AUTOMATED STAGING OF SPHENO-OCCIPITAL SYNCHONDROSIS FUSION VIA LANDMARK-GUIDED KNOWLEDGE DISTILLATION

Omid Halimi Milani[1], *Amanda Nikho*[2], *Lauren Mills*[2], *Marouane Tliba*[3], *Rashid Ansari*[1], *Ahmet Enis Cetin*[1], *Mohammed H Elnagar*[2]

[1]Department of Electrical and Computer Engineering, University of Illinois Chicago, Chicago, IL, USA
[2]Department of Orthodontics, College of Dentistry, University of Illinois Chicago, Chicago, IL, USA
[3]University of Orleans, France

ABSTRACT

A novel deep learning framework is proposed for automated staging of spheno-occipital synchondrosis (SOS) fusion, a critical diagnostic indicator in orthodontics and forensic anthropology. Clinicians rely on precise anatomical region defined by landmarks to determine SOS fusion stages. To emulate this diagnostic approach computationally, an object detection model (YOLOv11) was first trained to accurately localize and crop regions of interest (ROI) based on clinician-annotated anatomical landmarks. Utilizing these cropped regions, a specialized classification model was developed capable of accurately predicting the SOS fusion stages. To leverage both the localized expertise of this classifier and the global context provided by full medical images, knowledge distillation was applied, transferring specialized diagnostic knowledge from the cropped-region classifier (teacher model) to a holistic image classifier (student model). Our approach includes a regularization term that encourages alignment between the student's Grad-CAM activation maps and the bounding boxes provided by the teacher model, thus enhancing model interpretability and ensuring consistent diagnostic focus. Comprehensive evaluations demonstrate that our knowledge distillation-driven framework significantly outperforms conventional methods, providing an efficient and robust solution for automated SOS fusion staging.

1. INTRODUCTION

Accurately determining skeletal maturation is essential in both orthodontics and forensic anthropology. The SOS, a cartilaginous joint located in the cranial base, is the last synchondrosis to fuse and plays a pivotal role in postnatal craniofacial development [1]. Its fusion status is a crucial diagnostic indicator, facilitating orthodontic treatment planning [2, 3], predicting pubertal growth spurts [4], and estimating chronological age in forensic investigations [5]. Fig.1 illustrates the orientation and segmentation process, highlighting the areas where doctors focus to determine the fusion stage.

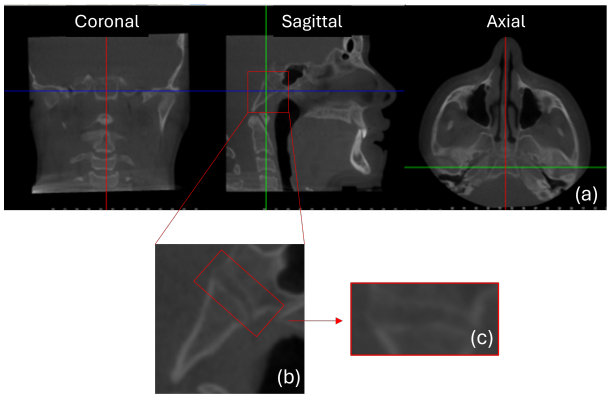


Fig. 1: Skull oriented in three planes (a). Occipital and sphenoid bones cropped (b). SOS rotated and segmented (c).

Despite its clinical significance, manual assessment of SOS fusion is frequently inconsistent due to variations in staging methodologies and imaging modalities. The cervical vertebrae maturation (CVM) method, commonly used for skeletal age estimation, is not universally applicable, particularly in situations where the cervical vertebrae fall outside the imaging field [6, 7]. Consequently, SOS fusion assessment emerges as a valuable alternative marker for skeletal maturity. Nonetheless, existing studies demonstrate significant variability in SOS classification schemes, employing anywhere from three to six stages of fusion [8, 9, 10]. Moreover, differences in imaging techniques—including histological sections, two-dimensional radiography, computed tomography (CT), and cone-beam computed tomography (CBCT)—exacerbate inconsistencies in clinical interpretations [11, 12].

To address these challenges, our work proposes a novel automated framework guided by expert anatomical annotations. Our approach integrates landmark-guided object detection with knowledge distillation. Specifically, an object detection model (YOLOv11) is first employed, precisely guided by expert annotations, to accurately localize critical anatomical landmarks relevant to SOS fusion assessment. Subsequently,

a specialized classification network is trained using these localized regions to determine the fusion stages.

This classification model, trained exclusively on expert-guided regions, serves as a teacher model within our knowledge distillation framework. The expert-driven diagnostic knowledge is then transferred to a student model designed to operate on the entire SOS image independently, eliminating the need for explicit landmark annotations at inference time. This student model effectively learns to identify clinically significant structures implicitly.

Our main contributions are summarized as follows.

(i) Our work introduces a framework designed specifically to automate and accurately classify the stages of SOS fusion, effectively addressing current inconsistencies in manual assessment methodologies.

(ii) Our work uses an innovative knowledge distillation strategy, guided by expert anatomical annotations, enhancing both accuracy and interpretability by implicitly embedding expert-level diagnostic insights into the holistic image classifier model.

(iii) Our work conducts extensive experiments to rigorously evaluate and compare state-of-the-art image classification models with our proposed training approach. The results validate the effectiveness of our knowledge distillation framework, demonstrating improvements in diagnostic accuracy, and interpretability.

2. METHODOLOGY

This section illustrates our extended proposed deep learning framework for automated SOS fusion staging [13], incorporating expert-guided annotations, knowledge distillation, and gradient-based attention. An overview of the overall approach is first presented, followed by a detailed description of the dataset and preprocessing pipeline, and finally, an explanation of our knowledge distillation strategy.

The primary goal of our framework is to accurately classify the stage of SOS fusion using deep neural networks. In clinical practice, SOS stage determination relies on local anatomical indicators, which motivates our use of expert-guided annotations during training. Our approach leverages a teacher-student paradigm: *(i)* a teacher model is trained on expertly cropped regions, capturing highly localized features relevant to SOS fusion, and *(ii)* a student model learns to classify the entire uncropped SOS image by distilling knowledge from the teacher. Furthermore, a regularization term is incorporated into our training strategy to explicitly align the student's attention maps, generated via Grad-CAM[14], with the expert-annotated landmark regions. This alignment ensures that the holistic classifier consistently focuses on diagnostically relevant areas, thereby enhancing interpretability, reliability, and overall diagnostic performance. The final framework yields a holistic classifier that requires no explicit landmark cropping at inference, yet benefits from localized diagnostic cues.

2.1. Dataset and Preprocessing for SOS Fusion Staging

This retrospective study utilized anonymized cone-beam computed tomography (CBCT) scans from 723 patients (260 males, 370 females, 93 unspecified), aged 7–68 years, in accordance with institutional ethical standards at the University of Illinois Chicago's Office for the Protection of Research Subjects (OPRS). Each SOS was categorized into five stages according to the classification system proposed by Bassed et al. [15], ranging from entirely open (Stage 1) to completely fused (Stage 5). This scheme was selected because its methodology most closely parallels ours, particularly in its use of three-dimensional acquisition. Three trained evaluators independently assigned the fusion stages, achieving high inter-rater reliability (Cronbach's alpha = 0.945).

All CBCT images were imported into Dolphin Imaging software and standardized in the axial, coronal, and sagittal planes. From each scan, the midsagittal slice was extracted, aligned, and cropped to a 2×1 aspect ratio centered on the SOS region. The final dataset comprised 158 scans at Stage 1, 88 at Stage 2, 92 at Stage 3, 124 at Stage 4, and 252 at Stage 5.

2.2. Knowledge Distillation Framework

To leverage localized anatomical insights for whole-image classification, a knowledge distillation framework was proposed that employs a teacher-student paradigm. The teacher model is trained on expertly cropped SOS regions identified via YOLOv11, producing logits $\mathbf{z}_{\text{teacher}}$ localized to the synchondrosis. The student model, in contrast, takes as input the uncropped images and outputs logits $\mathbf{z}_{\text{student}}$. Their training was formalized with the following combined loss:

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{reg}} + \beta\,\mathcal{L}_{\text{dist}} + \theta\,\mathcal{L}_{\text{cls}}, \tag{1}$$

where each term is described below.

Distillation Loss ($\mathcal{L}_{dist}$): To transfer knowledge from teacher to student, their output logits were aligned through the Kullback–Leibler (KL) divergence [16, 17]:

$$\mathcal{L}_{\text{dist}} == \sum_{k=1}^{K} \mathbf{z}_{\text{teacher}} \log\Big(\tfrac{\mathbf{z}_{\text{teacher}}}{\mathbf{z}_{\text{student}}}\Big), \tag{2}$$

with $\mathbf{z}_{\text{teacher}}$ and $\mathbf{z}_{\text{student}}$ denoting the teacher and student's softened probability distributions, respectively. By minimizing $\mathcal{L}_{\text{dist}}$, the student acquires the teacher's localized diagnostic representations.

Classification Loss ($\mathcal{L}_{cls}$): The standard cross-entropy loss was incorporated to drive the primary classification objective:

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{K} y_k\,\log\big(\hat{y}_k\big), \tag{3}$$

where y_k is the ground-truth label distribution and $\hat{y}_k$ is the predicted probability of class k.

A *Regularization Term ($\mathcal{L}_{reg}$)* was also used to the above cost function. An explicit alignment between the student's gradient-based activation (Grad-CAM) and expert-defined landmark masks was imposed to guide the network's attention toward clinically relevant regions:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{A}_i - M_i\right)^2, \tag{4}$$

where $\widehat{A}_i$ is the Grad-CAM map generated by the student for the i-th sample, and M_i is the corresponding expert-annotated region. This term ensures that even when learning from entire images, the student model concentrates on key anatomical structures.

Hyperparameter Tuning: Coefficients α, β, and θ in Eq. (1) are empirically tuned to balance the influence of each term. The following was set to achieve optimal convergence in practice: $\alpha = 1 \times 10^{-2}$, $\beta = 0.8$, and $\theta = 0.2$.

2.3. Student and Teacher Classification Backbone

A ConvNeXt base architecture was adopted [18] as the core feature extractor, leveraging its modernized convolutional design for efficient representation learning. To enhance the model's capacity for spatial awareness, a self-attention block at the final layer was inserted. Convolutional stages of ConvNeXt capture hierarchical features and global context, while the self-attention mechanism selectively reweights spatial regions to emphasize critical anatomical landmarks related to SOS fusion. Following attention-based feature aggregation, a fully connected layer outputs the final classification, and the new model is referred to as ConvNeXt+. This simple setup integrates seamlessly with our knowledge distillation framework, enabling both the teacher model (trained on expertly cropped regions) and the student model (operating on uncropped images) to benefit from discriminative, attention-driven feature representations. The cropped regions used to train the teacher model are automatically extracted using a YOLO object detector trained on expert-annotated landmarks, enabling consistent and efficient ROI generation without manual cropping.

3. RESULTS AND ANALYSIS

This section presents a detailed evaluation of our classification models on the SOS dataset, comparing various architectures and assessing the impact of our knowledge distillation framework.

3.1. Object Detection Accuracy

The object detection capabilities of YOLOv11 was assessed, which guides the model's spatial attention during training.

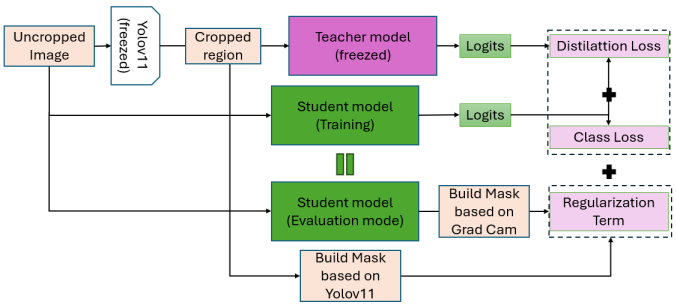


Fig. 2: A framework for training the student model using distillation loss and a regularization term.

YOLOv11 achieves an mAP@0.5 of 72.8%, indicating robust detection performance across a range of anatomical landmarks relevant to SOS fusion. This strong detection performance lays the foundation for accurately localizing the regions of interest used in our subsequent knowledge distillation framework. The teacher model was tested on the cropped region of interest from YOLOv11, achieving an accuracy of 82.49%, precision of 82.63%, recall of 82.50%, and an F1 score of 82.34%.

Table 2 summarizes the comparative performance of several architectures evaluated on the SOS dataset, and the most adopted architecture in medical imaging was selected for use. The following are key observations.

EfficientNet-B0 demonstrates a modest performance of 70.58% accuracy, indicating the challenges of representing the subtle cranial-base features in a lightweight model. ResNet34 and ResNet50 offer moderate gains of approximately 4–5% over EfficientNet-B0, suggesting that deeper residual networks capture more nuanced textures but show diminishing returns beyond a certain depth. ConvNeXt achieves a substantial jump to 78.99% accuracy, owing to its modernized convolutional design that balances global structure and fine details. Finally, our proposed ConvNeXt+ integrates targeted attention mechanisms into the ConvNeXt backbone, delivering an accuracy of 79.97% and an F1 score of 78.87%, highlighting the benefits of enhancing landmark localization without sacrificing global context.

ConvNeXt+ strikes the best trade-off between model capacity and attention-based feature refinement.

3.2. Impact of Knowledge Distillation

Table 1 details the quantitative benefits of applying our knowledge distillation (KD) approach to the ConvNeXt+ architecture:

i) Accuracy Gain: Incorporating KD boosts accuracy from 79.97% to 83.05%, demonstrating that information from the teacher model (trained on localized, expertly cropped SOS regions) effectively guides the student model to focus on relevant anatomical landmarks in the full image.

Table 1: Performance Comparison of ConvNeXt+ with and without Knowledge Distillation (KD)

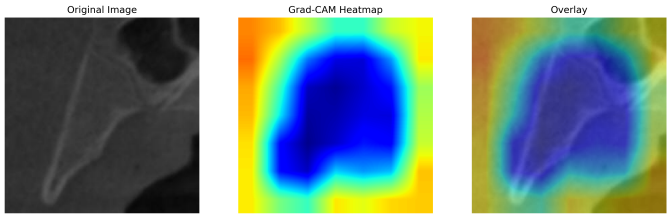
Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
ConvNeXt+	79.97	80.65	79.97	78.87
ConvNeXt+ (KD)	**83.05**	**83.26**	**83.05**	**82.07**

Table 2: Performance Comparison of Classification Models on our SOS Dataset

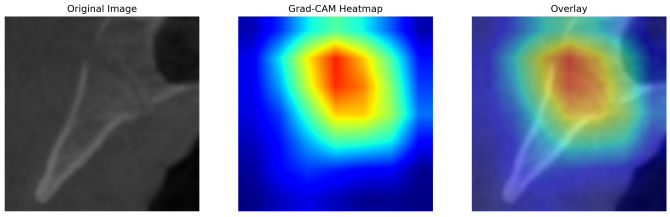
Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
EfficientNet-B0	70.58	69.52	70.58	69.13
ResNet34	75.07	75.01	75.07	74.05
ResNet50	74.79	74.08	74.79	73.79
ConvNeXt	78.99	78.57	78.99	78.35
ConvNeXt+ (Ours)	**79.97**	**80.65**	**79.97**	**78.87**

ii) Precision and Recall: By aligning the student's Grad-CAM outputs with the teacher's spatial knowledge, both precision and recall exceed 83%, indicating a balanced improvement in identifying each SOS fusion stage without inflating false positives or false negatives.

iii) F1 Score Increase: The F1 score rises from 78.87% to 82.07%, affirming that KD mitigates misclassifications where morphological changes might be subtle or confounded by surrounding structures. This indicates that the distilled student model is more robust and consistent across stages.



(a) Incorrect Focus



(b) Correct Focus

Fig. 3: Grad-CAM visualizations comparing incorrect (a) and correct (b) model focus [19]. In (a), the red areas show where the model focuses more, and the blue areas indicate less attention. In (b), after training with the proposed framework, the heatmap correctly highlights the region of interest, ensuring accurate and interpretable decision-making.

Qualitative Observations:

The KD-enabled student produces Grad-CAM heatmaps better aligned with the synchondrosis region (Fig. 3), supporting our hypothesis that localized knowledge transfer suppresses irrelevant background and enhances SOS classification.

4. CONCLUSION

In this study, a novel deep learning framework was proposed for automated SOS fusion staging that effectively balances localized anatomical precision and holistic image interpretation. Robust knowledge distillation is enabled through a teacher-student paradigm, where the teacher model is trained on expertly cropped ROIs, and the student model operates on uncropped images. Our inclusion of a Grad-CAM-based regularization term ensures that clinically relevant regions remain the focal point, thereby boosting interpretability and diagnostic accuracy. Experimental results demonstrate that our approach achieves state-of-the-art performance on a comprehensive SOS dataset, outperforming conventional models. The combination of YOLOv11-guided landmark detection, ConvNeXt+ backbones, and knowledge distillation led to significant gains in both classification metrics and visual interpretability. In future work, we aim to explore more advanced attention mechanisms and multi-task learning setups to further refine skeletal maturity assessments. Our findings pave the way for more reliable and scalable AI-driven orthopedic and forensic applications.

5. REFERENCES

[1] Thomas V Powell and Allan G Brodie, "Closure of the spheno-occipital synchondrosis," *The Anatomical Record*, vol. 147, no. 1, pp. 15–23, 1963.

[2] W. S. Al-Gumaei, R. Al-Attab, B. Al-Tayar, and et al., "Comparison of spheno-occipital synchondrosis maturation stages with three-dimensional assessment of mandibular growth," *BMC Oral Health*, vol. 22, no. 1, pp. 654, 2022.

[3] W. S. Al-Gumaei, H. Long, R. Al-Attab, and et al., "Comparison of three-dimensional maxillary growth across spheno-occipital synchondrosis maturation stages," *BMC Oral Health*, vol. 23, no. 1, pp. 100, 2023.

[4] Anwar Alhazmi, Eduardo Vargas, J Martin Palomo, Mark Hans, Bruce Latimer, and Scott Simpson, "Timing and rate of spheno-occipital synchondrosis closure and its relationship to puberty," *PLoS One*, vol. 12, no. 8, pp. e0183305, 2017.

[5] Natalie R Shirley and Richard L Jantz, "Spheno-occipital synchondrosis fusion in modern americans," *Journal of forensic sciences*, vol. 56, no. 3, pp. 580–585, 2011.

[6] Tiziano Baccetti, Lorenzo Franchi, and James A McNamara Jr, "The cervical vertebral maturation (cvm) method for the assessment of optimal treatment timing in dentofacial orthopedics," in *Seminars in Orthodontics*. Elsevier, 2005, vol. 11, pp. 119–129.

[7] Brent Hassel and Allan G Farman, "Skeletal maturation evaluation using cervical vertebrae," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 107, no. 1, pp. 58–66, 1995.

[8] Anwar Alhazmi, Mohammed Aldossary, J Martin Palomo, Mark Hans, Bruce Latimer, and Scott Simpson, "Correlation of spheno-occipital synchondrosis fusion stages with a hand-wrist skeletal maturity index: a cone beam computed tomography study," *The Angle Orthodontist*, vol. 91, no. 4, pp. 538–543, 2021.

[9] María José Fernández-Pérez, José Antonio Alarcón, James A McNamara Jr, Miguel Velasco-Torres, Erika Benavides, Pablo Galindo-Moreno, and Andrés Catena, "Spheno-occipital synchondrosis fusion correlates with cervical vertebrae maturation," *PLoS One*, vol. 11, no. 8, pp. e0161104, 2016.

[10] Nicolene Lottering, Donna M MacGregor, Clair L Alston, and Laura S Gregory, "Ontogeny of the spheno-occipital synchondrosis in a modern queensland, australian population using computed tomography," *American journal of physical anthropology*, vol. 157, no. 1, pp. 42–57, 2015.

[11] Kewal Krishan and Tanuj Kanchan, "Evaluation of spheno-occipital synchondrosis: A review of literature and considerations from forensic anthropologic point of view," *Journal of forensic dental sciences*, vol. 5, no. 2, pp. 72–76, 2013.

[12] Omid Halimi Milani, Lauren Mills, Amanda Nikho, Marouane Tliba, Veerasathpurush Allareddy, Rashid Ansari, Ahmet Enis Cetin, and Mohammed H Elnagar, "Automated classification of midpalatal suture maturation stages from cbcts using an end-to-end deep learning framework," *Scientific Reports*, vol. 15, no. 1, pp. 1–15, 2025.

[13] Omid Halimi Milani, Amanda N Nikho, Marouane Tliba, Lauren Mills, Ahmet Enis Cetin, and Mohammed H Elnagar, "Knowledge distillation approach for sos fusion staging: Towards fully automated skeletal maturity assessment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3473–3480.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[15] Richard B Bassed, Christopher Briggs, and Olaf H Drummer, "Analysis of time of closure of the spheno-occipital synchondrosis using computed tomography," *Forensic science international*, vol. 200, no. 1-3, pp. 161–164, 2010.

[16] Lin Wang and Kuk-Jin Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 3048–3068, 2021.

[17] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 11976–11986.

[19] Omid Halimi Milani, Amanda Nikho, Lauren Mills, Marouane Tliba, Ahmet Enis Cetin, and Mohammed H Elnagar, "Gradient attention map based verification of deep convolutional neural networks with application to x-ray image datasets," in *2025 IEEE 43rd VLSI Test Symposium (VTS)*. IEEE, 2025, pp. 1–5.