This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for RASH: Rheumatoid Arthritis Synthetic Hand Dataset for Joint-Level Inflammation Estimation

Rintaro Chiba, Shun Kato, Daiki Kitsunai, Yasushi Kondo, Yoshimitsu Aoki, and Mariko Isogawa

Keio University, Yokohama, Japan
rinrin819@keio.jp, mariko.isogawa@keio.jp

1 Overview of the Supplementary Material

This supplementary material complements the main paper. It provides an extended discussion of related work (Sec. 2), implementation details of the RGE-LA module (Sec. 3), the full hyperparameter configuration – separating values tuned automatically from those fixed a priori (Sec. 4), a sensitivity analysis under different hyperparameter settings (Sec. 5), and examples of images excluded during dataset preparation (Sec. 6).

2 Related Work

2.1 RA Assessment Using Specialized Medical Imaging Modalities

Conventional Rheumatoid Arthritis (RA) assessment relies on imaging modalities such as X-ray, MRI, and ultrasonography to complement clinical examination. X-ray is used for standardized scoring of structural damage, including joint space narrowing and bone erosions, and is suitable for longitudinal monitoring in clinical practice [1–3]. However, X-ray reflects accumulated structural changes rather than active inflammation and is therefore insensitive to early synovitis. MRI enables sensitive visualization of soft-tissue inflammation such as synovitis and bone marrow edema, and detects early lesions that precede irreversible joint damage [4, 5]. Ultrasonography is increasingly adopted in routine care to assess peripheral joints, as it captures synovial hypertrophy and vascular signals associated with active inflammation. Studies explore CNN-based automation of ultrasound scoring and demonstrate the feasibility of learning inflammation-related patterns even from limited annotated datasets [6].

Nevertheless, these modalities require dedicated equipment, controlled acquisition protocols, and expert operation and interpretation, which constrain scalability for frequent screening and out-of-clinic deployment.

2.2 RGB-based RA Screening

To improve accessibility, several studies investigate RA screening from RGB hand photographs captured by smartphones.

Pathak et al. [8] propose CNN-based pipelines for both whole-hand screening and joint-level classification using peri-joint crops extracted from standardized smartphone images. Their framework demonstrates that visual signs of inflammation can be detected from RGB photos with reasonable performance for the wrist and selected PIP joints, providing an important proof of concept for point-of-care screening. However, the evaluation is limited to a subset of joints, and performance is reported under relatively controlled acquisition conditions.

Kato et al. [7] further advance joint-level RGB-based assessment by introducing a dual-encoder architecture that integrates both global hand context and local joint appearance. Their method extracts keypoint-centered joint patches and processes them together with the full hand image using a DINO-pretrained ResNet18 backbone, enabling the fusion of global and local representations for joint-wise inflammation classification. This design improves the utilization of contextual information while maintaining sensitivity to local joint characteristics, demonstrating the feasibility of joint-level RA assessment from RGB images under limited annotated data. However, their evaluation mainly focuses on MCP joints, and the dependencies among multiple joints within the same hand are not explicitly modeled.

Beyond hand photographs, other imaging modalities have also been explored for accessible RA screening. Fayyaz et al. [9] apply classical digital image processing (based on the Pythagorean theorem) to detect joint-space changes indicative of RA from hand images. Ahalya et al. [10] instead use hand *thermal* images, proposing RANet, a custom CNN combined with a quanvolutional layer, to detect RA-related inflammation from surface temperature patterns. These approaches confirm the broader interest in low-cost, RGB-or-thermal-camera-based RA screening, but neither addresses joint-level inter-dependency or the severe class imbalance central to our setting.

Overall, prior RGB-based approaches demonstrate the feasibility of joint-level RA screening from hand photographs and establish important baselines. However, existing methods remain limited in joint coverage, inter-joint context modeling, and robustness to appearance variation in real-world settings.

2.3 RA Datasets

Public datasets for RA image analysis remain scarce and are predominantly built around X-ray radiographs rather than RGB photographs. Bo et al. [11] review and benchmark deep learning models for automating the modified Sharp/van der Heijde scoring of hand radiographs, highlighting the reliance on structural-damage scoring rather than joint-level active-inflammation labels. More recently, Yang et al. [12] introduced RAM-W600, a multi-task wrist radiograph dataset with joint-level annotations, benchmarking several detection and classification tasks. While valuable, these radiograph-based datasets characterize structural

damage rather than the RGB-visible inflammatory cues (erythema, swelling, gloss) that our framework targets, and none provide the joint-level, severity-controllable, skin-tone-diverse synthetic supervision that RASH offers. This gap motivates our synthetic data generation approach.

3 Detailed architecture of the RGE-LA module

This section provides implementation details of the lightweight CNN used in the RGE-LA module described in Sec. 2.2 of the main paper. In the main text, RGE-LA takes the cue maps and RGB patch to form a 6-channel input and uses a lightweight CNN with three parallel cue-specific branches to produce attention maps.

Specifically, the CNN operates on the cue-concatenated input $[R_{j,k}, G_{j,k}, E_{j,k}]$ and $[x_R, x_G, x_B]$, which combines the three cue maps with the RGB channels of the patch. The network consists of three parallel branches corresponding to redness, gloss, and edge cues. Each branch produces one attention map specialized for the corresponding inflammatory cue. Each branch contains two convolutional layers with 64 and 32 channels, respectively, using 3×3 kernels and ReLU activation. The resulting feature maps are then upsampled using bilinear interpolation to match the spatial resolution of the target backbone feature map.

After sigmoid activation, each attention map is further processed using adaptive histogram equalization to enhance local contrast, improving the visibility of subtle inflammatory regions.

The three enhanced attention maps are then aggregated through a learned weighted summation:

$$\boldsymbol{a}_{j,k} = w_R \cdot \sigma(R'_{j,k}) + w_G \cdot \sigma(G'_{j,k}) + w_E \cdot \sigma(E'_{j,k}), \tag{1}$$

where w_R, w_G, and w_E are learnable scalar weights initialized to $1/3$, and $\sigma(\cdot)$ denotes the sigmoid function.

The resulting integrated attention map is normalized to the range $[0, 1]$ and used to modulate the intermediate feature map as described in Sec. 2.2 of the main paper. This branch-wise architecture allows each branch to specialize in detecting its corresponding cue, while the learnable weights adaptively balance their contributions during end-to-end training.

4 Implementation Details

All methods were evaluated under the same protocol using patient-level 5-fold cross-validation (no patient's images appear in more than one split within a fold) repeated 30 times. In each fold, a validation set was split from the training data to tune hyperparameters and the decision threshold τ, and evaluation was performed on a held-out test set. We report mean test performance averaged over 5 folds and 30 trials. Training consisted of synthetic pretraining (batch size 32, up to 20 epochs) followed by up to 5 epochs of fine-tuning on real data, with

Table 1. Hyperparameters tuned automatically on the validation split via Optuna's TPE sampler (15 trials/fold). "Used" is the value selected for the run reported in the main paper.

Hyperparameter	Symbol	Search range	Used
Learning rate	–	$[8 \times 10^{-5}, 3 \times 10^{-4}]$ (log)	2.52×10^{-4}
Weight decay	–	$[5 \times 10^{-4}, 3 \times 10^{-3}]$ (log)	1.13×10^{-4}
Balanced MixUp (pretrain)	α	$[0.28, 0.45]$	0.36
Balanced MixUp (fine-tune)	α_{ft}	$[0.20, 0.45]$	0.36
Focal Loss focusing param.	γ	$[0.90, 1.60]$	1.19
CDA loss weight	λ_{cda}	$[0.10, 0.28]$	0.18

Table 2. Hyperparameters fixed a priori (pilot experiments / computational constraints).

Hyperparameter	Symbol	Value
GNN hidden dimension	–	320
GNN dropout	–	0.49
Camera perturbation range (pitch/yaw)	θ	15°
RGE-LA modulation strength	α	0.2
Joint patch size	$W \times H$	256×256
CAM masking ratio	N	50%

early stopping (on validation loss for fine-tuning, and on validation G-mean for pretraining) given the limited size of the real fine-tuning split (∼1,800 patches). We used AdamW [13] with horizontal flips, rotations, and color jitter for data augmentation. Experiments were conducted on Ubuntu 22.04 using an NVIDIA RTX 6000 Ada Generation GPU.

As described in Sec. 3.1 of the main paper, hyperparameters are separated into those tuned automatically on the validation split and those fixed a priori. Table 1 lists the search space used for automatic tuning (Optuna [15], Tree-structured Parzen Estimator sampler, 15 trials per fold) together with the resulting value used to obtain the main paper's reported results. Table 2 lists the hyperparameters fixed a priori, based on pilot experiments and computational constraints (e.g., the GNN hidden dimension and CAM masking ratio were capped to keep training time and GPU memory tractable given the 5-fold×30-repeat protocol), together with their fixed values. The decision threshold τ is not a model hyperparameter in this sense: it is re-tuned on the validation split of every fold/trial to maximize G-mean, and we report its mean value across folds and trials for reference.

5 Results under different hyperparameter settings

We further analyze the sensitivity of the proposed framework to two key hyperparameters that were fixed a priori (Table 2): the modulation strength α in RGE-LA and the Top-N% ratio used to generate the binary CAM mask in CDA.

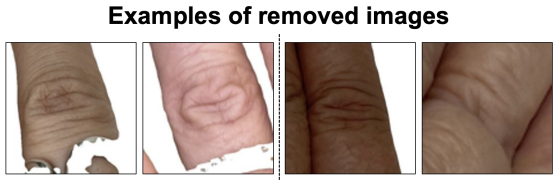


Fig. 1. Examples of removed images from the real dataset. Left: images containing unintended objects (e.g., rings). Right: images where the joint region is not properly isolated.

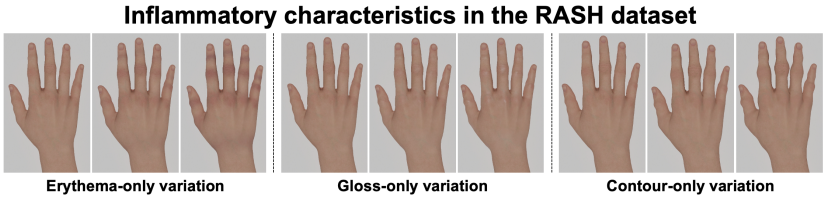


Fig. 2. Examples of inflammatory appearance variations in the RASH dataset. From left to right: erythema-only variation, gloss-only variation, and contour-only variation.

Specifically, the modulation strength α, which controls the feature modulation in RGE-LA, is varied from the default value of 0.2 to 0.1, 0.3, and 0.5. In addition, the Top-N% ratio used for CAM-based region selection in CDA is changed from the default value of 50 to 20 and 80.

The corresponding results are summarized in Table 3. Overall, the default configuration provides the most balanced performance, while excessively small or large values tend to slightly degrade the results. These observations indicate that moderate attention modulation and a moderate lesion-region selection ratio are important for stable joint-level inflammation estimation, and support fixing them a priori rather than including them in the automatic search of Table 1 (which would substantially increase the search space and computational cost under our 5-fold$\times$30-repeat protocol).

6 Examples of Removed Images

During the preparation of the real dataset, 96 images were excluded to ensure reliable joint-level analysis. The removed samples mainly fall into two categories, as illustrated in Fig. 1.

First, some images contain unintended objects that appeared during image acquisition, such as rings or other items around the finger region. These objects interfere with the visual appearance of the joint and may introduce misleading visual cues.

Second, several images fail to properly capture a single target joint region. In these cases, other fingers are partially visible, or the framing does not sufficiently

Table 3. Results under different hyperparameter settings. We vary the modulation strength α in RGE-LA and the Top-N% ratio used for CAM mask generation in CDA. **Bold** denotes the best result.

Setting	Accuracy[%] ↑	Sensitivity[%]↑	Specificity[%]↑	F1↑	Gmean↑
$\alpha = 0.1$	0.648	0.552	0.702	0.380	0.622
$\alpha = 0.2$ (default)	**0.663**	0.584	0.677	**0.392**	**0.628**
$\alpha = 0.3$	0.652	0.598	0.648	0.388	0.622
$\alpha = 0.5$	0.647	**0.623**	0.610	0.386	0.616
$N = 20$	0.650	0.548	**0.706**	0.381	0.622
$N = 50$ (default)	**0.663**	0.584	0.677	**0.392**	**0.628**
$N = 80$	0.646	0.612	0.620	0.385	0.616

isolate the intended joint area. Such images make it difficult to perform consistent joint-level analysis.

Removing these images improves the overall quality of the real dataset and reduces potential noise during model training and evaluation.

References

1. Mate, G.S., Kureshi, A.K., Singh, B.K.: An efficient CNN for hand X-ray classification of rheumatoid arthritis. Journal of Healthcare Engineering **2021**, Article ID 6712785 (2021). https://doi.org/10.1155/2021/6712785
2. Perez-Siguas, R., Matta-Solis, E., Remuzgo-Artezano, A., Matta-Solis, H., Matta-Perez, H., Perez-Siguas, L.: Hand X-ray image processing system for the early detection of rheumatoid arthritis. In: 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–4. IEEE (2023)
3. Huo, Y., Vincken, K.L., Viergever, M.A., Lafeber, F.P.: Automatic joint detection in rheumatoid arthritis hand radiographs. In: 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 125–128. IEEE (2013). https://doi.org/10.1109/ISBI.2013.6556428
4. Ostendorf, B., Peters, R., Dann, P., Becker, A., Scherer, A., Wedekind, F., Schneider, M.: Magnetic resonance imaging and miniarthroscopy of metacarpophalangeal joints: sensitive detection of morphologic changes in rheumatoid arthritis. Arthritis Rheum. **44**(11), 2492–2502 (2001)
5. Rubin, D.A.: MRI and ultrasound of the hands and wrists in rheumatoid arthritis. I. Imaging findings. Skeletal Radiol. **48**(5), 677–695 (2019)
6. Tang, J., Jin, Z., Zhou, X., Zhang, W., Wu, M., Shen, Q., Yuan, J.: Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data. Chinese Physics B **28**(3), 038701 (2019)
7. Kato, S., Kondo, Y., Saito, S., Aoki, Y., Isogawa, M.: Image-based joint-level detection for inflammation in rheumatoid arthritis from small and imbalanced data. arXiv preprint arXiv:2602.14365 (2026)
8. Pathak, S.: Deep learning approach to early detection of rheumatoid arthritis. In: IEEE Integrated STEM Education Conference (ISEC), pp. 1–2 (2024)
9. Fayyaz, S., Gondal, H.A., Khan, M.A., Sohail, A., Akram, Z.: Digital image processing based detection of rheumatoid arthritis using Pythagorean theorem. In: International Conference on Open Source Systems and Technologies (ICOSST), pp. 1–6 (2020)
10. Ahalya, R.K., et al.: RANet: a custom CNN model and quanvolutional neural network for the automated detection of rheumatoid arthritis in hand thermal images. Scientific Reports **13**(1), 15638 (2023)
11. Bo, Z., Coates, L.C., Papież, B.W.: Deep learning models to automate the scoring of hand radiographs for rheumatoid arthritis. In: Proceedings of the Annual Conference on Medical Image Understanding and Analysis (MIUA). Springer, Cham (2024)
12. Yang, S., et al.: RAM-W600: A multi-task wrist dataset and benchmark for rheumatoid arthritis. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
14. Galdran, A., Carneiro, G., González Ballester, M.A.: Balanced-mixup for highly imbalanced medical image classification. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, pp. 323–333. Springer, Cham (2021)
15. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the ACM SIGKDD In-

ternational Conference on Knowledge Discovery & Data Mining (KDD), pp. 2623–2631 (2019)