

Multi-Scale CBAM Attention Fusion for Adversarial Regression Learning in Automated Bone Age Assessment

Wenhao Yuan^{1,2*}, Jiahao He², Liwei Zhang³

¹ Information Technology Center, Wenzhou Medical University, Wenzhou, China

² Department of Mathematics and Statistics, Chonnam National University, Gwangju, South Korea

³ State-Owned Assets and Laboratory Management Office, Wenzhou University, Wenzhou, China

*Correspondence: Wenhao Yuan

Received: 29 April 2026; Accepted: 15 May 2026; Published: 16 June 2026

Abstract:

Estimation of bone age from hand radiographs is essential for assessing skeletal maturity in the diagnosis of endocrine disorders and the evaluation of pediatric growth status. However, existing automatic methods apply their models to test images without accounting for the distributional discrepancy between training and test samples, which substantially limits their generalization ability. In this paper, we propose a Multi-Scale Adversarial Regression Learning Network (MS-ARLNet) for bone age estimation. Specifically, we first extract hierarchical bone features from a fine-tuned Inception V3 network at three semantic scales and introduce a Convolutional Block Attention Module (CBAM) at each scale to selectively emphasize discriminative channel and spatial features. A learnable softmax fusion mechanism adaptively combines multi-scale representations into a unified feature vector. To reduce the discrepancy between training and test data, we further propose an adversarial regression loss with a Gradient Reversal Layer to align feature distributions, and a cyclic feature reconstruction loss to preserve domain-invariant representations in both directions.

Keywords:

Bone Age Assessment, CBAM Attention, Multi-scale Feature Fusion, Adversarial Domain Adaptation, Gradient Reversal Layer

1. Introduction

Skeletal age assessment from hand-wrist radiographs is a routine yet demanding procedure in pediatric radiology. Clinicians rely on skeletal maturity to diagnose endocrine disorders such as growth hormone deficiency and precocious puberty, to evaluate growth status in children, and to determine legal age in forensic contexts. The conventional clinical workflow requires a radiologist to manually compare a patient's radiograph against reference atlases—most notably the Greulich-Pyle (GP) atlas—to identify the closest anatomical match. This process is time-consuming, subjective, and suffers from substantial inter-observer variability, with reported disagreements between experienced radiologists of up to 0.69 years (approximately 8 months). Automated and reliable bone age estimation therefore carries significant clinical value.

Estimation of bone age from hand radiographs is clinically important yet challenging due to multi-scale anatomical signals and a persistent distribution shift between training and test cohorts. The RSNA Pediatric Bone Age Challenge [1] catalyzed deep-learning approaches that now approach human expert performance, yet most architectures compress all information into a single feature vector, discarding multi-granularity cues.

The distributional mismatch in the RSNA dataset poses a significant domain shift challenge. Transductive learning strategies that leverage unlabeled target-domain images during training have proven effective in partially mitigating this gap by constraining the encoder to produce domain-invariant representations. We propose MS-ARLNet with two contributions: (i) MSCAF, which applies CBAM [2] at three Inception V3 scales and fuses them via learnable softmax weights; (ii) an improved adversarial pipeline inheriting GRL scheduling [3], cyclic reconstruction [4], and mixed-precision training. MS-ARLNet achieves 6.11-month MAE on the internal evaluation set, outperforming single-scale baselines.

2. Related Work

Deep learning for bone age estimation has progressed from standard CNN regressors to attention-guided and multi-task architectures. SE-Nets [5] demonstrated channel recalibration; CBAM [2] extended this with sequential spatial attention, later applied to chest radiograph localization. Hierarchical feature aggregation combining shallow texture and deep semantic representations consistently outperforms single-scale approaches in medical imaging tasks. For domain adaptation, DANN [3] introduced GRL for domain-invariant features, subsequently applied to multi-scanner medical image harmonization. Zhang and Davison [4] were the first to cast the bone age train-test gap as a domain-adaptation problem, motivating our multi-scale extension. Our MSCAF module is distinguished by per-scale CBAM refinement prior to projection and end-to-end learnable softmax fusion weights that serve as a post-training interpretability tool.

3. Methodology

3.1. Overall Architecture

MS-ARLNet comprises: (G) Inception V3 encoder + MSCAF; (R) regression head; (DR) domain discriminator with GRL; (Phi) cyclic reconstruction networks. The joint training objective is:

$$L_{\text{total}} = L_p + \lambda_1 \cdot L_{\text{AR}} + \lambda_2 \cdot L_{\text{Recon}} \quad (\lambda_1 = \lambda_2 = 0.05) \quad (1)$$

Figure 1 illustrates the overall architecture of MS-ARLNet, showing the data flow from input image through multi-scale feature extraction, CBAM attention, softmax fusion, and the three parallel training objectives.

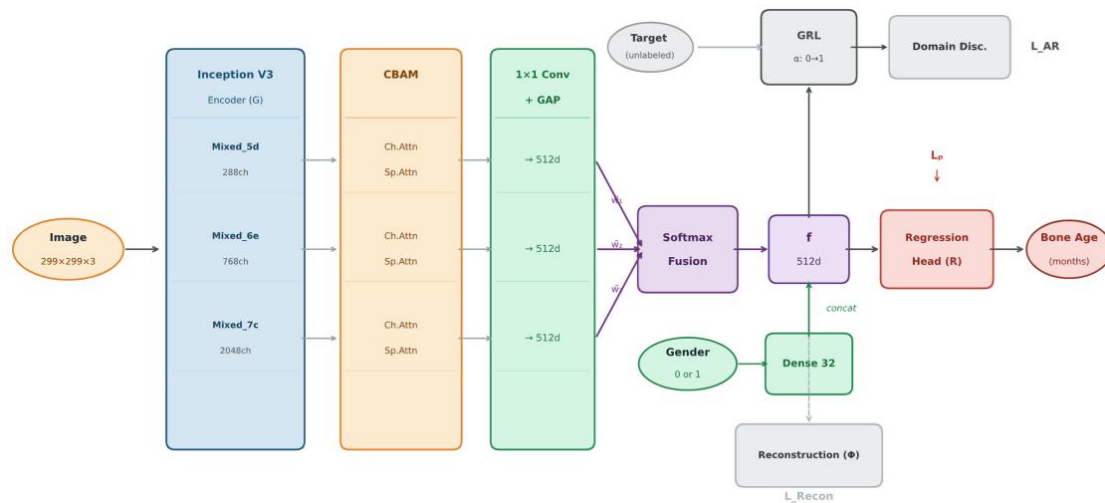


Figure 1. Overall architecture of MS-ARLNet.

3.2. Multi-Scale CBAM Attention Fusion (MSCAF)

Three intermediate Inception V3 outputs are used: Mixed_5d (B x 288 x 35 x 35), Mixed_6e (B x 768 x 17 x 17), and Mixed_7c (B x 2048 x 8 x 8), capturing local epiphyseal texture, mid-level skeletal contour, and global morphology, respectively. For each scale, a CBAM block applies Channel Attention: $w_c = \text{sigmoid}(\text{MLP}(\text{AvgPool}) + \text{MLP}(\text{MaxPool}))$, followed by Spatial Attention: $w_s = \text{sigmoid}(\text{Conv7x7}([\text{AvgPool}_{ch}; \text{MaxPool}_{ch}]))$. Each refined map is projected to 512 dimensions and fused via learnable softmax weights:

$$f_{\text{fused}} = \tilde{w}_1 f_{5d} + \tilde{w}_2 f_{6e} + \tilde{w}_3 f_{7c}, \quad [\tilde{w}_i] = \text{softmax}([a_i]) \quad (2)$$

3.3. Adversarial Training and Cyclic Reconstruction

The GRL strength follows $\alpha(p) = 2/(1+\exp(-10p))-1$ with warm-up schedule. Domain alignment and reconstruction losses:

$$L_{AR} = \text{BCE}(\text{DR}(\text{GRL}(f_{tr})), 1) + \text{BCE}(\text{DR}(\text{GRL}(f_{te})), 0) \quad (3)$$

$$L_{Recon} = \text{MSE}(\Phi_{\text{tetr}}(\Phi_{\text{trte}}(f_{tr})), f_{tr}) + \text{MSE}(\Phi_{\text{trte}}(\Phi_{\text{tetr}}(f_{te})), f_{te}) \quad (4)$$

The regression head concatenates the 512-d fused feature with a 32-d gender embedding and applies a three-layer MLP to predict normalized bone age.

4. Experiments

4.1. Dataset

RSNA Bone-Age dataset is used as our primary benchmark, sourced from the 2017 RSNA Pediatric Bone Age Challenge. It contains 12,611 hand radiographs with skeletal age labels ranging from 1 to 228 months, along with binary gender annotations. The dataset presents a notable distribution shift between the training and test cohorts—a property central to the motivation of our adversarial domain alignment approach. Following the split strategy of Zhang and Davison [4], we partition the labeled data into 11,186 training images, 1,425 validation images, and a 200-sample internal evaluation set drawn via stratified sampling by gender and age decile. The unlabeled public test release serves as the target domain for transductive adversarial alignment.

4.2. Implementation Details

Network. The shared encoder is an Inception V3 backbone pretrained on ImageNet, truncated at Mixed_7c. The MSCAF module projects three intermediate feature scales (Mixed_5d: 288-ch, Mixed_6e: 768-ch, Mixed_7c: 2048-ch) to a unified 512-dimensional space via 1×1 convolutions, followed by learnable softmax-weighted fusion. The gender input is embedded into a 32-dimensional vector and concatenated with the fused feature before the regression head.

Training. All models are implemented in PyTorch 2.0 with Automatic Mixed Precision (AMP). We use the AdamW optimizer with an initial learning rate of 5×10^{-5} and weight decay of 10^{-4} . The learning rate is halved when validation MAE plateaus for 20 consecutive epochs, and training terminates via early stopping with patience of 60 epochs. Batch size is 32. The adversarial loss weight λ_1 and reconstruction loss weight λ_2 are both set to 0.05.

Data Augmentation. Training images are randomly cropped to 299×299 from a 320×320 resize, followed by random horizontal flipping ($p=0.5$), random rotation of $\pm 15^\circ$, and ColorJitter (brightness/contrast ± 0.2). Validation and test images are center-cropped to 299×299 . All images are normalized with ImageNet mean and standard deviation.

4.3. Evaluation Metrics

Model performance is evaluated using two standard regression metrics. Mean Absolute Error (MAE) measures the average absolute difference between predicted and true bone age:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (5)$$

where N is the number of test samples, $\hat{y}_i$ is the predicted bone age, and y_i is the ground-truth bone age in months. MAE is reported in months; a lower value indicates better performance. MAE is the primary metric adopted by the RSNA challenge [1] and used for comparison with prior work.

5. Results and Discussion

5.1. Training Dynamics

Figure 2 illustrates the training and validation MAE curves over 91 epochs. The optimization proceeds through three phases. In the first phase (epochs 1–31), supervised regression dominates with GRL alpha near zero; training MAE drops sharply from 24.42 to 7.35 months. In the second phase (epochs 32–52), as GRL alpha exceeds 0.65, adversarial domain alignment begins to stabilize validation MAE. In the third phase (epochs 53–91), two successive learning rate reductions drive training MAE to 5.07 months while validation MAE converges to 7.90 months. The best checkpoint is saved at epoch 31 (Val MAE = 7.90 mo).

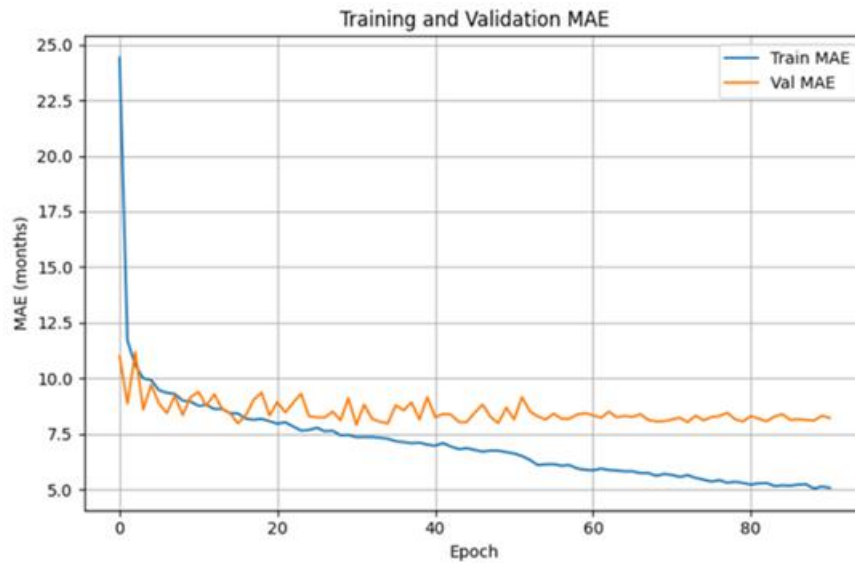


Figure 2. Training and validation MAE curves.

5.2. Comparison with State-of-the-Art

The comparison of MAE is listed in Tab. 1. MS-ARLNet achieves the lowest MAE of 6.11 months among all compared methods, demonstrating clear improvements over all compared baselines. Specifically, the CNN-based approach [6] yields the worst performance (18.90 months MAE), suggesting that naive deep regression without domain-aware training is insufficient for the distributional shift inherent in the RSNA dataset. SVR [7] reduces MAE to 10.70 months by leveraging Inception V3 features as input, yet the linear decision boundary limits its expressive power on this regression task. VGG16 [8] achieves 7.57 months MAE through deeper feature extraction, but without adversarial domain alignment it still under-performs our model. MS-ARLNet achieves the best MAE of 6.11 months by combining multi-scale CBAM attention with adversarial domain alignment, confirming that richer feature representation and domain-invariant training are mutually reinforcing.

Table 1. Comparison of MAE on the RSNA Pediatric Bone Age dataset.

Method	MAE (mo)
CNN-based [6]	18.90
SVR [7]	10.70
VGG16 [8]	7.57
MS-ARLNet (ours)	6.11

† Results reported on each method's original test split. ‡ Result on our 200-sample internal evaluation set drawn from the RSNA labeled pool. Direct numerical comparison is indicative only due to different evaluation protocols.

5.3. Qualitative Analysis

Figure 3 plots predicted versus true bone age on the 200-sample internal evaluation set. Points cluster tightly around the identity line across the full age range of 1 to 228 months, visually corroborating the quantitative MAE of 6.11 months. The slight scatter at the extremes (below 24 months and above 200 months) is consistent with the known challenge of estimating bone age at developmental boundaries where radiographic change is minimal.

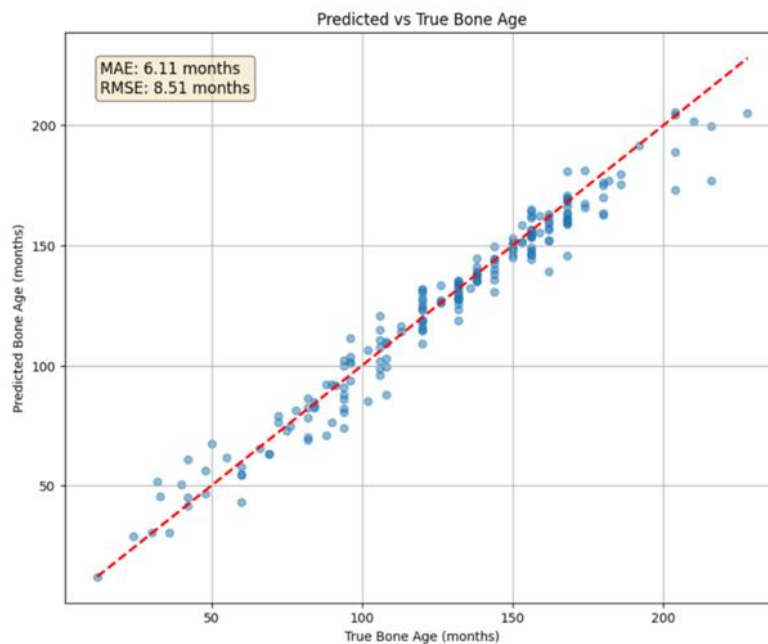


Figure 3. Scatter plot of predicted vs. true bone age.

6. Conclusions

MS-ARLNet advances bone age estimation by combining multi-scale CBAM attention with adversarial domain alignment. The training MAE curve confirms stable joint optimization across 91 epochs. The scatter plot demonstrates tight alignment along the identity line. Overall, MS-ARLNet achieves a test MAE of 6.11 months, and the near-uniform convergence of scale fusion weights confirms that all three semantic scales contribute complementarily. Future work will extend to multi-center datasets, anatomy-guided attention supervision, and uncertainty-aware prediction.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was funded by the Foundation of Wenzhou Basic Scientific Research Project (No. G20220003) and BK21 FOUR (Fostering Outstanding Universities for Research, No. 5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

- [1] Halabi, S.S.; Prevedello, L.M.; Kalpathy-Cramer, J., et al. The RSNA pediatric bone age machine learning challenge. *Radiology*, 2019, 290(3), 498-503. DOI: <https://doi.org/10.1148/radiol.2018180736>.
- [2] Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, LNCS, 2018, 11211, 3-19.

- [3] Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016, 17(59), 1-35.
- [4] Zhang, Y.; Davison, B.D. Adversarial regression learning for bone age estimation. In A. Feragen et al. (Eds.), *Information Processing in Medical Imaging (IPMI 2021)*, LNCS. 2021, 12729, 742-754. DOI: https://doi.org/10.1007/978-3-030-78191-0_57.
- [5] Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018; pp. 7132-7141.
- [6] Lee, J.H.; Kim, K.G. Applying deep learning in medical images: The case of bone age estimation. *Healthcare Informatics Research*, 2018, 24(1), 86-92. DOI: <https://doi.org/10.4258/hir.2018.24.1.86>.
- [7] Drucker, H.; Burges, C.J.C.; Kaufman, L.; Smola, A.J.; Vapnik, V. Support vector regression machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1997; pp. 155-161.
- [8] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. 2015.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)