DentalNet: Geometric Aware Multi-View Transformer for Occlusion Grade Prediction in Dental 3D Scans

Arnesh Batra[1†] **Arush Gumber**[1†] **Vaibhav Sharma**[1,2] **Peemit Rawat**[3]
Rinkle Sardana[3] **Tulika Tripathi**[3] **Anubha Gupta**[1*]

[1] SBI Lab, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India
[2]All India Institute of Medical Sciences (AIIMS), New Delhi, India
[3]Maulana Azad Institute of Dental Sciences (MAIDS), New Delhi, India

[†] Equal contribution [*]Corresponding Author

Abstract

The assessment of orthodontic treatment need, standardized by the Index of Orthodontic Treatment Need (IOTN), is a cornerstone of clinical dentistry. The conventional workflow for grading its Dental Health Component (DHC) is a laborious process, where physical study models must be prepared for the manual measurement of geometric discrepancies. To overcome this, we formulate the task of automated IOTN-DHC classification directly from scanned dental 3D models. Furthermore, we conduct a study of retrospectively collected real clinical cases, with ground truth labels annotated by an orthodontic specialist, to serve as the benchmark for this problem. We propose DentalNet, a multi-modal deep learning architecture which fuses information from 2D rendered views and 3D point clouds. It achieves this by integrating a Vision Transformer for visual context and a Point Transformer for precise geometry, via a cross-attention mechanism to learn the critical inter-arch relationships that define malocclusion. On our 4-class classification task, DentalNet outperforms the state-of-the-art models, achieving a mean F1-score of 67.03% which significantly surpasses the best performing 2D and 3D baselines i.e. ConvNeXtV2 (49.79%) and PointNet++ (55.36%) demonstrating the utility of our multimodal approach.

1 Introduction

The Index of Orthodontic Treatment Need (IOTN) is a widely used standard to grade the severity of dental malocclusions. It comprises two components: the Dental Health Component (DHC), which focuses on clinical factors such as overjet, overbite, crossbite, and tooth displacement, and the Aesthetic Component (AC), which reflects the visual appearance of the dentition. IOTN-DHC is the objective component of the IOTN index while IOTN-AC is the subjective component of the index. Despite its value, calculating IOTN still remains a manual task that can be labor-intensive and prone to inter-operator variability.

Although digital dentistry and 3D intraoral scanning have become increasingly common, the automated classification of IOTN grades remains relatively underexplored. Obtaining expertly labeled data is a significant bottleneck in this domain. Our study therefore, makes use of a novel clinical dataset of 123 3D intraoral scans comprising samples of all 5 grades of DHC of IOTN. For the purpose of developing a robust model, we have merged grade 1 and grade 2 into one class as the DHC scale is inherently ordinal and a model must understand that misclassifications between adjacent

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Imageomics: Discovering Biological Knowledge from Images Using AI.

grades (e.g., Grade 3 vs. 4) are a less severe error than one between distant grades (e.g., Grade 2 vs. 5).

The development of a robust automated system for IOTN grading holds significant clinical implications and would alleviate the time-consuming nature of manual assessment. This work establishes the foundational elements for a new direction in automated orthodontic diagnostics. We provide the first comprehensive solution to this problem, including a novel task formulation, a benchmark dataset, and a high-performing baseline model. Our primary contributions are as follows:

- **A Novel Multi-Modal Architecture (DentalNet)**: Proposes a deep learning framework that fuses information from 2D rendered images and 3D point clouds. A cross-attention mechanism is used to model the critical relationship between 2D and 3D mapping them to give a holistic prediction.
- **A New Clinical Study**: Conducted a new study of retrospectively collected clinical cases, this study is based on 123 3D intraoral scans with ground truth labels annotated by an orthodontic specialist.
- **Sets a New Benchmark**: Sets a new benchmark for the task with a mean F1-score of 67.03%. This performance significantly surpasses the best 2D (ConvNeXtV2 at 49.79%) and 3D (PointNet++ at 55.36%) baseline models.

2 Related Works

The Index of Orthodontic Treatment Need (IOTN) [4] is the clinical standard for grading malocclusion, but its manual application is time-consuming and prone to inter-observer variability [5]. While deep learning has been successfully applied to 2D dental images like radiographs for tasks such as tooth detection [14] and landmark localization [1, 12], these approaches are inherently limited for IOTN assessment. They cannot capture the precise three-dimensional morphology and crucial inter-arch relationships that define the Dental Health Component (DHC) [8]. The advent of 3D intraoral scanners has enabled the use of point cloud architectures like PointNet++ [17] for geometric analysis [10, 9]. However, even when applied to full dental arches, these unimodal 3D methods often lack an explicit mechanism to model the complex occlusal relationship between the two jaws, which is fundamental to IOTN grading.

Multimodal fusion is a well-established strategy in medical imaging [11, 7] for integrating complementary data sources to improve diagnostic accuracy [2, 18]. In dentistry, however, this approach remains comparatively underexplored, with limited studies combining data types like CBCT and photographs for treatment planning [13, 20]. Crucially, to the best of our knowledge, no prior work has developed a multimodal fusion framework specifically for the automated classification of IOTN-DHC. This critical gap motivates our work, which combines the rich contextual information from 2D rendered views with the precise geometric data from 3D point clouds in a unified architecture to address the limitations of unimodal approaches.

3 Methodology

The clinical assessment of IOTN grades relies on a dual understanding of the patient's dentition: a precise geometric measurement of malocclusion (e.g., overjet, overbite, crowding) and a visual appraisal of aesthetic features. Standard deep learning models that process only 2D images or 3D geometry in isolation struggle with this task. 2D models fail to capture precise 3D spatial relationships, while 3D models often analyze jaws independently, missing the critical inter-arch relationship that defines occlusion. Our proposed model, DentalNet, is a bespoke multi-modal architecture engineered to emulate a clinician's holistic reasoning by simultaneously interpreting 2D appearance and 3D relational geometry. The overall architecture is depicted in Figure 1.

3.1 Data Modalities and Preprocessing

To create our multimodal dataset, we preprocess primary 3D STL mesh files for the upper and lower jaws into two distinct modalities. First, we generate 2D data by rendering eight canonical views (518×518 pixels) to capture textural and holistic morphological information; a key step involves

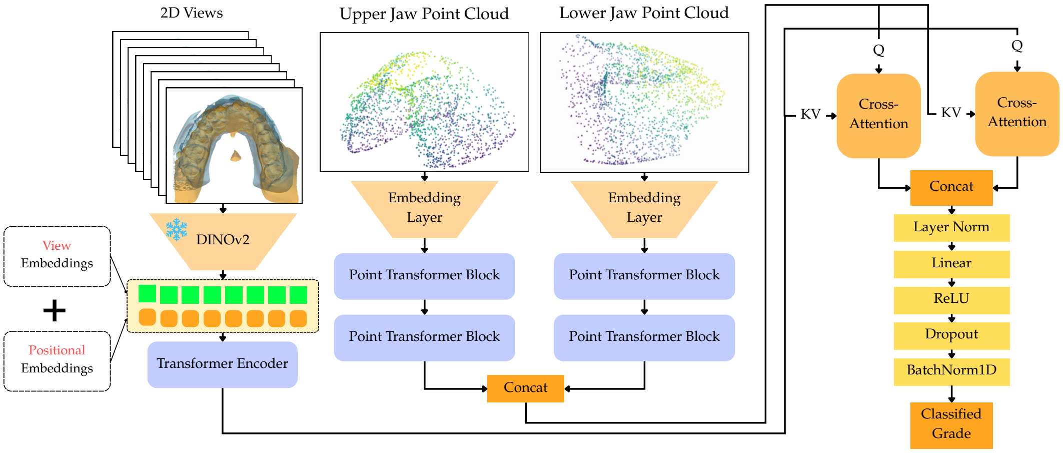

Figure 1: The DentalNet architecture. A multi-view 2D branch (left) and a dual 3D point cloud branch (center) extract features in parallel. These are integrated via a bidirectional cross-attention block (right) that explicitly models the inter-arch relationship for final classification.

overlaying the jaws with transparency to visualize their relative distances. Second, we create 3D point clouds by uniformly sampling $N = 2048$ points from each mesh (P_{upper}, P_{lower}), which are then normalized by centering and scaling to fit within a unit sphere.

3.2 DentalNet Architecture

DentalNet is composed of three key components: a 2D branch for robust view-based feature extraction, a 3D branch for fine-grained geometric analysis, and a fusion block designed to explicitly model the occlusal relationship.

3.2.1 Multi-View 2D Branch

To extract a global, view-invariant feature from the eight rendered images, we leverage the power of a pre-trained DINOv2 [16] Vision Transformer (ViT-Base) [6] which is trained on the LVD-142M dataset. To mitigate overfitting on our small medical dataset, we employ a strong regularization strategy by freezing all backbone weights except for the final transformer block, which is fine-tuned. The features from the eight views $\{f_1, ..., f_8\}$ are infused with learnable positional embeddings to retain viewpoint information. We add a 2-layer Transformer Encoder to help understand the sequential nature of the views.

3.2.2 Geometric 3D Branch

The 3D branch is engineered to learn directly from the raw point cloud representations of the dental arches, P_{upper} and P_{lower}. For this, we employ a Point Transformer encoder [21], an architecture designed to process unordered point sets. The core of this encoder is a vector self-attention mechanism [19, 3] applied within local point neighborhoods. This allows the network to dynamically learn context-aware weights for aggregating features, making it highly effective at capturing intricate geometric details like tooth curvature, cusp morphology, and inter-arch displacement without reliance on a mesh topology. The encoder operates hierarchically, progressively downsampling the point cloud to learn features at multiple scales, from fine-grained details to the overall arch form. To ensure a consistent and comparable feature space, a single, shared-weight encoder processes both jaws. This network culminates in a global pooling layer that produces compact 256-dimensional feature vectors, F_{upper} and F_{lower}, which robustly encapsulate the intrinsic geometries of their respective jaws.

3.2.3 Cross-Modal Fusion to model Dental occlusion

This stage is the core of our contribution, designed to explicitly model occlusal state. A purely geometric assessment of malocclusion requires understanding how the upper and lower jaws relate to one another. We achieve this by first concatenating the individual jaw features (F_{upper}, F_{lower})

Table 1: A comparison of model performance on the IOTN classification task. Results are reported as mean F1 Score and Accuracy (± standard deviation in %) over 5-fold cross-validation.

Model	**Accuracy (%)**	**F1 Score (%)**
Random Baseline	25.00	25.00
2D Models		
ResNet18	45.50 ± 4.33	37.96 ± 4.05
ViT	48.00 ± 5.00	44.95 ± 5.02
DINOv2	43.90 ± 7.00	39.69 ± 7.32
ConvNeXtV2	51.17 ± 8.72	49.79 ± 8.65
3D Models		
PointNet	55.82 ± 9.42	55.67 ± 9.43
PointMLP	52.80 ± 6.37	50.40 ± 7.75
PointNet++	56.56 ± 10.95	55.36 ± 11.32
DGCNN	52.31 ± 3.99	50.22 ± 4.18
PointNeXt	55.19 ± 4.39	54.55 ± 4.40
PointGPT	48.96 ± 6.91	46.83 ± 4.80
2D+3D Hybrid Model		
DentalNet (ours)	**67.60** ± 11.31	**67.03** ± 10.96

and passing them through a linear layer to produce a unified geometric representation, $F_{3D} \in \mathbb{R}^{512}$, which captures the combined state of both arches.

Next, a **bidirectional cross-attention mechanism** integrates the 2D and 3D representations. This allows each modality to query the other, enriching its own features with relevant context. For example, the 3D geometry can be informed by visual cues of discoloration or wear from the 2D views, while the 2D features can be grounded in the precise spatial information from the 3D model. This fusion produces two refined feature vectors, F'_{2D} and F'_{3D}, which are finally concatenated and passed to the main classifier. This explicit modeling of the inter-arch relationship is what allows DentalNet to learn the subtle geometric discrepancies that define IOTN grades, a capability lacking in unimodal approaches.

4 Results

Analysis of the IOTN-DHC classification results reveals a clear performance hierarchy by modality. 2D models, exemplified by ConvNeXtV2, consistently underperformed, achieving a maximum F1 score of just 49.79%. This limitation can be attributed to the loss of crucial 3D depth and relational information during the projection of dental geometry onto 2D planes. 3D models, such as PointNet++, which operate directly on point clouds, demonstrated a superior aptitude for geometric feature extraction achieving an F1 score of 55.36%. Our proposed DentalNet, based on the multimodal fusion strategy achieved the highest F1 score of 67.03%. This architecture surpasses the best 3D and 2D baselines by 11.67 and 17.24 percentage points, respectively. This outcome confirms that bidirectional cross-attention mechanism successfully creates a synergistic representation that integrates contextual visual understanding with precise geometric analysis.

5 Conclusion

We introduce DentalNet, a deep learning framework setting a new state-of-the-art benchmark for automated IOTN-DHC grade classification. It uses multimodal data (3D point clouds and 2D images) with 2D/3D branches and a cross-modal fusion block to model jaw relationships. This work also utilizes a novel in-house 3D intraoral scan dataset, created due to the lack of publicly available alternatives. Future work will aim to refine the architecture and expand the dataset.

6 Acknowledgements

The authors would like to thank the Centre of Excellence in Healthcare (CoEHe), IIIT-Delhi for supporting this study.

References

[1] Sercan Ö Arık, Bulat Ibragimov, and Lei Xing. Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging*, 4(1):014501–014501, 2017.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Arnesh Batra, Arush Gumber, and Anushk Kumar. M-scan: A multistage framework for lumbar spinal canal stenosis grading using multi-view cross attention. *arXiv preprint arXiv:2503.01634*, 2025.

[4] Peter H Brook and William C Shaw. The development of an index of orthodontic treatment priority. *The European Journal of Orthodontics*, 11(3):309–320, 1989.

[5] Charles Daniels and Stephen Richmond. The development of the index of complexity, outcome and need (icon). *Journal of orthodontics*, 27(2):149–162, 2000.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Patrick Godau, Veronika Cheplygina, Michal Kozubek, Sharib Ali, et al. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568*, 2022.

[8] Ana-Maria Haude, Thomas Lehmann, Christoph-Ludwig Hennig, and Collin Jacobs. Comparison of conventional two-dimensional and digital three-dimensional imaging in orthodontics: A systematic review and meta-analysis. *Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie*, pages 1–18, 2025.

[9] Golriz Hosseinimanesh, Ammar Alsheghri, Julia Keren, Farida Cheriet, and Francois Guibault. Personalized dental crown design: A point-to-mesh completion network. *Medical Image Analysis*, 101:103439, 2025.

[10] Joon Im, Ju-Yeong Kim, Hyung-Seog Yu, Kee-Joon Lee, Sung-Hwan Choi, Ji-Hoi Kim, Hee-Kap Ahn, and Jung-Yul Cha. Accuracy and efficiency of automatic tooth segmentation in digital dental models using deep learning. *Scientific reports*, 12(1):9429, 2022.

[11] Pulkit Kumar, Pravin Nagar, Chetan Arora, and Anubha Gupta. U-segnet: fully convolutional neural network based automated brain tissue segmentation tool. In *2018 25th IEEE International conference on image processing (ICIP)*, pages 3503–3507. IEEE, 2018.

[12] Jeong-Hoon Lee, Hee-Jin Yu, Min-ji Kim, Jin-Woo Kim, and Jongeun Choi. Automated cephalometric landmark detection with confidence regions using bayesian convolutional neural networks. *BMC oral health*, 20(1):270, 2020.

[13] Amornrut Manosudprasit, Arshan Haghi, Veerasathpurush Allareddy, and Mohamed I Masoud. Diagnosis and treatment planning of orthodontic patients with 3-dimensional dentofacial records. *American Journal of Orthodontics and Dentofacial Orthopedics*, 151(6):1083–1091, 2017.

[14] Yuma Miki, Chisako Muramatsu, Tatsuro Hayashi, Xiangrong Zhou, Takeshi Hara, Akitoshi Katsumata, and Hiroshi Fujita. Classification of teeth in cone-beam ct using deep convolutional neural network. *Computers in biology and medicine*, 80:24–29, 2017.

[15] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[18] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Dan Zhao, Morteza Homayounfar, Zhe Zhen, Mei-Zhen Wu, Shuk Yin Yu, Kai-Hang Yiu, Varut Vardhanabhuti, George Pelekos, Lijian Jin, and Mohamad Koohi-Moghadam. A multimodal deep learning approach to predicting systemic diseases from oral conditions. *Diagnostics*, 12(12):3192, 2022.

[21] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

7 Appendix: Supplementary Diagrams and Details

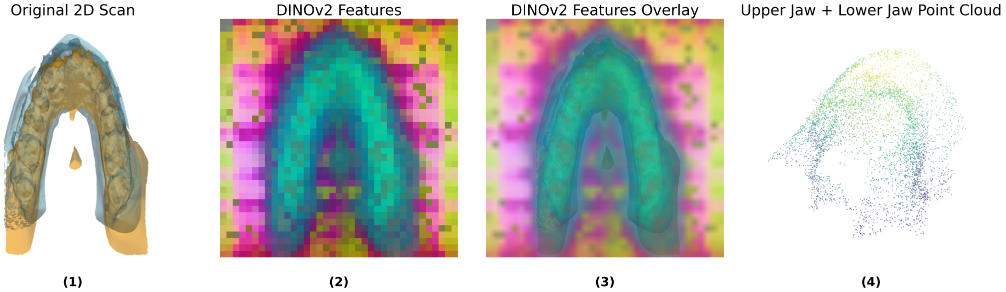

Figure 2: Qualitative analysis of DentalNet's multi-modal feature learning. From left to right: **(1)** A sample 2D rendered view used as input. **(2)** Visualization of the DINOv2 patch features via Principal Component Analysis (PCA), showing that the model learns to focus on clinically relevant dental anatomy. **(3)** The feature map overlaid on the original scan, confirming that attention is concentrated on the occlusal and incisal surfaces. **(4)** The corresponding 3D point cloud representation of the complete dental arches.

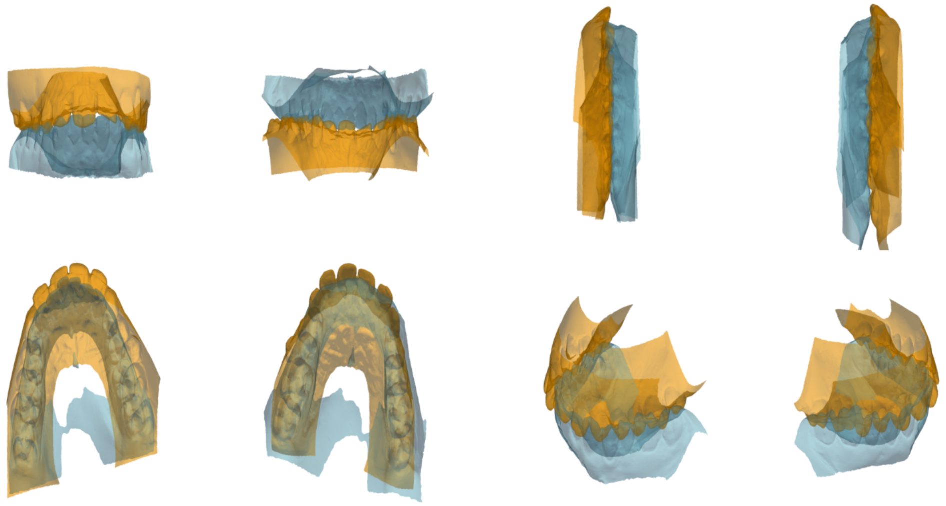

Figure 3: The eight canonical views rendered from a 3D dental model to serve as input for the 2D branch of DentalNet. This multi-view approach ensures that all clinically relevant angles, such as frontal (occlusion), buccal (side bite), and occlusal (top/bottom), are captured.

7.1 Heavily Regularized Training Objective

Training a complex, multi-modal model on a small dataset presents a significant risk of overfitting. To combat this, we employ a heavily regularized training objective based on deep supervision. In addition to the primary classifier that acts on the final fused features, we attach three auxiliary classification heads directly to the outputs of the unimodal branches (F_{2D}, F_{upper}, and F_{lower}).

This creates a multi-task learning environment where the total loss $\mathcal{L}_{total}$ forces each component of the network to become a proficient feature extractor in its own right:

$$\mathcal{L}_{total} = \mathcal{L}_{CE}(\hat{y}_{final}, y) + \alpha \sum_{b \in \{2D, upper, lower\}} \mathcal{L}_{CE}(\hat{y}_b, y) \tag{1}$$

Here, $\mathcal{L}_{CE}$ is the Cross-Entropy loss with label smoothing ($\epsilon = 0.1$) [15], y is the ground truth label, $\hat{y}$ are the predictions from the final and auxiliary branch (b) classifiers, and α is a hyperparameter balancing the regularization strength. This objective ensures that the feature extractors do not become lazy, improving the quality of the features available for the final fusion and enhancing overall model robustness.

Table 2: Dataset distribution by IOTN grade.

IOTN Grade	Count
5	38
4	32
3	34
1,2	19

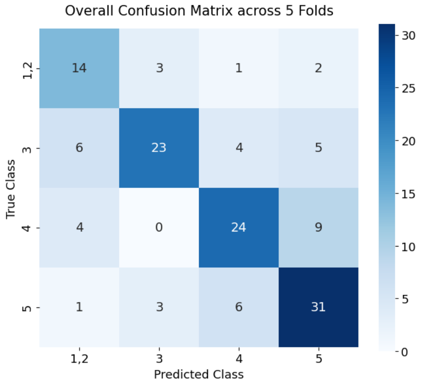

Figure 4: Confusion matrix of DentalNet predictions.