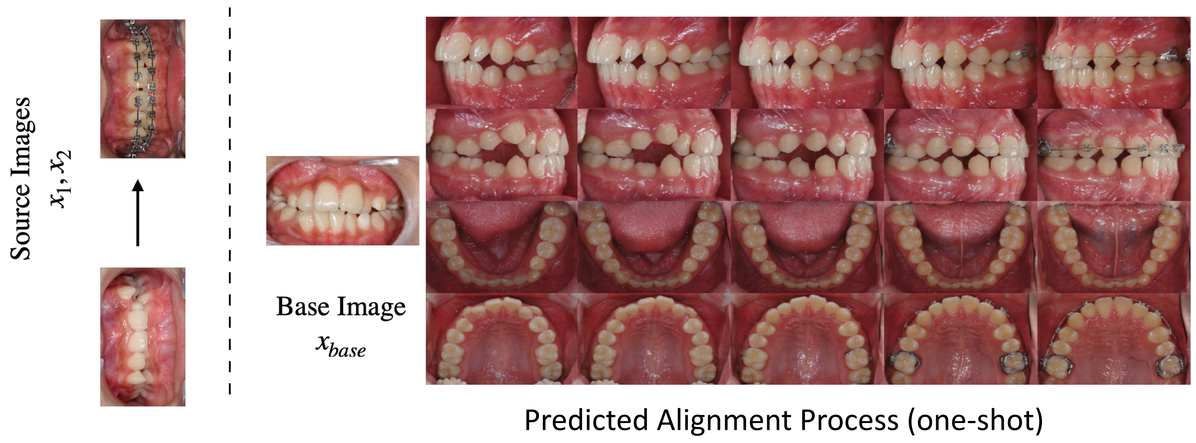
CLIP-IVP: CLIP-based Intraoral View Prediction Using Conditional GANs

Yewon Lim[1], Heejin Yu[1], Jongeun Choi[1,*], Jinwoo Kim[2,*]

[1]Yonsei University, Seoul, Republic of Korea
[2]Ewha Womans University, Seoul, Republic of Korea

TL;DR

- ***We propose a novel method called CLIP-based Intraoral View Prediction (CLIP-IVP) for predicting novel views of intraoral structure using only a single front teeth image of a patient.***
- ***Our approach leverages pre-trained CLIP image encoder to represent front intraoral images in zero-shot manner, reducing time and resources for training.***
- ***Our model can also be used to predict the orthodontic treatment process in a one-shot manner, which might be useful in treatment planning and prediction.***



Predicted Alignment Process (one-shot)

Abstract

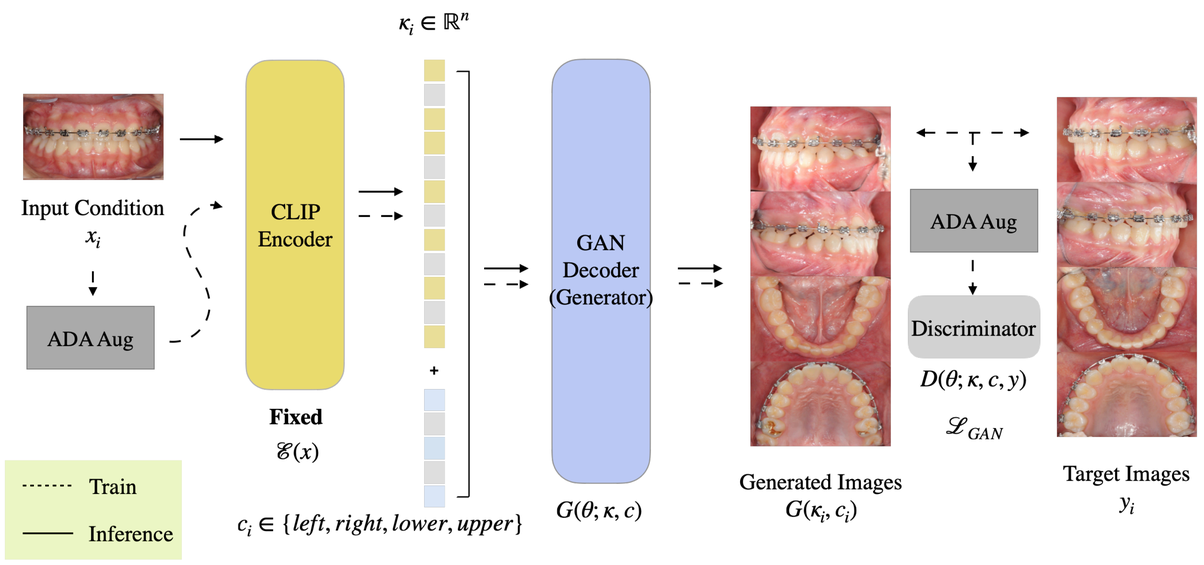
In this study, we propose a novel method called CLIP-based Intraoral View Prediction (CLIP-IVP) for predicting novel views of intraoral structure using only a single front teeth image of a patient. Our approach leverages pre-trained CLIP image encoder to represent front intraoral images, reducing time and resources for training. Our model achieves a Frechet Inception Distance (FID) score of 3.4 on the intraoral view prediction task. Our model can also be used to predict the orthodontic treatment process in a one-shot manner, which might be useful in treatment planning and prediction. The official implementation of our method is available on https://github.com/three0-s/clip-ivp.

Introduction

- While intraoral images play a pivotal role in orthodontics, obtaining a complete set of standardized intraoral images in real-world clinical settings remains challenging.
- Therefore, we propose a novel method called CLIP-based Intraoral View Prediction (CLIP-IVP) that utilizes a pre-trained CLIP image encoder to predict medically standardized full set of intraoral images from front intraoral images without additional clinical information.

Method

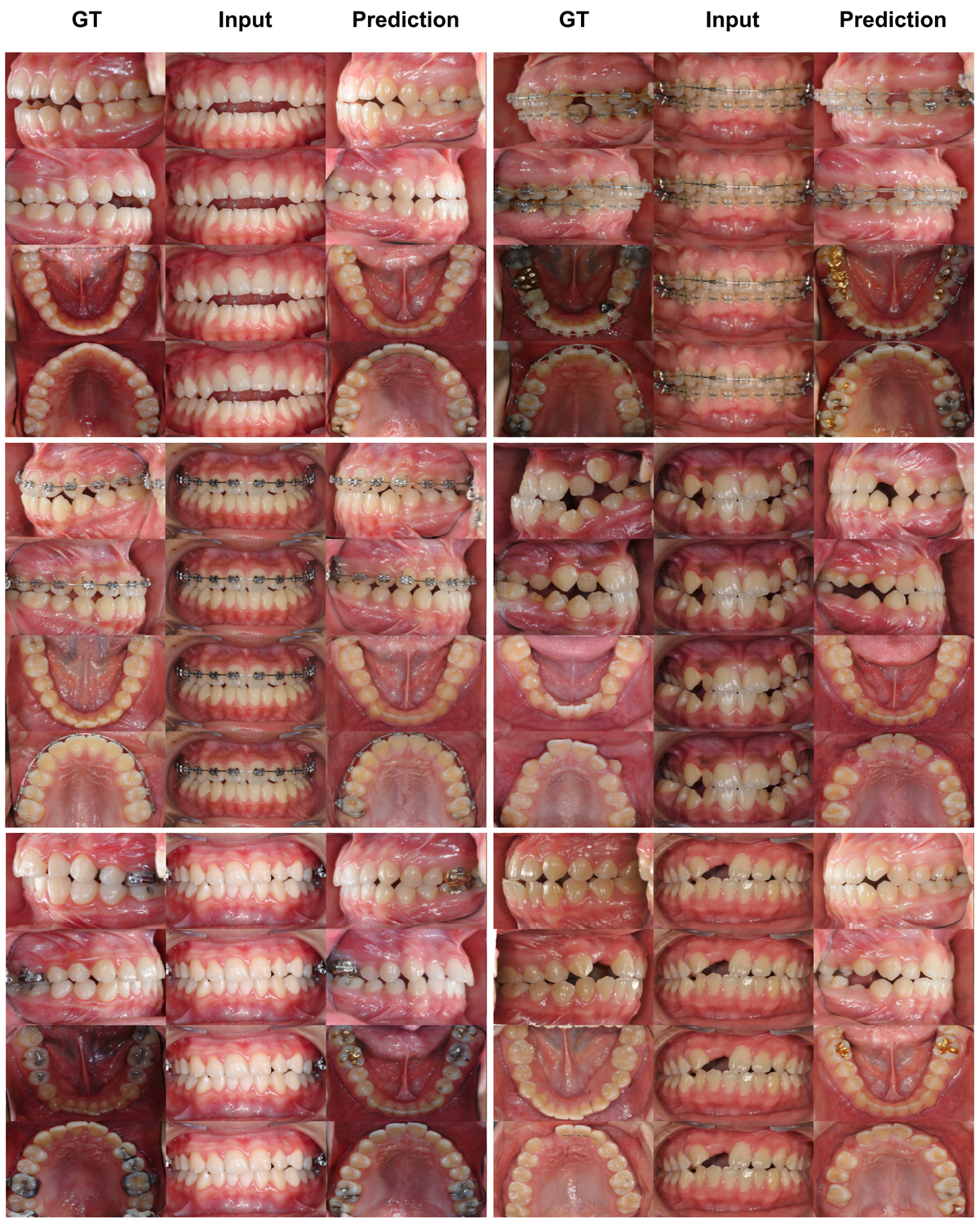
- Training dataset should include tuples (x_i, y_i, c_i), where x_i, y_i and c_i, respectively represents the front intraoral image, the corresponding view of intraoral structures, and the corresponding class label. By forwarding x_i through the CLIP image encoder, we obtain the corresponding CLIP image latent vector κ_i. During training, the decoder is conditioned on κ_i and c_i.
- Our CLIP Image latent decoder builds on the network proposed by Karras et al. [1].



Suggested network diagram.

Experiment

- We train our model using an internal dataset consisting of 7,000 pairs of intraoral images, while the test set and the validation set had 2,048 pairs and 512 pairs, respectively.
- We achieve an FID of 3.4 on image-to-image translation.
- By manipulating an image latent based on the disparity between pre-treatment and post-treatment image latent, our model can also be used to progressively predict the orthodontic treatment process of a patient using only a pair of source images.
- For one-shot *semantic nudge*, we manipulate the base image to reflect the semantic difference between a pair of source images, such as pre- and post treatment intraoral images. With semantic nudge, we can progressively predict the orthodontic treatment process of a patient. The predicted images show the process of leveling and alignment.



Conclusions & Acknowledgements

Our model provides a promising solution for intraoral view prediction and has the potential to improve the efficiency and accuracy of orthodontic treatment planning.

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI23C0162). This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 9991006713, KMDF_PR_20200901_0040)

References

[1] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)