Simplify Implant Depth Prediction as Video Grounding: A Texture Perceive Implant Depth Prediction Network

Xinquan Yang[1,2,3], Xuguang Li[4], Xiaoling Luo[1,2,3], Leilei Zeng[1,2,3], Yudi Zhang[1,2,3], Linlin Shen[1,2,3†], Yongqiang Deng[4†],

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University, Shenzhen, China
[3] Guangdong Provincial Key Laboratory, Shenzhen, China
[4] Department of Stomatology, Shenzhen University General Hospital, Shenzhen, China
xinquanyang99@gmail.com, llshen@szu.edu.cn, qiangyongdeng@sina.com

Abstract. Surgical guide plate is an important tool for the dental implant surgery. However, the design process heavily relies on the dentist to manually simulate the implant angle and depth. When deep neural networks have been applied to assist the dentist quickly locates the implant position, most of them are not able to determine the implant depth. Inspired by the video grounding task which localizes the starting and ending time of the target video segment, in this paper, we simplify the implant depth prediction as video grounding and develop a Texture Perceive Implant Depth Prediction Network (TPNet), which enables us to directly output the implant depth without complex measurements of oral bone. TPNet consists of an implant region detector (IRD) and an implant depth prediction network (IDPNet). IRD is an object detector designed to crop the candidate implant volume from the CBCT, which greatly saves the computation resource. IDPNet takes the cropped CBCT data to predict the implant depth. A Texture Perceive Loss (TPL) is devised to enable the encoder of IDPNet to perceive the texture variation among slices. Extensive experiments on a large dental implant dataset demonstrated that the proposed TPNet achieves superior performance than the existing methods.

Keywords: Dental Implant · Deep Learning · Implant Depth Prediction

1 Introduction

Tooth loss is a common problem among middle-aged and elderly people, and artificial dental implantation is one of the most appropriate treatment methods. In clinical, to ensure implant accuracy and accelerate the implantation process,

---

[†] Corresponding Author

(a) The task of video grounding.
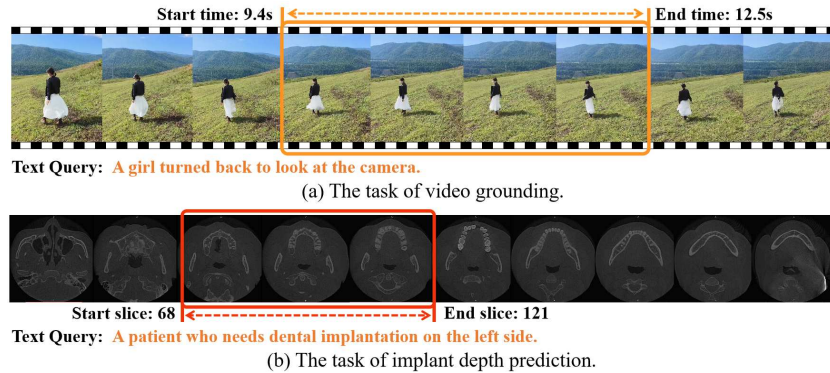
(b) The task of implant depth prediction.

Fig. 1. Comparison of the video grounding task and implant depth prediction task.

dentists usually use the surgical guide plate during surgery. However, the design of the surgical guide plate require to manually simulate the implant position (e.g., implantation angle and depth) by loading the Cone-beam computed tomography (CBCT) data into the design software, which is labour-intensive and time-consuming. With the development of deep learning, using artificial intelligence methods to speed up such process is promising.

Recently, a number of literature works have been proposed to assist the dentist quickly locating the implant position. ImplantFormer [12] proposed to predict the implant position using the 2D axial view of tooth crown images and projects the prediction results back to the tooth root by the space transform algorithm. Following this paradigm, a series of improved works, TSIRP [11], TCEIP [14], and TCSloT [13] are proposed to improve the accuracy of implant position prediction. Although these methods demonstrate excellent performance, they are semi-automated as the dentist are required to manually set the implant depth, which is inefficient for the clinic application. To solve this problem, some researchers try to detect the alveolar bone and mandibular canal based on the sagittal view of CBCT to determine the height and width of the alveolar bone, which predicts a approximate implant depth [9]. Kurt et al. [3] utilised multiple pre-trained convolutional networks to segment the teeth and jaws to locate the missing tooth and determine the implant depth by measuring oral tissues (e.g., mandibular canal, maxillary sinus, and jaw bone edge). However, these methods are too complicated for the clinic application and can not provide a precise implant depth.

Video grounding is an important yet challenging task in computer vision, which requires the machine to watch a video and localize the starting and ending time of the target video segment that corresponds to the given query [15]. In this paper, we found that the task of implant depth prediction is similar to video grounding, if we consider the 3D CBCT data as a video and the beginning and ending slice of implant as the starting and ending time of the target video segment, as shown in Fig. 1. By this means, the implant depth can be directly
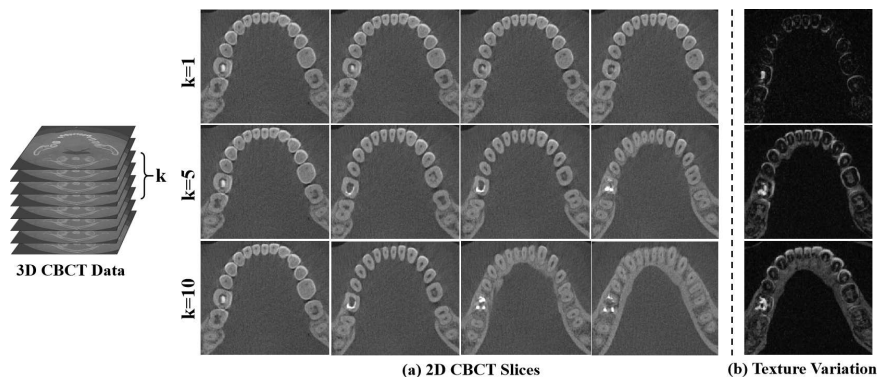
Fig. 2. (a) Comparison of 2D slices with different sampling intervals, k represents the sampling interval. (b) Texture variation computed from the 2D slices.

determined during inference, without requiring additional measurements of oral tissues.

Motivated by the above observation, in this paper, we develop a Texture Perceive Implant Depth Prediction Network (TPNet), which consists of an implant region detector (IRD) and an implant depth prediction network (IDPNet). IRD is an object detector designed to locate the implant region. We crop a sub-volume from the CBCT data according to the detection result of IRD. By this means, the irrelevant information of CBCT for implantation will be removed and the input data size can be substantially reduced. Then, the sub-volume is taken as the input of IDPNet. IDPNet is devised to regress the precise implant depth, which is a single encoder-decoder regression network. As the determination of implant depth relies on the texture of neighboring teeth, a Texture Perceive Loss (TPL) is proposed to enable the encoder to perceive the texture variation among slices, which greatly helps the IDPNet predicts more accurate implant depth.

Main contributions of this paper can be summarized as follows:1) To the best of our knowledge, we are the first one to model the task of implant depth prediction as video grounding, which enables us to directly predict the implant depth, without requiring additional computation. 2) An implant region detector (IRD) is introduced to remove the irrelevant information of CBCT, which sharply reduces the input data size and save computational costs. 3) A Texture Perceive Loss (TPL) is devised to enable the encoder to capture more fine-grained features by perceiving the texture variation among slices. 4) Extensive experiments on a large dental implant dataset demonstrated the proposed TPNet achieves superior performance than the existing methods.

2   Method

Given a patient's CBCT data, TPNet aims to predict a precise implant depth, i.e., the index of start slice and end slice. An overview of TPNet is presented
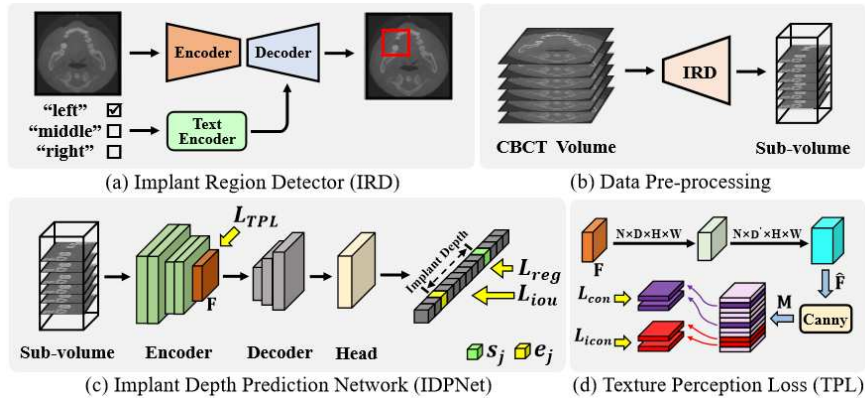
(a) Implant Region Detector (IRD)

(b) Data Pre-processing

(c) Implant Depth Prediction Network (IDPNet)

(d) Texture Perception Loss (TPL)

Fig. 3. The architecture of the proposed texture perceive implant depth prediction framework.

in Fig. 3. It mainly consists of two parts: i) Implant Region Detector (IRD), ii) Implant Depth Prediction Network (IDPNet). IRD first locates the implant region to crop a sub-volume from the CBCT data, and the IDPNet takes the sub-volume as input to predict the implant depth. Next, we will introduce them in detail.

2.1   Implant Region Detector

The CBCT data contains complete information about the maxillary and mandibular bones, in which the maxillary and mandibular sinuses are irrelevant for predicting implant depth. Therefore, it is computationally intensive to train IDPNet using the whole CBCT data. Using an IRD to detect the implant region and crop a sub-volume according to the detection result can significantly reduce the CBCT size. Inspired by the previous methods, we introduce a text guided implant position prediction network - TCEIP [14] as IRD. TCEIP integrates the direction embedding form CLIP [7] to guide the prediction model to locate the implant position, thus perform well in the patient who have multiple missing teeth. Consider clinical practicality, in this paper, we design a lightweight TCEIP as IRD, in which the knowledge alignment module and cross-modal attention module are discarded.

The architecture of IRD is shown in Fig. 3(a), which consists of an encoder, a decoder and a text encoder of CLIP. Firstly, ResNet-50 [2] is used as the encoder for feature extraction, and three deconvolution layers are adopted as the decoder to recover high-resolution features. Then, we extract the conditional text embedding from CLIP by inputting an additional text, e.g., 'left', 'middle', or 'right' into the CLIP text encoder. In the end, the conditional text embedding is concatenated with the last feature map of decoder to generate a gaussian heatmap for implant position regression. We follow TCEIP to use the focal loss [4] and L1 loss for supervision. After obtaining the implant position, we

generate a 256×256 box centered on the implant position as the implant region, which ensures that the texture of neighboring teeth is included. We then crop a $352 \times 256 \times 256$ sub-volume from the original CBCT along the axial view, and the sub-volume will be taken as the input data of IDPNet.

2.2   Implant Depth Prediction Network

The architecture of IDPNet is shown in Fig. 3(c). It mainly consists of an encoder, a decoder and a regression head. Firstly, the encoder extracts features from the sub-volume and the middle feature map $\mathbf{F} \in \mathbb{R}^{N \times C \times D \times H \times W}$ will be extracted. We use the proposed Texture Perception Loss (TPL) to supervise $\mathbf{F}$ that enables the encoder can capture more fine-grained features by perceiving the texture variation among slices. Then, the decoder recovers the encoder features to high-resolution and the regression head predicts the implant depth. Next, we will introduce them in detail.

Encoder and Decoder. We employ the widely used resblock to construct the encoder of IDPNet. Specifically, the encoder consists of two 3D Resblocks [8] and two 2D Resblocks [2]. The architecture of encoder and decoder are shown in the Fig. 3(c). The 3D resblock first takes the sub-volume as input and learns context information among slices. Then, these temporal features are fed into the 2D resblock to learn the texture feature in different slices. The output of encoder is a feature map $\mathbf{F} \in \mathbb{R}^{N \times C \times D \times H \times W}$. Considering that the regression of implant depth heavily relies on clearly neighboring tooth texture, which requires high-resolution feature representations. Hence, we adopt three deconvolution layers as decoder to consecutively upsamples the feature map.

Texture Perceive Loss. Clinically, dentists determine the implant depth according to the texture of neighboring teeth, e.g., the bottom of the implant does not exceed the root of neighboring teeth. Therefore, IDPNet should possess the ability to perceive the texture variation among slices. In Fig. 2, we visualize 2D slices sampled with different sampling and compute the texture variation among these slices by standard deviation. We can observe from the figure that the larger the sampling interval, the more obvious such texture variations. This observation indicates that the neighboring 2D slices have a similar feature, while the distant slices have a big difference in features. Drawing inspiration from this observation, in this paper, we propose a Texture Perceive Loss (TPL), which assists the encoder learns more robust features.

The details of TPL is given in the Fig. 3(d). Specifically, we first reduce the channel C of $\mathbf{F}$ to 1 and reshape the channel D to $D^{'}$, to restore the information of channel D. By this means, the pre-processed $\hat{\mathbf{F}} \in \mathbb{R}^{N \times D^{'} \times H \times W}$ is obtained. Then, we apply the Canny operator for $\hat{\mathbf{F}}$ to extract textures along the channel $D^{'}$. After obtaining a series of texture matrix $\mathbf{M} \in \mathbb{R}^{N \times D^{'} \times H \times W}$, we perform the consistency loss $\mathcal{L}_{con}$ for the neighboring matrix to close these features, and

the inconsistency loss $\mathcal{L}_{icon}$ for the distant matrix to distinguish these features. $\mathcal{L}_{TPL}$ is the summation of $\mathcal{L}_{con}$ and $\mathcal{L}_{icon}$, and we implement $\mathcal{L}_{con}$ and $\mathcal{L}_{icon}$ by L2 loss. In our implementation, we set the sampling interval k of the distant matrix as 10.

Regression Head. The regression head is designed to predict the implant depth, which is implemented by two convolutions followed by the activation function, i.e., ReLU. We use the L1 loss to optimize the regression head:

$$\mathcal{L}_{reg} = \sum_{j=1}^{N_p} |y_j - \hat{y}_j|, \tag{1}$$

where j is the patient index in a mini-batch and N_p is the total number of patient. $y_j = (s_j, e_j)$ and $\hat{y}_j = (\hat{s}_j, \hat{e}_j)$ is the predicted and ground-truth index of start and end implant slice, respectively.

As discussed in previous sections, we model implant depth prediction as the task of video grounding. Therefore, we follow the video grounding to introduce the temporal iou loss [18] to supervise the regression head:

$$\mathcal{L}_{tiou} = 1 - \frac{\hat{y}_j \cap y_j}{\hat{y}_j \cup y_j}. \tag{2}$$

The rationale of $\mathcal{L}_{tiou}$ is to maximize the overlapping between the predicted slice index and its ground truth. The overall training loss of IDPNet is:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \mathcal{L}_{tiou} + \mathcal{L}_{TPL} \tag{3}$$

3   Experiment

3.1   Dataset and Implementation Details

We evaluate the proposed TPNet on a collected dental implant dataset. The dataset contains 400 patients, in which 80% data were selected as the training set and the remaining 20% as the testing set. All the CBCT data were captured using the KaVo 3D eXami machine, manufactured by Imagine Sciences International LLC. The original CBCT size is $432 \times 776 \times 776$. For the traing of IRD, we follow TCEIP to use the 2D slice of CBCT and resize them to 512×512 for training and inference. After the data pre-processing of IRD, the size of CBCT data for each patient is reduced to $352 \times 256 \times 256$.

For the training of IRD, we use a batch size of 8, Adam optimizer and a learning rate of 0.001. Total training epochs is 80 and the learning rate is divided by 10 when epoch $= \{40, 60\}$. Three data augmentation methods, i.e. random crop, random scale and random flip are employed. For the training of IDPNet, we use a batch size of 1, SGD optimizer and a learning rate of 0.001. As the asymmetric structure of the upper and lower jaws, only the horizontal flip is applied for data augmentation. IDPNet is trained for 40 epochs and the learning rate is divided by 10 at 20th and 30th epochs, respectively. All the models are trained and tested on the platform of NVIDIA A100 GPU.

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

Network	Knowledge Alignment	Cross-modal Attention	AP_{75}↑	FLOPs(G) ↓
IRD	✓	✓	**18.4**	67.48
	✓	✗	17.1	56.88
	✗	✓	16.8	66.81
	✗	✗	16.2	**56.21**

Table 2. Performance comparison of different loss function.

$\mathcal{L}_{reg}$	$\mathcal{L}_{tiou}$	$\mathcal{L}_{TPL}$	Acc(R@1, IoU=m)		
			m=0.6	m=0.7	m=0.8
✓	✗	✗	28.8	23.7	15.3
✓	✓	✗	**35.6**	**28.8**	16.9
✓	✓	✓	33.9	25.4	**20.3**

### 3.2   Performance Analysis

In our task, the kernel of the implant should not invade the mandibular nerve canal and should maintain a minimum safety distance of 1.5mm. Therefore, as long as the center point of the implant root conforms to this rule, it is a good prediction. We follow previous work [1, 5, 6] to adopt Acc(R@1, IoU=m) as the performance evaluation metric, which represents the percentage accuracy of top-1 predicted moments whose IoU with the ground-truth moment is larger than m. We set the IoU threshold values m={0.6, 0.7, 0.8}.

Ablation Studies of IRD. As the IRD is used as pre-processing method to crop CBCT data for IDPNet, an approximate planting area is sufficient but requires quick inference speed. To evaluate the effectiveness of the proposed IRD, we conduct ablation experiments to investigate the effect of removing components in IRD, results are given in Table 1. AP_{75} and FLOPs are used as evaluation metrics. We can observe from the table that removing both modules will result in a 2.2% performance decrease, but FLOPs is decreased by 11.27. This results meet with the requirement and demonstrate that the proposed IRD is effective and lightweight for clinical practice.

Ablation Studies of Loss Function. To demonstrate the effectiveness of the proposed loss function, we conduct ablation experiments to investigate the effect of each loss function in Table 2. We can observe from the second row of the table that using temporal iou loss alone will lead to regression failure. When combining both regression loss and temporal iou loss, the accuracy (m=0.8) improves by 1.6%. When the TPL loss is introduced, the improvement reaches to 5.0%. Although the accuracy of smaller IoU thresholds has decreased, high IoU threshold is required in clinical practice. This results demonstrate the effectiveness of TPL loss, which enables the encoder to perceive the texture variation among slices.

Table 3. Performance comparison to the video grounding methods.

Method	Visual Feature	Acc(R@1, IoU=m) m=0.6	m=0.7	m=0.8
TSP-PRL [10]	C3D	33.1	**26.8**	18.6
MAN [16]	I3D	32.6	23.1	15.8
VSLNet [17]	I3D	31.2	23.5	17.1
DRN [15]	I3D	**34.5**	21.7	16.3
TPNet(ours)	-	33.9	25.4	**20.3**
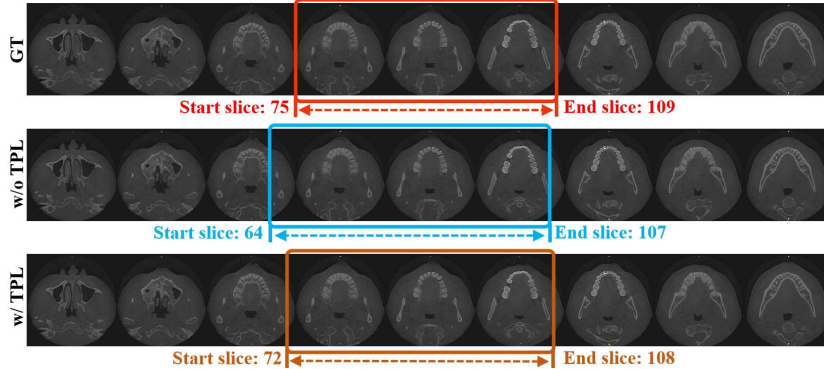
Fig. 4. Detection results of TPNet trained with or without TPL loss.

Visual Comparison. To further validate the effectiveness of the proposed TPL loss, in Fig. 4, we visualize the prediction result of TPNet with or without training by the TPL loss. From the figure we can observe that the introduction of TPL predict more precise start and end slices of the implant, due to the perception capability of texture variation.

Comparison to the Video Grounding Methods. As previously discussed, we model the task of implant depth prediction as video grounding. To demonstrate the superior performance of the proposed method, we compare TPNet with other state-of-the-art video grounding methods in Table 3. Specifically, we choose different visual feature based methods, e.g., the C3D-based method, TSP-PRL, and the I3D-based methods, MAN, VSLNet and DRN. From the table we can observe that, the C3D-based method perform better than the I3D-based networks in high iou threshold (e.g., TSP-PRL achieved 18.6% Acc, which is 1.5% higher than the best-performing I3D-based network - VSLNet). The proposed TPNet achieves the best accuracy of 20.3%, among all benchmarks. The experimental results proved the effectiveness of our method.

4   Conclusions

In this paper, we simplify the task of implant depth prediction as video grounding, and develop a texture perceive implant depth prediction network (TPNet).

TPNet consists of an implant region detector (IRD) and an implant depth prediction network (IDPNet). IRD is an object detector designed to reduce the size of CBCT by cropping a probable implant region from CBCT data. IDPNet is devised to regress the precise implant depth. A texture consistency (TC) loss is designed to enable the image encoder to capture more fine-grained features. Extensive experiments on a large dental implant dataset demonstrated that the proposed TPNet achieves superior performance than the existing methods.

Disclosure of Interests. The authors declare no competing interests.

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
2. 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)
3. Kurt Bayrakdar, S., Orhan, K., Bayrakdar, I.S., Bilgir, E., Ezhov, M., Gusarev, M., Shumilov, E.: A deep learning approach for dental implant planning in cone-beam computed tomography images. BMC Medical Imaging **21**(1), 86 (2021)
4. 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)
5. Liu, P., Zhang, J., Wu, X., Liu, S., Wang, Y., Feng, L., Diao, Y., Liu, Z., Lyu, G., Chen, Y.: Benchmarking supervised and self-supervised learning methods in a large ultrasound muti-task images dataset. IEEE Journal of Biomedical and Health Informatics (2024)
6. Liu, S., Liu, Y., Xu, X., Chen, R., Liang, D., Jin, Q., Liu, H., Chen, G., Zhu, Y.: Accelerated cardiac diffusion tensor imaging using deep neural network. Physics in Medicine & Biology **68**(2), 025008 (2023)
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
8. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)

9. Widiasri, M., Arifin, A.Z., Suciati, N., Fatichah, C., Astuti, E.R., Indraswari, R., Putra, R.H., Za'in, C.: Dental-yolo: Alveolar bone and mandibular canal detection on cone beam computed tomography images for dental implant planning. IEEE Access **10**, 101483–101494 (2022)

10. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12386–12393 (2020)

11. Yang, X., Li, X., Li, X., Chen, W., Shen, L., Li, X., Deng, Y.: Two-stream regression network for dental implant position prediction. Expert Systems with Applications **235**, 121135 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2023.121135, https://www.sciencedirect.com/science/article/pii/S0957417423016378

12. Yang, X., Li, X., Li, X., Wu, P., Shen, L., Deng, Y.: Implantformer: vision transformer-based implant position regression using dental cbct data. Neural Computing and Applications pp. 1–16 (2024)

13. Yang, X., Xie, J., Li, X., Li, X., Shen, L., Deng, Y.: Tcslot: Text guided 3d context and slope aware triple network for dental implant position prediction. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 726–732. IEEE (2023)

14. Yang, X., Xie, J., Li, X., Li, X., Li, X., Shen, L., Deng, Y.: Tceip: Text condition embedded regression network for dental implant position prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 317–326. Springer (2023)

15. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10287–10296 (2020)

16. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1247–1257 (2019)

17. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)

18. Zhang, Y., Chen, X., Jia, J., Liu, S., Ding, K.: Text-visual prompting for efficient 2d temporal video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14794–14804 (2023)