TCEIP: Text Condition Embedded Regression Network for Dental Implant Position Prediction

Xinquan Yang[1,2,3], Jinheng Xie[1,2,3,5], Xuguang Li[4], Xuechen Li[1,2,3], Xin Li[4],
Linlin Shen[1,2,3](✉), and Yongqiang Deng[4]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
yangxinquan2021@email.szu.edu.cn, llshen@szu.edu.cn

[2] AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University, Shenzhen, China

[3] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China

[4] Department of Stomatology, Shenzhen University General Hospital, Shenzhen, China

[5] National University of Singapore, Singapore , Singapore

Abstract. When deep neural network has been proposed to assist the dentist in designing the location of dental implant, most of them are targeting simple cases where only one missing tooth is available. As a result, literature works do not work well when there are multiple missing teeth and easily generate false predictions when the teeth are sparsely distributed. In this paper, we are trying to integrate a weak supervision text, the target region, to the implant position regression network, to address above issues. We propose a text condition embedded implant position regression network (TCEIP), to embed the text condition into the encoder-decoder framework for improvement of the regression performance. A cross-modal interaction that consists of cross-modal attention (CMA) and knowledge alignment module (KAM) is proposed to facilitate the interaction between features of images and texts. The CMA module performs a cross-attention between the image feature and the text condition, and the KAM mitigates the knowledge gap between the image feature and the image encoder of the CLIP. Extensive experiments on a dental implant dataset through five-fold cross-validation demonstrated that the proposed TCEIP achieves superior performance than existing methods.

Keywords: Dental Implant · Deep Learning · Text Guided Detection · Cross-Modal Interaction

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_31.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 317–326, 2023.
https://doi.org/10.1007/978-3-031-43987-2_31

1 Introduction

According to a systematic research study [2], periodontal disease is the world's 11th most prevalent oral condition, which potentially causes tooth loss in adults, especially the aged [8]. One of the most appropriate treatments for such a defect/dentition loss is prosthesis implanting, in which the surgical guide is usually used. However, dentists must load the Cone-beam computed tomography (CBCT) data into the surgical guide design software to estimate the implant position, which is tedious and inefficient. In contrast, deep learning-based methods show great potential to efficiently assist the dentist in locating the implant position [7].

Recently, deep learning-based methods have achieved great success in the task of implant position estimation. Kurt et al. [4] and Widiasri et al. [11] utilized the convolutional neural network (CNN) to locate the oral bone, e.g., the alveolar bone, maxillary sinus and jaw bone, which determines the implant position indirectly. Different from these implant depth measuring methods, Yang et al. [13] developed a transformer-based implant position regression network (ImplantFormer), which directly predicts the implant position on the 2D axial view of tooth crown images and projects the prediction results back to the tooth root by the space transform algorithm. However, these methods generally consider simple situations, in which only one missing tooth is available. When confronting some special cases, such as multiple missing teeth and sparse teeth disturbance in Fig. 1(a), the above methods may fail to determine the correct implant position. In contrast, clinically, dentists have a subjective expertise about where the implant should be planted, which motivates us that, additional indications or conditions from dentists may help predict an accurate implant position.

In recent years, great success has been witnessed in Vision-Language Pretraining (VLP). For example, Radford [9] proposed Contrastive Language-Image Pretraining (CLIP) to learn diverse visual concepts from 400 million image-text pairs automatically, which can be used for vision tasks like object detection [17] and segmentation [12]. In this paper, we found that CLIP has the ability to learn the position relationship among instances. We showcase examples in Fig. 1(b) that the image-text pair with the word 'left' get a higher matching score than others, as the position of baby is on the left of the billboard.

Motivated by the above observation in dental implant and the property of CLIP, in this paper, we integrate a text condition from the CLIP to assist the implant position regression. According to the natural distribution, we divide teeth regions into three categories in Fig. 1(c), i.e., left, middle, and right. Specifically, during training, one of the text prompts, i.e., 'right', 'middle', and 'left' is paired with the crown image as input, in which the text prompt works as a guidance or condition. The crown image is processed by an encoder-decoder network for final location regression. In addition, to facilitate the interaction between features in two modalities, a cross-modal interaction that consists of cross-modal attention (CMA) and knowledge alignment module (KAM), is devised. The CMA module fuses conditional information, i.e., text prompt, to the encoder-decoder. This brings additional indications or conditions from the dentist to help

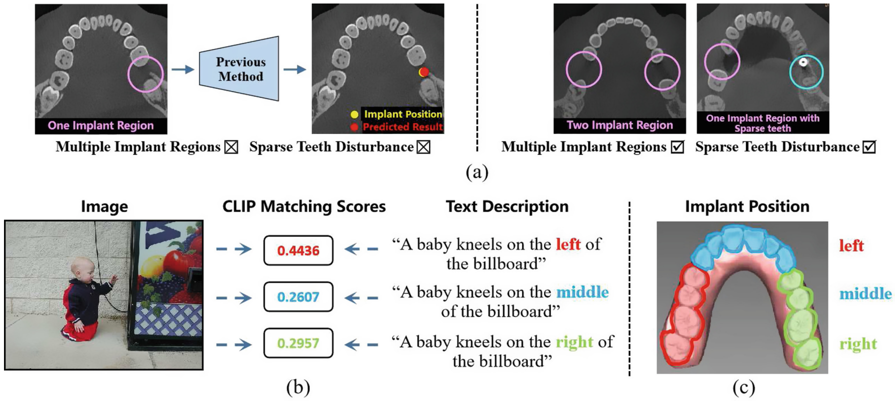
One Implant Region
Previous Method
Implant Position
Predicted Result
Multiple Implant Regions ☒
Sparse Teeth Disturbance ☒
Two Implant Region
One Implant Region with Sparse teeth
Multiple Implant Regions ☑
Sparse Teeth Disturbance ☑
(a)
Image
CLIP Matching Scores
Text Description
0.4436
0.2607
0.2957
"A baby kneels on the left of the billboard"
"A baby kneels on the middle of the billboard"
"A baby kneels on the right of the billboard"
(b)
Implant Position
left
middle
right
(c)

Fig. 1. (a) The 2D axial view of tooth crown images captured from different patients, where the pink and blue circles denote the implant and sparse teeth regions, respectively. (b) The matching score of the CLIP for a pair of image and text. (c) The division teeth region. (Color figure online)

the implant position regression. However, a knowledge gap may exist between our encoder-decoder and CLIP. To mitigate the problem, the KAM is proposed to distill the encoded-decoded features of crown images to the space of CLIP, which brings significant localization improvements. In inference, given an image, the dentist just simply gives a conditioning text like "let's implant a prosthesis on the left", the network will preferentially seek a suitable location on the left for implant prosthesis.

Main contributions of this paper can be summarized as follows: 1) To the best of our knowledge, the proposed TCEIP is the first text condition embedded implant position regression network that integrates a text embedding of CLIP to guide the prediction of implant position. (2) A cross-modal interaction that consists of a cross-modal attention (CMA) and knowledge alignment module (KAM) is devised to facilitate the interaction between features that representing image and text. (3) Extensive experiments on a dental implant dataset demonstrated the proposed TCEIP achieves superior performance than the existing methods, especially for patients with multiple missing teeth or sparse teeth.

2 Method

Given a tooth crown image with single or multiple implant regions, the proposed TCEIP aims to give a precise implant location conditioned by text indications from the dentist, i.e., a description of position like 'left', 'right', or 'middle'. An overview of TCEIP is presented in Fig. 2. It mainly consists of four parts: i) Encoder and Decoder, ii) Conditional Text Embedding, iii) Cross-Modal Interaction Module, and iv) Heatmap Regression Network. After obtaining the predicted coordinates of the implant at the tooth crown, we adopt the space transformation algorithm [13] to fit a centerline of implant to project the coordinates to the tooth root, where the real implant location can be acquired. Next, we will introduce these modules in detail.

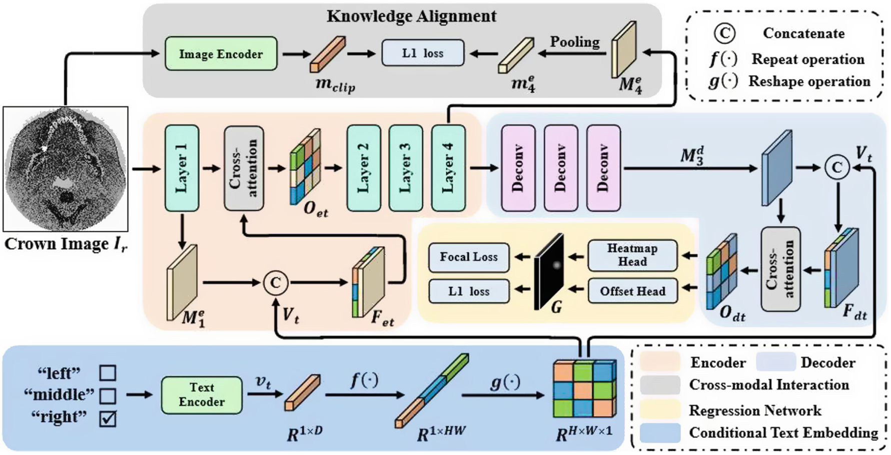

Fig. 2. The network architecture of the proposed prediction network.

2.1 Encoder and Decoder

We employ the widely used ResNet [3] as the encoder of TCEIP. It mainly consists of four layers and each layer contains multiple residual blocks. Given a tooth crown image $\mathbf{I}_r$, a set of feature maps, i.e., $\{\mathbf{M}_1^e, \mathbf{M}_2^e, \mathbf{M}_3^e, \mathbf{M}_4^e\}$, can be accordingly extracted by the ResNet layers. Each feature map has a spatial and channel dimension. To ensure fine-grained heatmap regression, three deconvolution layers are adopted as the Decoder to recover high-resolution features. It consecutively upsamples feature map $\mathbf{M}_4^e$ as high-resolution feature representations, in which a set of recovered features $\{\mathbf{M}_1^d, \mathbf{M}_2^d, \mathbf{M}_3^d\}$ can be extracted. Feature maps $\mathbf{M}_1^e$, $\mathbf{M}_4^e$ and $\mathbf{M}_3^d$ will be further employed in the proposed modules, where $\mathbf{M}_1^e$ and $\mathbf{M}_3^d$ have the same spatial dimension $\mathbb{R}^{128\times 128\times C}$ and $\mathbf{M}_4^e \in \mathbb{R}^{16\times 16\times \hat{C}}$.

2.2 Conditional Text Embedding

To integrate the text condition provided by a dentist, we utilize the CLIP to extract the text embedding. Specifically, additional input of text, e.g., 'left', 'middle', or 'right', is processed by the CLIP Text Encoder to obtain a conditional text embedding $\mathbf{v}_t \in \mathbb{R}^{1\times D}$. As shown in Fig. 2, to interact with the image features from ResNet layers, a series of transformation $f(\cdot)$ and $g(\cdot)$ over $\mathbf{v}_t$ are performed as follow:

$$\mathbf{V}_t = g(f(\mathbf{v}_t)) \in \mathbb{R}^{H\times W\times 1}, \tag{1}$$

where $f(\cdot)$ repeats text embedding $\mathbf{v}_t$ from $\mathbb{R}^{1\times D}$ to $\mathbb{R}^{1\times HW}$ and $g(\cdot)$ then reshapes it to $\mathbb{R}^{H\times W\times 1}$. This operation ensures better interaction between image and text in the same feature space.

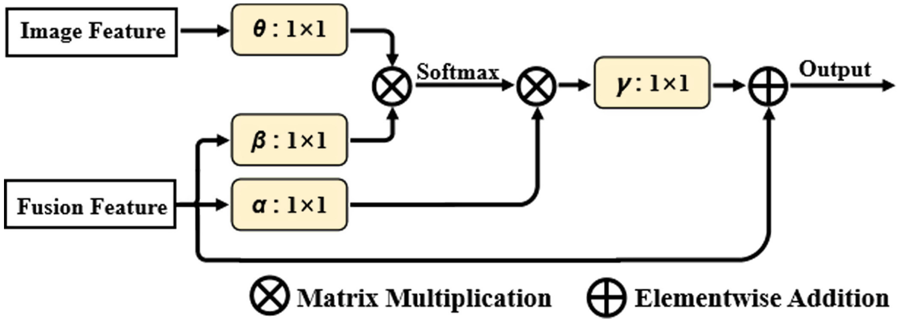

Fig. 3. The architecture of the proposed cross-modal attention module.

2.3 Cross-Modal Interaction

High-resolution features from the aforementioned decoder can be directly used to regress the implant position. However, it cannot work well in situations of multiple teeth loss or sparse teeth disturbance. In addition, although we have extracted the conditional text embedding from the CLIP to assist the network regression, there exists a big difference with the feature of encoder-decoder in the feature space. To tackle these issues, we propose cross-modal interaction, including i) Cross-Modal Attention and ii) Knowledge Alignment module, to integrate the text condition provided by the dentist.

Cross-Modal Attention Module. To enable the transformed text embedding $\mathbf{V}_t$ better interact with intermediate features of the encoder and decoder, we design and plug a cross-modal attention (CMA) module into the shallow layers of the encoder and the final deconvolution layer. The architecture of CMA is illustrated in Fig. 3. Specifically, the CMA module creates cross-attention between image features $\mathbf{M}_1^e$ and fusion feature $\mathbf{F}_{et} = [\mathbf{M}_1^e|\mathbf{V}_t]$ in the encoder, and image features $\mathbf{M}_3^d$ and fusion feature $\mathbf{F}_{dt} = [\mathbf{M}_3^d|\mathbf{V}_t]$ in the decoder, where $\mathbf{F}_{et}, \mathbf{F}_{dt} \in \mathbb{R}^{H\times W\times(C+1)}$. The CMA module can be formulated as follows:

$$\mathbf{O} = \gamma(\mathrm{Softmax}(\theta(\mathbf{M})\beta(\mathbf{F}))\alpha(\mathbf{F})) + \mathbf{F}, \tag{2}$$

where four independent 1×1 convolutions α, β, θ, and γ are used to map image and fusion features to the space for cross-modal attention. At first, $\mathbf{M}$ and $\mathbf{F}$ are passed into $\theta(\cdot)$, $\beta(\cdot)$ and $\alpha(\cdot)$ for channel transformation, respectively. Following the transformed feature, $\mathbf{M}_\theta$ and $\mathbf{F}_\beta$ perform multiplication via a Softmax activation function to take a cross-attention with $\mathbf{F}_\alpha$. In the end, the output feature of the cross-attention via $\gamma(\cdot)$ for feature smoothing is added with $\mathbf{F}$. Given the above operations, the cross-modal features $\mathbf{O}_{et}$ and $\mathbf{O}_{dt}$ are obtained and passed to the next layer.

Knowledge Alignment Module. The above operations only consider the interaction between features in two modalities. A problem is that text embeddings from pre-trained text encoder of CLIP are not well aligned with the image features initialized by ImageNet pre-training. This knowledge shift potentially weakens the proposed cross-modal interaction to assist the prediction of implant position. To mitigate this problem, we propose the knowledge alignment module (KAM) to gradually align image features to the feature space of pre-trained CLIP. Motivated by knowledge distillation [10], we formulate the proposed knowledge alignment as follows:

$$\mathcal{L}_{align} = |\mathbf{m}_4^e - \mathbf{m}_{clip}|, \tag{3}$$

where $\mathbf{m}_4^e \in \mathbb{R}^{1\times D}$ is the transformed feature of $\mathbf{M}_4^e$ after attention pooling operation [1] and dimension reduction with convolution. $\mathbf{m}_{clip} \in \mathbb{R}^{1\times D}$ is the image embedding extracted by the CLIP Image Encoder. Using this criteria, the encoder of TCEIP approximates the CLIP image encoder and consequently aligns the image features of the encoder with the CLIP text embeddings.

2.4 Heatmap Regression Network

The heatmap regression network is used for locating the implant position, which consists of the heatmap and the offset head. The output of the heatmap head is the center localization of implant position, which is formed as a heatmap $\mathbf{G} \in [0,1]^{H\times W}$. Following [5], given coordinate of the ground truth implant location $(\tilde{t}_x, \tilde{t}_y)$, we apply a 2D Gaussian kernel to get the target heatmap:

$$\mathbf{G}_{xy} = \exp(-\frac{(x-\tilde{t}_x)^2 + (y-\tilde{t}_y)^2}{2\sigma^2}) \tag{4}$$

where σ is an object size-adaptive standard deviation. The predicted heatmap is optimized by the focal loss [6]:

$$\mathcal{L}_h = \frac{-1}{N}\sum_{xy}\begin{cases} (1-\hat{\mathbf{G}}_{xy})^\lambda \log(\hat{\mathbf{G}}_{xy}) & \text{if}\, \mathbf{G}_{xy} = 1 \\ (1-\hat{\mathbf{G}}_{xy})^\varphi \log(\hat{\mathbf{G}}_{xy})^\lambda \log(1-\hat{\mathbf{G}}_{xy}) & \text{otherwise} \end{cases} \tag{5}$$

where λ and φ are the hyper-parameters of the focal loss, $\hat{\mathbf{G}}$ is the predicted heatmap and N is the number of implant annotation in image. The offset head computes the discretization error caused by the downsampling operation, which is used to further refine the predicted location. The local offset loss $\mathcal{L}_o$ is optimized by the L1 loss. The overall training loss of network is:

$$\mathcal{L} = \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_{align} \tag{6}$$

2.5 Coordinate Projection

The output of TCEIP is the coordinate of implant at the tooth crown. To obtain the real implant location at the tooth root, we fit a centerline of implant using the predicted implant position of TCEIP and then extend the centerline to the root area, which is identical as [13]. By this means, the intersections of implant centerline with 2D slices of root image, i.e. the implant position at the tooth root area, can be obtained.

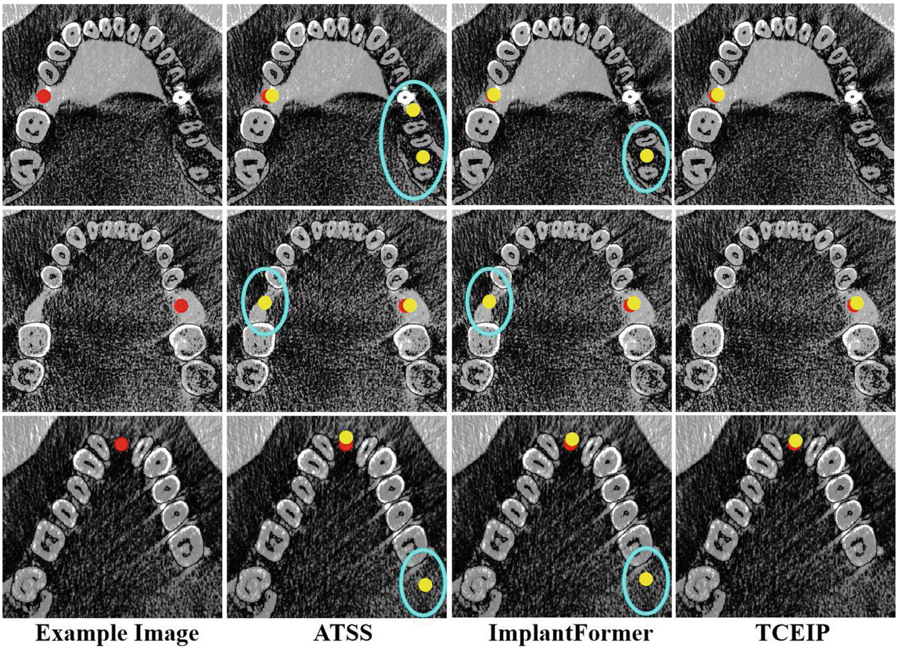

Fig. 4. Visual comparison of the predicted results with different detectors. The yellow and red circles represent the predicted implant position and ground-truth position, respectively. The blue ellipses denote false positive detections. (Color figure online)

3 Experiments and Results

3.1 Dataset and Implementation Details

The dental implant dataset was collected by [13], which contains 3045 2D slices of tooth crown images. The implant position annotations are annotated by three experienced dentists. The input image size of network is set as 512 × 512. We use a batch size of 8, Adam optimizer and a learning rate of 0.001 for the network training. Total training epochs is 80 and the learning rate is divided by 10 when epoch $= \{40, 60\}$. The same data augmentation methods in [13] was employed.

3.2 Performance Analysis

We use the same evaluation criteria in [13], i.e., average precision (AP) to evaluate the performance of our network. As high accurate position prediction is required in clinical practice, the IOU threshold is set as 0.75. Five-fold cross-validation was performed for all our experiments.

Table 1. The ablation experiments of each components in TCEIP.

Network	KAM	Text Condition	Feature Fusion	CMA	AP_{75}%
TCEIP					10.9 ± 0.2457
		✓			14.6 ± 0.4151
	✓	✓			15.7 ± 0.3524
	✓	✓		✓	16.5 ± 0.3891
	✓	✓	✓		17.1 ± 0.2958
	✓	✓	✓	✓	**17.8 ± 0.3956**

Table 2. Comparison of the proposed method with other mainstream detectors.

Methods	Network	Backbone	AP_{75}%
Transformer-based	ImplantFormer	ViT-Base-ResNet-50	13.7 ± 0.2045
	Deformable DETR [19]	ResNet-50	12.8 ± 0.1417
CNN-based	CenterNet [18]		10.9 ± 0.2457
	ATSS [16]		12.1 ± 0.2694
	VFNet [15]		11.8 ± 0.8734
	RepPoints [14]		11.2 ± 0.1858
	TCEIP		**17.8 ± 0.3956**

Ablation Studies. To evaluate the effectiveness of the proposed network, we conduct ablation experiments to investigate the effect of each component in Table 1. We can observe from the second row of the table that the introduction of text condition improves the performance by 3.7%, demonstrating the validity of using text condition to assist the implant position prediction. When combining both text condition and KAM, the improvement reaches 4.8%. As shown in the table's last three rows, both feature fusion operation and CAM improve AP value by 1.4% and 0.8%, respectively. When combining all these components, the improvement reaches 6.9%.

Comparison to the Mainstream Detectors. To demonstrate the superior performance of the proposed TCEIP, we compare the AP value with the mainstream detectors in Table 2. Only the anchor-free detector is used for comparison, due to the reason that no useful texture is available around the center of the implant. As the teeth are missing, the anchor-based detectors can not regress the implant position successfully. From the table we can observe that, the transformer-based methods perform better than the CNN-based networks (e.g., ImplantFormer achieved 13.7% AP, which is 1.6% higher than the best-performed anchor-free network - ATSS). The proposed TCEIP achieves the best AP value - 17.8%, among all benchmarks, which surpasses the ImplantFormer with a large gap. The experimental results proved the effectiveness of our method.

In Fig. 4, we choose two best-performed detectors from the CNN-based (e.g., ATSS) and transformer-based (e.g., ImplantFormer) methods for visual comparison, to further demonstrate the superiority of TCEIP in the implant position prediction. The first row of the figure is a patient with sparse teeth, and the second and third rows are a patient with two missing teeth. We can observe from the figure that both the ATSS and ImplantFormer generate false positive detection, except for the TCEIP. Moreover, the implant position predicted by the TCEIP is more accurate. These visual results demonstrated the effectiveness of using text condition to assist the implant position prediction.

4 Conclusions

In this paper, we introduce TCEIP, a text condition embedded implant position regression network, which integrate additional condition from the CLIP to guide the prediction of implant position. A cross-modal attention (CMA) and knowledge alignment module (KAM) is devised to facilitate the interaction between features in two modalities. Extensive experiments on a dental implant dataset through five-fold cross-validation demonstrated that the proposed TCEIP achieves superior performance than the existing methods.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 82261138629; Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688 and 2021A1515220072; Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ2022053110141 2030 and JCYJ20220530155811025.

References

1. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. Elani, H., Starr, J., Da Silva, J., Gallucci, G.: Trends in dental implant use in the US, 1999–2016, and projections to 2026. J. Dent. Res. **97**(13), 1424–1430 (2018)

3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
4. Kurt Bayrakdar, S., et al.: A deep learning approach for dental implant planning in cone-beam computed tomography images. BMC Med. Imaging **21**(1), 86 (2021)
5. Law, H., Deng, J.: Cornernet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
6. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
7. Liu, Y., Chen, Z.C., Chu, C.H., Deng, F.L.: Transfer learning via artificial intelligence for guiding implant placement in the posterior mandible: an in vitro study (2021)
8. Nazir, M., Al-Ansari, A., Al-Khalifa, K., Alhareky, M., Gaffar, B., Almas, K.: Global prevalence of periodontal disease and lack of its surveillance. Sci. World J. 2020 (2020)
9. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
10. Rasheed, H., Maaz, M., Khattak, M.U., Khan, S., Khan, F.S.: Bridging the gap between object and image-level representations for open-vocabulary detection. arXiv preprint arXiv:2207.03482 (2022)
11. Widiasri, M., et al.: Dental-yolo: alveolar bone and mandibular canal detection on cone beam computed tomography images for dental implant planning. IEEE Access **10**, 101483–101494 (2022)
12. Xie, J., Hou, X., Ye, K., Shen, L.: Clims: cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4483–4492 (2022)
13. Yang, X., et al.: ImplantFormer: vision transformer based implant position regression using dental CBCT data. arXiv preprint arXiv:2210.16467 (2022)
14. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: RepPoints: point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9657–9666 (2019)
15. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: VarifocalNet: an IoU-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8514–8523 (2021)
16. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9759–9768 (2020)
17. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 1363, pp. 350–368. Springer, Cham (2022)
18. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
19. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)