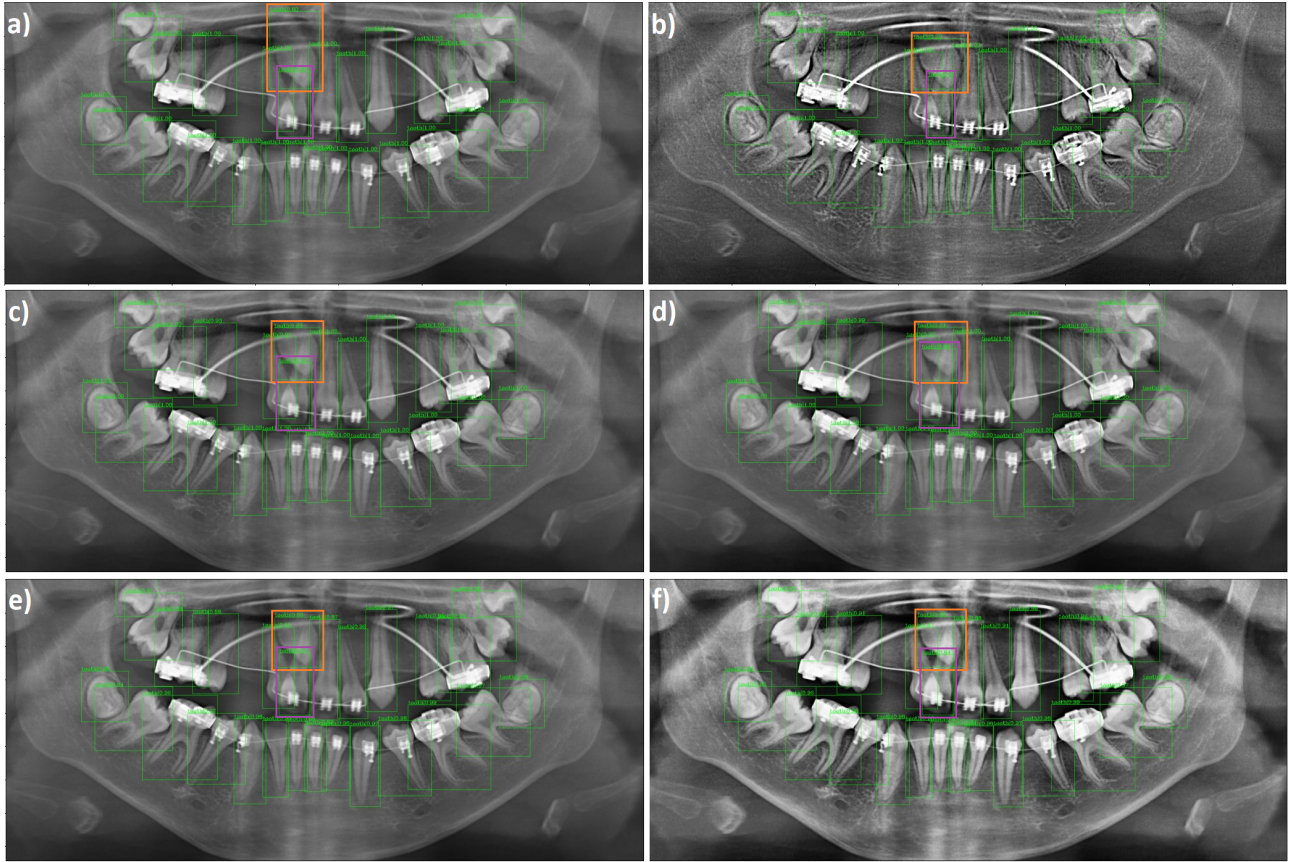


Figure A.6: Inference performed by four models on an image belonging to category 9. **a)** shows the inference performed by the baseline model and **b)** the MTHT model of Faster R-CNN. While **c)** and **d)** shows the inference performed by the baseline and QHELC models of Cascade R-CNN, respectively. At the bottom, **e)** and **f)** are the RetinaNet baseline and CLAHE models, respectively.

Figure A.7 shows the inference performed in an image of category 10. This image represents a challenge for most of the models in any architecture due to the present of a tooth near the nasal septum and three teeth fragments. Images **a)** and **b)** show the inference performed by the baseline and MTHT models of Faster R-CNN, respectively. In **a)**, the baseline model misses the detection of the tooth near the nasal septum, which is an impacted tooth, and a tooth fragment, both are marked with blue boxes. On the other hand, in **b)**, the MTHT model only misses the tooth near the nasal septum, being able to detect the tooth fragment. In fact, the models of this architecture were not able to detect at least one of the tooth fragments in this image. The images in the middle, **c)** and **d)**, show the inference performed by **c)** baseline and **d)** QHELC models of Cascade R-CNN. None of these models are able to detect the tooth near the nasal septum and a tooth fragment, marked with the blue boxes, and the double root, marked with the orange boxes, is detected with a bounding box does not contain the smaller root in its entirety. This occurs in all models of this architecture. The bottom images, **e)** and **f)**, show the inference performed by RetinaNet baseline and QHELC models. The baseline

model misses a tooth fragment, marked with the blue box in **e)**, while QHELC detect all teeth in **f)**. In these images, the double root is again detected with bounding boxes that do not contain it entirely, although only for a small range. In fact, RetinaNet QHELC is the only model, across all models and architectures, to be able to detect all teeth in this image.

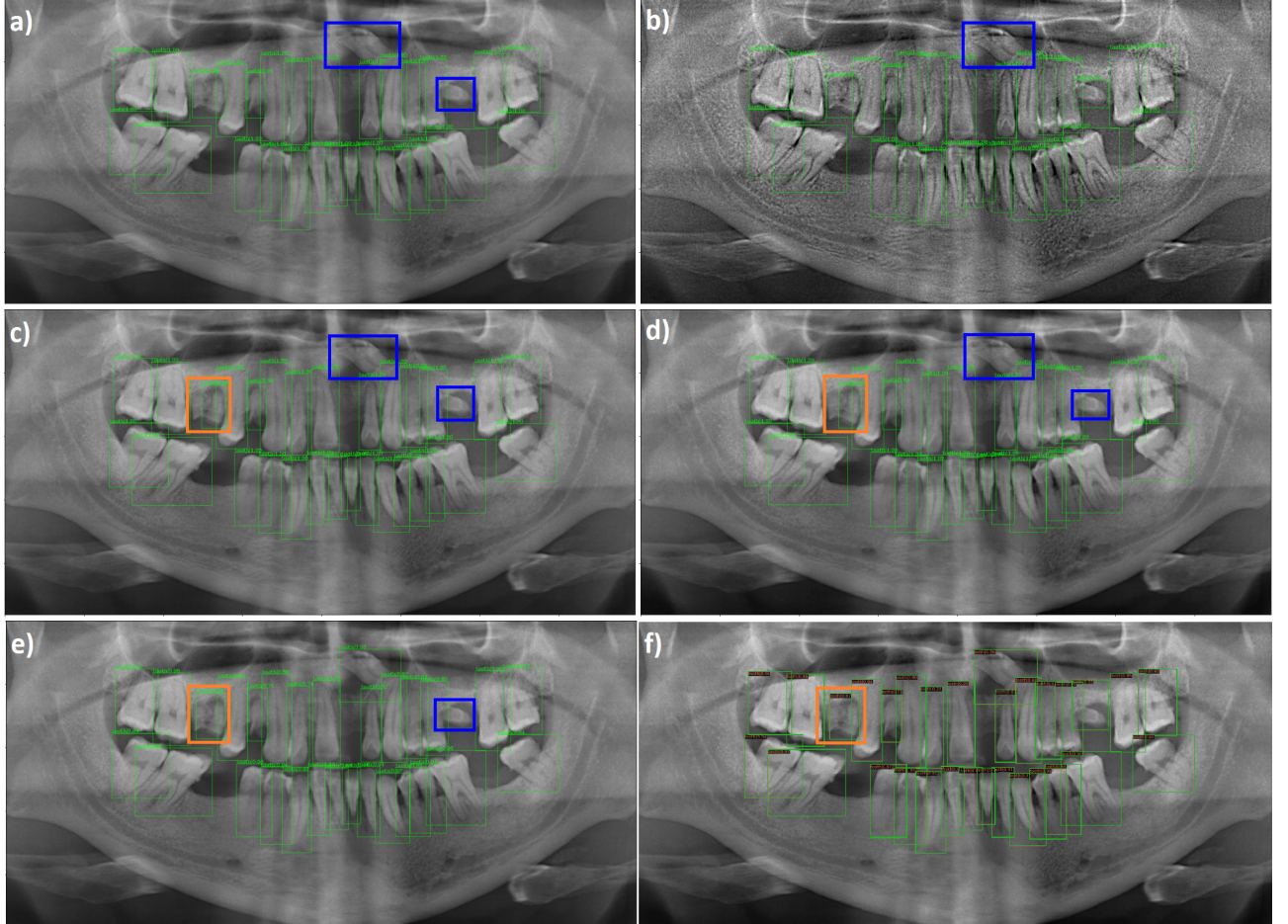


Figure A.7: Inference performed by six models on an image belonging to category 10. **a)** shows the inference performed by the baseline model and **b)** the MTHT model of Faster R-CNN. While **c)** and **d)** shows the inference performed by the baseline and QHELC models of Cascade R-CNN, respectively. At the bottom, **e)** and **f)** are the RetinaNet baseline and QHELC models, respectively.

The image in the following Figure A.8 also belongs to category number 10. This image contains 3 root fragments that represent a challenge for the models, in fact, none of the models in any architecture were able to detect them. The figure shows the inference performed by the baseline and best models of Faster R-CNN, RetinaNet, and FCOS. At the top of the figure, images **a)** and **b)** show the inference performed by the baseline and MTHT model of Faster R-CNN, respectively. The baseline model is unable of detecting these three root fragments, marked with the blue boxes, whereas MTHT detects one of the root fragments, the bigger one, which can be seen in the left blue box of image **b)**. The images showing the inference performed by RetinaNet baseline and GRMMCE models are **c)** and **d)**. No model of this architecture is able of detecting these root fragments, and some models detect some teeth with bounding boxes

smaller than the teeth. This is the case of the bounding box marked with the orange box in **c)**, whereas in **d)**, the QHELC model predicts a box that fits this tooth better. Finally, images in **e)** and **f)** present the inference of the FCOS baseline and QHELC models, respectively. These images show a similar scenario to images **c)** and **d)**, where none of the root fragments is predicted and one tooth in **e)** is predicted with a box smaller than the corresponding box in **f)**.

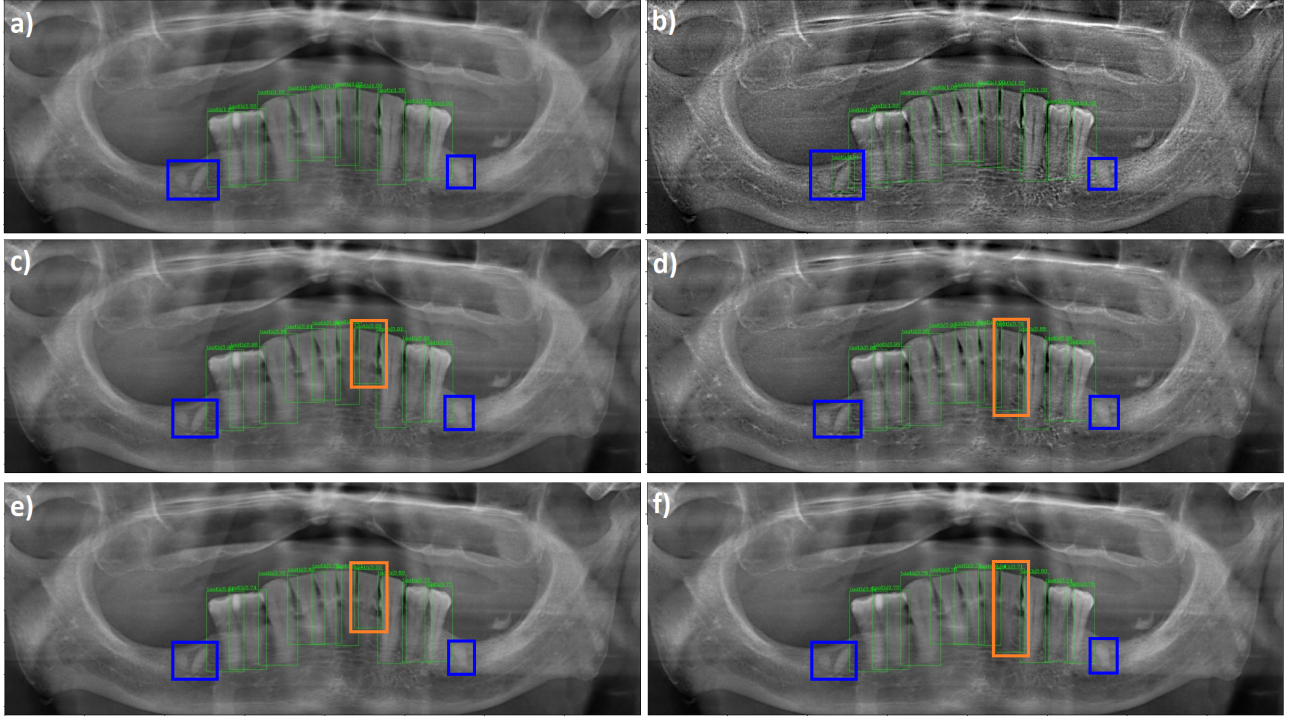


Figure A.8: Inference performed by six models on an image belonging to category 10. **a)** shows the inference performed by the baseline model and **b)** the MTHT model of Faster R-CNN. While **c)** and **d)** shows the inference performed by the baseline and GRMMCE models of RetinaNet, respectively. At the bottom, **e)** and **f)** are the images of the FCOS baseline and QHELC models, respectively.

The 0.75 threshold requires the models to have more accurate detections to achieved good performance. As was mentioned in the discussion in 4.2.1, the architectures' performance, except for Faster R-CNN, decreases when changing from 0.5 to 0.75. The architecture with the greatest change is FCOS, when most models are unable to detect teeth that are detected for all models in every architecture. This is the case of the image show in Figure A.9. The teeth of this image, belonging to category 10, are all detected with the eleven models of Faster R-CNN and Cascade R-CNN, and for most models of RetinaNet and YOLOv3, with at most 3 teeth not being detected in these two architectures. On the other hand, 12 teeth are undetected by the baseline model of FCOS when the threshold is changed to 0.75. These undetected teeth are marked with blue boxes in image **b)**, while in **a)**, the baseline model detect all teeth when the threshold used is 0.5.

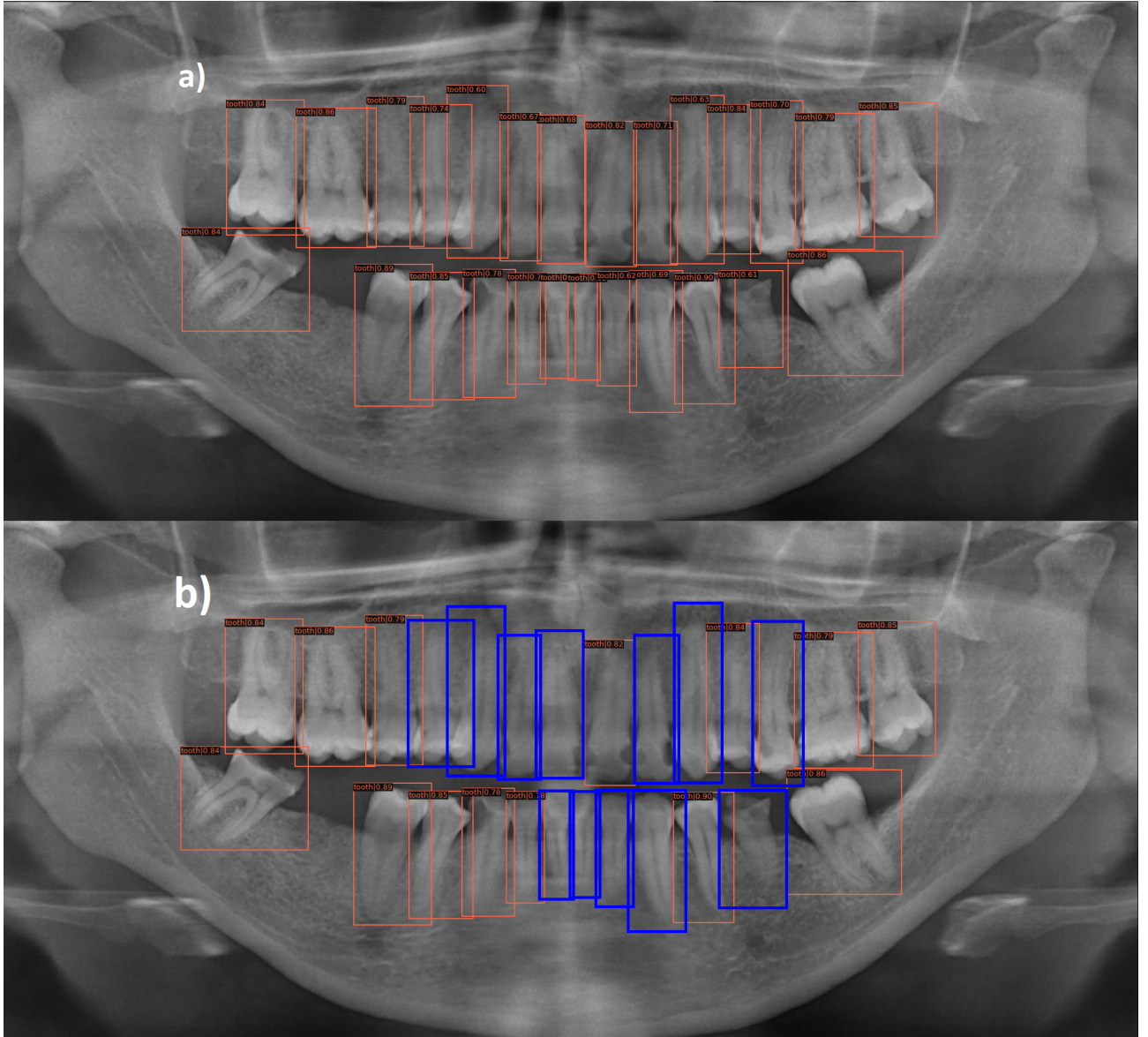


Figure A.9: Inference performed by the baseline models of FCOS with the threshold set to a) 0.5 and b) 0.75.

A.3 Presentation in International Conferences

This section presents two related works submitted to the *XL Congresso Nacional de Matemática Aplicada e Computacional* (CNMAC) 2021 and *18th World Congress of Medical and Health Informatics* (MedInfo) 2021. In CNMAC, a summary of the results of the correlation found between five IQM measured in the enhanced panoramic by four IET and the detection results of four models trained using Cascade R-CNN was submitted and accepted. The accepted work is attached next.

Correlation between image quality measurement of different image enhancement algorithms and Cascade R-CNN detection results applied to teeth detection in panoramic X-Ray images

Claudia L. Giardina¹, Jose Luis Vázquez Noguera, Horacio Legal-Ayala

Facultad Politécnica, Universidad Nacional Asunción, Central, Paraguay

Vicente R. Fretes

Faculdade de Odontologia de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto - SP 14040-904, Brazil. Facultad de Odontología, Universidad Nacional Asunción, Central, Paraguay

Diego Defazio

Centro de Diagnóstico y Tratamiento Periodontal, Asunción, Paraguay

Luis Salgueiro

Universitat Politècnica de Catalunya, Barcelona, España

Automatic teeth detection and segmentation in dental radiographs play a significant part in forensic identification and are considered the first step towards more complex systems for oral healthcare [3]. The reading and interpretation of X-ray images are time-consuming tasks, mainly performed by dentists where misdiagnosis can occur due to inexperience, fatigue or professional bias [3]. Numerous computer-assisted systems have been proposed to reduce the workload of professionals. Indeed, most of the published work based on Neural Networks (NN) focuses on tooth detection and segmentation, some of them achieving results above 90% [5]. In this work, we applied different image enhancement techniques as a pre-processing step to determine which features of the image are correlated to a better detection performance of teeth in panoramic images using an object detection algorithm. We used Cascade NN [2] for training 5 models, one trained with the original dataset and 4 trained with the dataset obtained after the application of different image enhancement techniques. The scheme can be seen in Figure 1 A). The images used to train the models belong to a publicly available dataset of 1500 images [6], of which 300 were annotated and validated by professionals dentists.

The 300 images were split into 192 for training, 48 for validation, and 60 for testing. The image enhancement techniques considered were the smoothing-edge preservation filters Anisotropic Diffusion [1], Bilateral Filtering [1], and the contrast enhancement algorithms BBHE [4] and QHELC [4]. Each model was trained for 8 epochs, with a learning rate of 0.009 and a batch-size of 2 in Google Colab. We evaluated the results considering the image processing outcome and the detection performance. For evaluating the image enhancement, we considered quality metrics like SSIM, PSNR, Edge Preservation Index (EPI), global contrast and Contrast Improvement Ratio (CIR) [4]. Whereas the detection performance was evaluated by measuring the accuracy, precision, and recall after considering as true detection those teeth that present bounding boxes with an IoU ≥ 0.5 . Table 1 shows the scores achieved by each model. Applying QHELC presents the best detection performance since it slightly increases the global contrast and preserves the main features of the

¹cgiardina@fpuna.edu.py

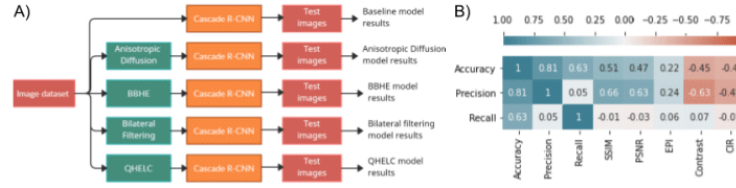


Figure 1: A) Experimental pipeline overview. B) Correlation heat map.

images, which translate to a high SSIM value. On the contrary, applying BBHE doubles the global contrast and raises the local contrast significantly, also producing a low PSNR, in detrimental of detection performance.

Table 1: IQM of the testing group. Accuracy, precision and recall of the models.

	SSIM	PSNR	(IQM)		CIR	Models' performance trained with IQM		
			EPI	Contrast		Accuracy (%)	Precision (%)	Recall (%)
Original				33.12		99.43	99.94	99.49
Anisotropic diffusion	0.963	41.003	0.423	32.71	0.742	99.37	99.89	99.49
BBHE	0.710	14.810	0.423	74.15	39.204	95.35	95.76	99.54
Bilateral Filtering	0.970	42.996	0.448	32.86	0.680	99.37	99.94	99.43
QHELC	0.993	36.906	0.913	36.33	0.188	99.49	99.94	99.54

We compute the cross-correlation between the performance of the detection models and the Image Quality Metrics (IQM) measured per image. Figure 1 B) shows that accuracy and precision have a negative relation with the global contrast and CIR, which means that significantly increasing the global and local contrast of the original images would negatively affect the detection performance metrics. On the other hand, detection performance would increase if the image presents less distortion and noise, as shown by the positive correlation with SSIM and PSNR.

References

- [1] Barash, D. Bilateral Filtering and Anisotropic Diffusion: Towards a Unified Viewpoint, *Scale-Space and Morphology in Computer Vision*, volume 2106, 2001. DOI: 10.1007/3-540-47778-0_24.
- [2] Cai, Z. and Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1483–1498, 2019. DOI: 10.1109/TPAMI.2019.2956516
- [3] Chen, H., Zhang, K., Lyu, P., Li, H., Zhang, L., Wu, J. and Lee, C. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, *Scientific Reports*, 9, 2019. DOI: 10.1038/s41598-019-40414-y.
- [4] Mello, J. C., Fretes, V. R., Adorno, C. G., Gariba, R. Vázquez, J. L., Legal-Ayala, H., Mello-Román, J. D., Escobar, R. D. and Facon, J., Panoramic Dental Radiography Image Enhancement Using Multiscale Mathematical Morphology, *Sensors*, 21, 2001. DOI: 10.3390/s21093110.
- [5] Schwendicke, F., Golla, T., Dreher, M. and Krois, J. Convolutional neural networks for dental image diagnostics: A scoping review, *Journal of Dentistry*, 91, 2019. DOI: 10.1016/j.jdent.2019.103226.
- [6] Silva, G., Oliveira, L., and Pithon, M. Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives, *Expert Systems with Applications*, 107:15–31, 2018. DOI: 10.1016/j.eswa.2018.04.001.

In MedInfo2021, the correlation between the detection results of 8 models and the IQM AMBE, SSIM, PSNR, Entropy, global contrast, CIR, and EPI were presented in a 5 pages article. The detection results of these models correspond to the five architectures using an IoU threshold of 0.5. The accepted article is attached below.

Correlation Between Quality Evaluation Metrics and Teeth Detection Results in Panoramic X-rays using Deep Learning

Claudia L. Giardina^a, Horacio Legal^a, José Luis Vázquez Noguera^a, Luis Salgueiro^b, Vicente R. Fretes^c,
Diego Defazio^d

^aFacultad Politécnica, Universidad Nacional de Asunción, Central, Paraguay

^bUniversitat Politècnica de Catalunya, Barcelona, España

^cFaculdade de Odontologia de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brasil

^cFacultad de Odontología, Universidad Nacional de Asunción, Asunción, Paraguay

^dCentro de Diagnóstico y Tratamiento Periodontal, Asunción, Paraguay

Abstract

Panoramic images are one of the most requested exams by dentists for allowing the visualization of the entire mouth. Interpreting X-ray images is a time-consuming task in which misdiagnoses can occur due to the inexperience or fatigue of professionals. In this work, we applied different image enhancement techniques as a pre-processing step to determine which image features correlate with improvements in teeth detection in panoramic images using deep learning architectures. We contrasted the performance of five object-detection architectures using 300 panoramic images of a public dataset. We evaluated the enhancement in the pre-processing step and the detection performance. Quality and detection metrics were considered, and the cross-correlation between them was computed for every object-detection method contemplated. We observe the dependence of the detection performance with some image enhancement techniques, especially those that introduce less noise and preserve the global contrast of the image.

Keywords:

Panoramic images, Deep Learning, Teeth detection

Introduction

Panoramic radiography is an extra-oral type of dental X-ray image that has become commonly used in dental practice and can be a valuable diagnostic tool in the dentist's armamentarium. It allows the visualization of the maxillary and mandibular teeth, surrounding bones, chin, and part of the cervical spine [1], complementing the clinical examination while exposing the patient to a small dose of radiation [1,2]. The interpretation of X-ray images is a time-consuming task mainly performed by dentists where misdiagnosis can occur due to the inexperience, fatigue, or bias of the professional [3]. Hence, a computer-assisted application is a helpful tool for assisting diagnosis and reducing the workload of the professionals.

Teeth detection can be done by Image-Processing and Deep Learning (DL) algorithms. In the field of dentistry, neural

networks are being mainly used for the detection and segmentation of teeth, with a few works focusing on the classification of dental pathologies as caries and periodontal bone loss [2]. Recent tooth detection and classification works achieve results comparable to professionals. Tuzoff et al. [4] achieved a sensitivity of 99.41% and a precision of 99.45% using Faster-RCNN [5] architecture for teeth detection in panoramic radiographs. Chen et al. [2] developed three post-processing techniques to improve the Faster-RCNN detections, achieving a precision and sensitivity of 98.8% and 98.5%, respectively. In [6] a fully convolutional neural network based on U-Net [7] is used to perform semantic segmentation on panoramic images, achieving a final accuracy of 95.06%. The success of a DL method not only depends on the configuration of the network (the architecture), but also, is related to other aspects like pre-processing, data augmentation, and hyper-parameter optimization. In this work, we explored some pre-processing, which has proven to be a differentiating factor between DL algorithms with the same architecture [8].

The rest of the article is organized as follows: in Methods we present the description of the dataset, the Image Enhancement Techniques (IET), the object detection architectures, and the experimental setup. In Results, we show the object detection performance of each architecture, as well as the correlation between the detection results and the Image Quality Metrics (IQM). Finally, we present the discussion and conclusion remarks.

Methods

Dataset

We use a public dataset of 1500 panoramic images generated by Gil Silva et al. [9] originally conceived for semantic segmentation. From the 1500 images, 300 images were selected and divided into 192 for training, 48 for validation, and 60 for testing. Bounding boxes were drawn for each tooth, implant, prosthesis, or dental root fragments as the ground-truth labels. These annotations were reviewed, edited, and validated by two experienced dentists.

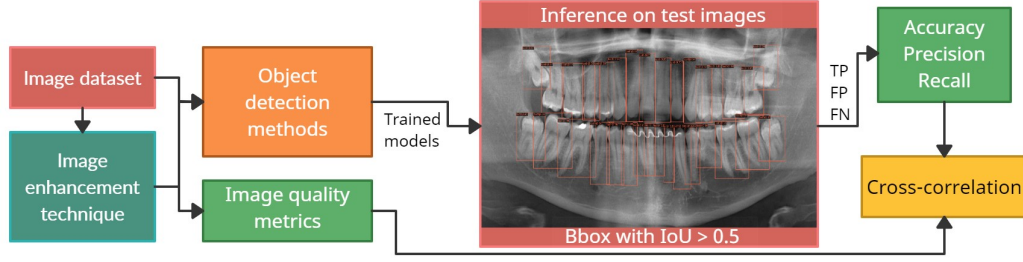


Figure 1– Diagram of steps followed in this work.

Table 1– Hyperparameters values.

Architecture	batch size	learning rate	weight decay	warmup iterations	steps
Faster R-CNN	2	0.01	0.0001	48	6
RetinaNet	2	0.008	0.0001	4	4
Cascade R-CNN	2	0.009	0.001	24	3, 5
FCOS	2	0.02	-	48	6, 13
YOLOv ₃	8	0.001	0.001	-	48, 96

Image enhancement techniques (IET)

The image enhancement techniques considered were the smoothing-edge preservation filters Anisotropic Diffusion (AD) [10] and Bilateral Filtering (BF) [11], in addition to the contrast enhancement algorithms BBHE [12], CLAHE [13], GRMMCE [14], MTHT [15], and QHELC [16]. Before applying the mentioned techniques, the images were first trimmed to exclude extra toothless regions by having as references the hard palate and lower mandibular contour. For evaluating the image enhancement, we considered quality metrics like Absolute Mean Brightness Error (AMBE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), standard deviation or global contrast (GC), Contrast Improvement Ratio (CIR) [17], Entropy, and Edge Preservation Index (EPI) [18].

AMBE measures if a method preserves the original brightness of an image. The better the brightness preservation, the lower AMBE would be. PSNR measures the distorting noise introduced in an image, where a higher image quality yields a higher PSNR. SSIM measures the structural similarity between the original image and the output image with a better value closer to 1. CIR measures the enhancement in the local contrast of the images. A higher value of CIR indicates a higher difference between the local contrast of the original image and the output image [17]. Entropy is used to quantify the information content of an image; a more detailed image has the higher entropy. EPI index quantifies the edge preservation after applying the enhancement to the image. EPI equals 1 indicates a similar value between the input and enhanced images [18].

Object detection architectures

We selected Faster R-CNN [5], Cascade R-CNN [19], RetinaNet [20], Fully Convolutional One-Stage Object Detection (FCOS) [21], and You Only Look Once (YOLOv3) [22] for being among the most widely used architectures. Pre-trained models were chosen from the Model Zoo of MMDetection [23], an open-source object detection toolbox based on Pytorch. Backbones used were ResNext [24] for Faster R-CNN, RetinaNet, and Cascade R-CNN, ResNet 101

for FCOS, and DarkNet-53 for YOLOv3. All pre-trained models were trained on the COCO dataset.

The best weight of each method, trained with the panoramic images, was chosen considering the highest mean Average Precision (mAP) with $IoU \geq 0.5$. The IoU is defined as:

$$IoU = \frac{Area_{DB} \cup Area_{GTB}}{Area_{DB} \cap Area_{GTB}}, \text{ where } Area_{DB} \text{ is the area of}$$

detected boxes and $Area_{GTB}$ is the area of ground truth boxes.

With the models trained and the weights selected, the performance of the models on the test set was evaluated considering as true detection a prediction with $IoU \geq 0.5$. We then measured the accuracy (ACC), precision (PRE), and recall, metrics that are calculated using:

$$Accuracy = \frac{TP}{TP + FP + FN}, \text{ Precision} = \frac{TP}{TP + FP}, \text{ and}$$

$$Recall = \frac{TP}{TP + FN}, \text{ where TP or True Positive is the number}$$

of teeth correctly detected, FP or False Positive is the number of teeth incorrectly detected, and FN or False Negative is the number of teeth that are not detected or are detected with and $IoU < 0.5$. Figure 1 shows a diagram of the steps followed in this work.

Experimental setup

Table 1 shows the parameters used for training the models, these values were set after hyperparameter tuning. The baseline model was trained using the original dataset, and seven more models were trained with the dataset obtained after the application of different IET. We chose the best weight for each architecture, after training for at most 100 epochs.

Results

Image quality metrics (IQM)

Table 2 shows the IQM for the pre-processed images as well as the Entropy and GC measured on the original dataset. We noticed that the IET that introduces less noise to the images was the bilateral filtering, achieving the highest PSNR among the techniques. While the change in the contrast made by BBHE, introduces the most distortion and noise to the images, having low values of SSIM and PSNR. CLAHE also introduces noise to the images, but unlike BBHE, CLAHE preserves the edges of the objects, having the closest value of EPI to one, while also raising the entropy of the images. Applying QHELC produces an image with an EPI close to one, which also has the highest SSIM value, indicating a low distortion introduced to the images.

Teeth detection

Table 3 shows the detection performance of the methods. Faster R-CNN results show that five models have a better accuracy than the baseline model, being the MTHT model the one with the best results, and the BF model the second best. On the contrary, QHELC model is the one with the lowest detection performance, having the worst accuracy and precision among the models. CLAHE is the model with the highest recall across all models.

In RetinaNet results, we noticed that GRMMCE, bilateral filtering, and QHELC models have better accuracy than the baseline. Being GRMMCE the one with the best accuracy, whereas BBHE has the lowest accuracy and precision among the models.

According to Cascade R-CNN results, QHELC model is the only model that has a slightly better performance than the

baseline model, while the model with the worse performance is again the BBHE model, having a precision 5.37% less than the precision of the QHELC model.

In FCOS results, the accuracy and recall have the same value since the precision is 100%. For this architecture, QHELC model achieves the higher accuracy, and it is followed by the baseline model, while the BBHE model presents the lowest accuracy among the models.

Finally, in YOLOv3 detection results, we noticed that there were no improvements introduced by any IET. Conversely, the BBHE model presents the lowest accuracy among the models.

Cross-correlation

We computed the correlation heat maps presented in Figure 2 using the IQM in Table 2 and the detection results of Table 3.

Figure 2-A shows the correlation heat map of Faster R-CNN. The relationships between the detection metrics and the IMQ are mostly neutral in this heat map. The relationship that slightly stands out is the negative relation of the precision with AMBE and GC. RetinaNet models' correlation heat map is shown in Figure 2-B. The map shows that the precision has a negative relationship of -0.27 with GC, and -0.22 with CIR. In Figure 2-C, the heat map of Cascade R-CNN models shows that the accuracy and the precision have a negative relation with AMBE, the GC, and CIR. While they have a positive relationship with both SSIM and PSNR. The heat map of FCOS in Figure 2-D does not have the precision since the value of this metric is 100% for every model. According to the heat maps in Figure 2-D and 2-E, the teeth detection results of FCOS and YOLOv3 models have a neutral relationship with the IQM considered.

Table 2 – Average of IQM measured on the testing group.

	AMBE	Entropy	SSIM	PSNR	EPI	Contrast	CIR
Original dataset		6.998				33.120	
AD	9.4E-07	6.982	0.963	41.003	0.423	32.715	0.742
BF	2.2E-06	6.987	0.970	42.996	0.448	32.863	0.680
BBHE	5.7E-02	6.829	0.710	14.810	0.423	74.154	39.204
CLAHE	5.9E-02	7.647	0.866	19.038	0.951	50.484	1.699
GRMMCE	5.2E-04	7.058	0.959	35.632	0.672	34.578	4.356
MTHT	6.0E-04	7.156	0.767	27.793	0.309	37.119	35.199
QHELC	1.1E-03	6.982	0.993	36.906	0.913	36.334	0.188

Table 3– Teeth detection performance of the methods.

	Faster R-CNN			RetinaNet			Cascade R-CNN			FCOS		YOLOv3		
	ACC	PRE	Recall	ACC	PRE	Recall	ACC	PRE	Recall	ACC	PRE	ACC	PRE	Recall
Baseline	98.92	99.49	99.43	98.81	99.43	99.37	99.43	99.94	99.49	97.83	100	96.29	99.94	96.34
AD	99.09	99.66	99.43	98.69	99.37	99.31	99.37	99.89	99.49	97.43	100	94.18	99.82	94.34
BF	99.20	99.71	99.49	98.92	99.43	99.49	99.37	99.94	99.43	97.54	100	95.03	99.88	95.14
BBHE	98.81	99.32	99.49	97.37	97.92	99.43	95.35	95.76	99.54	97.37	100	95.84	99.76	96.05
CLAHE	99.15	99.49	99.66	98.70	99.09	99.60	97.75	98.25	99.49	97.71	100	95.66	99.82	95.83

GRMMCE	99.03 99.54 99.49	98.98 99.43 99.54	99.37 99.83 99.54	97.54 100	95.95 99.82 96.11
MTHT	99.26 99.71 99.54	98.41 99.03 99.37	99.31 99.94 99.37	97.71 100	95.04 99.64 95.37
QHELC	98.64 99.31 99.31	98.92 99.37 99.54	99.49 99.94 99.54	98.11 100	95.83 99.94 95.88

Discussion

The detection performance of the models, shown in Table 3, indicates that applying an enhancement to the dataset does not always result in an improvement of the detection. This behavior is found in Cascade R-CNN, FCOS, and YOLOv3, where the baseline model has the second best or the best result. We can also observe that the QHELC model achieves the best results in Cascade R-CNN and FCOS, and the second-best result in RetinaNet. The BBHE model shows the weakest or one of the weakest performances in the five architectures. The heat map with the strongest relationships belongs to Cascade R-CNN in Figure 2-C. The negative relationship of the precision with GC and CIR, also presented in the RetinaNet heat map, means that the precision will be negatively affected by an increment of GC and local contrast. According to Table 2, BBHE is the enhancement technique that increases the most GC and the local contrast of the images, which results in the BBHE model having the lowest precision among the models of RetinaNet and Cascade R-CNN. We can see this in Figure 3, where in A we have the inference of the original model of RetinaNet, and in B we see the BBHE model. In Figure 3-A, a fragment of a tooth is missed, while in Figure 3-B the fragment is detected but there are two extra boxes, these boxes affect the precision of the model. The heat map of Figure 2-C also shows that the

accuracy and the precision have a positive relationship with SSIM and PSNR. This means that techniques that less noise and distortion introduced to the images, like QHELC, have a greater chance to have higher accuracy and precision.

The correlations found in this work might vary if the dataset has a higher number of images, or with greater diversity in the features of the images. For that reason, increasing the number of images in the dataset would elevate the importance of the results.

Conclusions

We evaluated the detection performance of several deep learning object-detection architectures with images modified by seven IET as a pre-processing step in the context of teeth detection. Some models show dependency on the modification of the image while others do not. We also correlated the features obtained by the IET with improvements of the detection performance showing that images with the highest values of SSIM and PSNR, i.e., those IET which preserves better the structure and introduce less noise to the images reach better performance and are suitable to be used for data augmentation, while incrementing CIR and GC will be in detrimental of the detection performance.

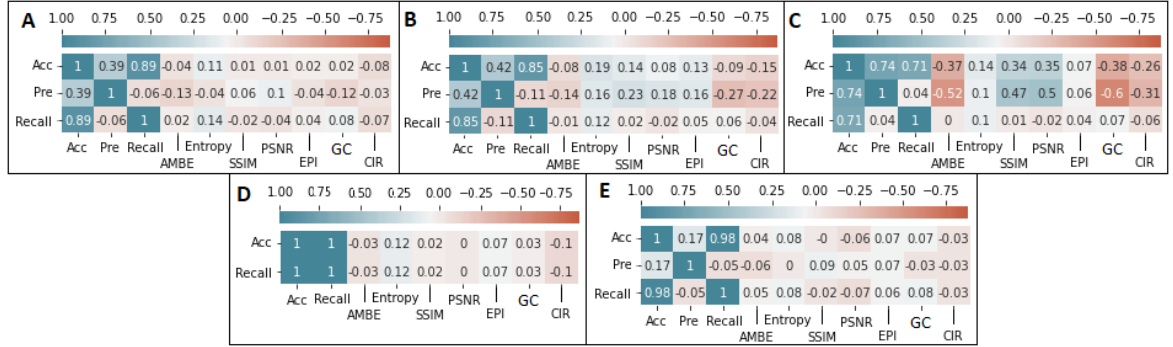


Figure 2— Correlation heat maps of A: Faster R-CNN, B: RetinaNet, C: Cascade R-CNN, D: FCOS, and E: YOLOv3.

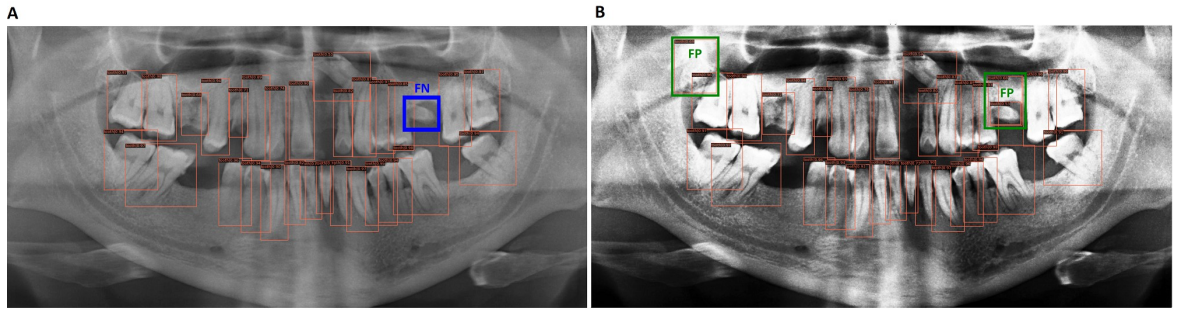


Figure 3— A: Baseline model inference missed a tooth marked with a blue box. B: BBHE model inference has two extra boxes marked with the green boxes.

References

- [1] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, Deep instance segmentation of teeth in panoramic x-ray images, In Conference on Graphics, Patterns and Images (SIBGRAPI) (2018), 400-407.
- [2] F. Schwendicke, T. Golla, M. Dreher, and J. Krois, Convolutional neural networks for dental image di-agnostics: A scoping review, *Journal of Dentistry* 91 (2019), 103226.
- [3] H. Chen, K. Zhang, P. Lyu, H. Li, L. Zhang, J. Wu, and C.-H. Lee, A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, *Scientific Reports* 9 (1) (2019).
- [4] D. Tuzoff, L. Tuzova, M. Bornstein, A. Krasnov, M. Kharchenko, S. Nikolenko, M. Sveshnikov, and G. Bednenko, Tooth detection and numbering in panoramic radiographs using convolutional neural networks, *Dentomaxillofacial Radiology* 48 (4) (2019), 20180051.
- [5] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6) (2015).
- [6] T. Koch, M. Perslev, C. Igel, and S. Brandt, Accurate segmentation of dental panoramic radiographs with u-nets, In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019), 15-19.
- [7] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, *LNCS 9351* (2015), 234-241.
- [8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, A survey on deep learning in medical image analysis, *Medical Image Analysis* 42 (2017), 60-88.
- [9] G. Silva, L. Oliveira, and M. Pithon, Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives, *Expert Systems with Applications* 107 (2018), 15-31.
- [10] G. Sapiro and D. Ringach, Anisotropic diffusion of multivalued images with applications to color filtering., *IEEE Transactions on Image Processing* 5 (1996), 1582-1586.
- [11] C. Tomasi and R. Manduchi, Bilateral filtering for gray and color images, In Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271) (1998), 839-846.
- [12] Y.-T. Kim, Contrast enhancement using brightness preserving bi-histogram equalization, *IEEE Transactions on Consumer Electronics* 43 (1997), 1-8.
- [13] K. Zuiderveld, Contrast Limited Adaptive Histogram Equalization, *Graphics Gems*, 4th Edition, Elsevier, 1994, 474-485.
- [14] J. Mello, R. Escobar, F. Martínez, J. Vázquez Noguera, H. Legal-Ayala, and D. Pinto-Roa, Medical image enhancement with brightness and detail preserving using multiscale top-hat transform by reconstruction, *Electronic Notes in Theoretical Computer Science* 349 (2020), 69-80.
- [15] J. Mello, H. Legal-Ayala, and J. Vázquez Noguera, Image color contrast enhancement using multiscale morphology, *Computational Interdisciplinary Sciences* 8 (3) (2017).
- [16] I. Brizuela, R. D. Medina Caballero, J. Silva, J. Roman, and J. Vazquez Noguera, Quadri-histogram equalization using cutoff limits based on the size of each histogram with preservation of average brightness, *Signal, Image and Video Processing* 13 (2) (2019).
- [17] Y.-P. Wang, Q. Wu, K. Castleman, and Z. Xiong, Chromosome image enhancement using multiscale differential operators, *IEEE Transactions on Medical Imaging* 22 (5) (2003), 685-693.
- [18] F. Sattar, L. Floreby, G. Salomonsson, and B. Lovstrom, Image enhancement based on a nonlinear multiscale method, *IEEE Transactions on Image Processing* 6 (6) (1997), 888-895.
- [19] Z. Cai and N. Vasconcelos, Cascade R-CNN: high quality object detection and instance segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2019), 1483-1498.
- [20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017), 1-1.
- [21] Z. Tian, C. Shen, H. Chen, and H. Tong, Fcos: Fully convolutional one-stage object detection, In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019), 9626-9635.
- [22] J. Redmon and A. Farhadi, Yolov3: An incremental improvement, (2018).
- [23] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, (2019).
- [24] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, Aggregated Residual Transformations for Deep Neural Networks, In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 5987-5995.

Address for correspondence

Claudia L. Giardina, email address cgiardina@fpuna.edu.py.