

Hierarchical Cross-Modal Fusion with Modality-Specific Training Paradigms for High-Precision Non-Invasive Aging Assessment

Yihang Wu¹, Junnan Xu¹, Guocong Li¹, Yiwen Zhang¹, Bo Cao²(✉), and Shuaihan Huang¹(✉)

¹ Zhejiang University, Hangzhou, China
huangshuaihan@outlook.com

² Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
init_cb@163.com

Abstract. Accurate and clinically relevant non-invasive aging assessment remains a key challenge in gerontology. While computer vision enables functional phenotypic analysis, existing methods primarily rely on unimodal analysis, struggle to encapsulate the systematic and complex nature of the aging process. We propose a hierarchical cross-modal fusion framework that integrates three complementary aging-related modalities: optical coherence tomography angiography (OCTA) and optical corneal thickness (OCT) from the ocular modality, millisecond-level behavioral syllables from the behavioral modality, and a high-dimensional gait matrix from the gait modality. Our framework introduces specialized training paradigms for each modality, including a three-stage progressive self-masked training paradigm, an active learning behavioral representation framework, and a multi-stage gait quantization workflow. This fusion architecture systematically progresses from unimodal encoding and intra-domain alignment of OCTA-OCT to inter-domain feature enhancement, and employs an entropy-guided probabilistic weighting strategy to integrate peripheral biological structural changes with central functional outputs. We performed aging assessments using a novel dataset of 166 mice aged 8 weeks to 18 months, divided into four age groups, with concurrent ocular, behavioral, and gait measurements. Experimental results demonstrate that our method achieves an accuracy (ACC) of 92.31% and an area under the curve (AUC) of 97.73% in aging assessment, significantly outperforming both single-modal baseline methods and state-of-the-art approaches. Furthermore, it exhibits superior capability in the precise extraction of age-related features. This work establishes a clinically interpretable framework for non-invasive aging assessment by bridging structural and functional changes at different biological scales. Code is available at <https://github.com/hangxingwu/HCMFMSTP>.

Keywords: Aging Prediction · Cross-Modal Fusion · Ocular Imaging · Behavioral and Gait Representations · Hierarchical Integration

1 Introduction

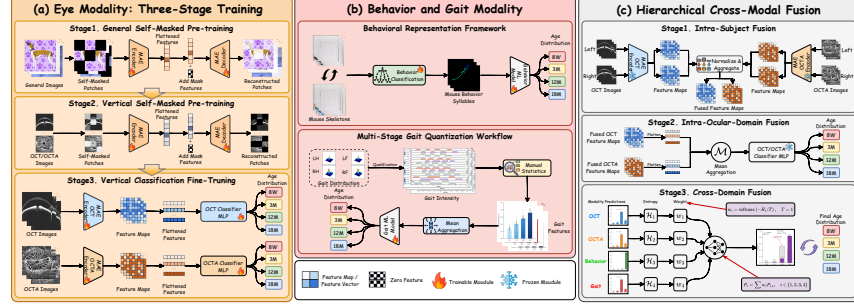


Fig. 1: Overview of the cross-modal training and fusion inference framework.

Accurately assessing individual aging is a key challenge in gerontological research. Traditional methods are limited by the invasiveness of molecular biomarker analysis and the subjectivity of clinical assessment[18]. Recent advances in computer vision and cross-modal fusion have made non-invasive aging assessment through functional phenotype analysis possible. However, existing methods primarily focus on single-modal analysis of ocular imaging or dynamic behavior, lacking a refined cross-modal alignment framework that effectively utilizes the systematic and complex nature of the aging process[8].

Behavior serves as a crucial modality externalizing individual aging. The rapid progression of pose recognition and behavior classification techniques has furnished robust tools for aging research centered on functional behaviors. Currently, the research focus in pose recognition has shifted from early autoregressive methods such as DeepPose and MTCNN to confidence map-based methods such as AlphaPose[16,19,5]. Simultaneously, behavior classification techniques have evolved from video sequence-based models like MotionMapper and DBscorer to more refined models based on skeleton sequences such as VAME[13,15,11]. Recent research, such as KeypointMoSeq[17], has further achieved frame-level high-precision behavior classification.

The eyeball is another important modality for externalizing individual aging, but existing methods are mostly limited to a single modality, resulting in limited prediction accuracy. LensAge and eyeAge[9,1] have established correlations between ocular features and biological age. FLEX[12] has proposed a bidirectional transfer learning framework based on the Swin Transformer to predict organ aging status. VisionFM and RETFound[14,20] have further enhanced model performance via large-scale pre-training on ocular medical images. However, these methods are limited to single image modality information, and existing cross-modal fusion frameworks typically adopt late-stage feature concatenation, ignoring hierarchical relationships between modalities and sacrificing complementary information available in early processing stages.

To overcome these limitations, we propose a hierarchical cross-modal fusion framework specifically for aging representation and prediction. As shown in Fig 1, our method introduces modality-specific training paradigms that respect the domain features of each modality while establishing cross-modal relationships through progressive fusion. For the ocular modality, we have developed a self-masked dual-stream encoder-decoder architecture combined with domain-adaptive masking prediction, which effectively extracts anatomical features from limited labeled data while modeling optical corneal thickness (OCT) and optical coherence tomography (OCTA) features. For the behavioral modality, we have implemented an active learning-enhanced pose estimation framework combined with behavioral syllable parsing to detect age-related millisecond-level behavioral changes. For the gait modality, we designed a multi-stage quantization workflow to transform the raw footprint distribution into a high-dimensional, clinically interpretable gait intensity matrix. Crucially, our fusion architecture starts with single-modality encoding, proceeds through intra-domain OCTA-OCT alignment, and finally achieves inter-domain feature enhancement. To this end, we employ a probability distribution-weighted fusion strategy based on feature distribution entropy to integrate peripheral structural features with central functional features.

The main contributions of this paper are as follows: (1) To address the scarcity of cross-modal mouse aging data, we construct the first synchronized ocular-behavioral-gait dataset for mouse aging assessment. The dataset was collected over one year in three batches and includes 166 surviving mice from approximately 200 initially enrolled animals, covering four age groups: 8 weeks, 3 months, 12 months, and 18 months. (2) We propose an aging characterization and prediction framework based on cross-modal biological signal fusion, which synergistically integrates three complementary modalities closely related to the degree of aging. (3) Extensive experiments demonstrate that our method significantly outperforms existing methods and single-modal methods in all aspects, effectively extracting fine-grained features from aging-sensitive regions such as the retinal vascular network and corneal epithelium.

2 Methods

2.1 Cross-modal Data Collection and Preprocessing

We constructed a synchronized cross-modal mouse aging dataset covering four age groups: 8 weeks, 3 months, 12 months, and 18 months. Data collection lasted one year and was conducted in three independent batches. Approximately 200 mice were initially enrolled; due to anesthesia- and breeding-related mortality, 166 surviving mice were included in the final analysis. As shown in Fig 2(a), for ocular imaging, each mouse was anesthetized for 10–15 minutes, and both left and right eyes were repeatedly acquired 3–5 times for manual quality screening. This process ultimately yielded 324 high-quality OCT and OCTA images. Each image was annotated with an ROI bounding box, and data augmentation, including random scaling, rotation, and elastic transformation, was used to improve generalization under the small-sample setting. As shown in Fig 2(b), for behavioral analysis,

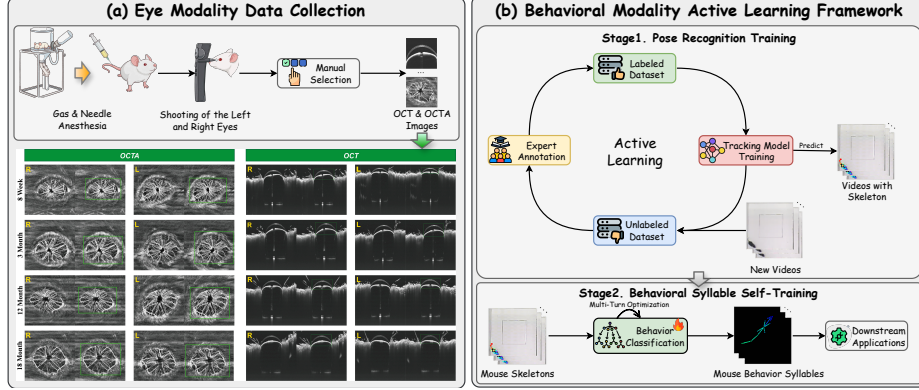


Fig. 2: Overview of the cross-modal data collection workflow: (a) shows the collection pipeline for ocular modality data. (b) shows the framework for constructing and processing behavioral modality data.

we collected 5 minutes of video at a frequency of 25Hz from the open field experiment and processed it using the DeepLabCut posture recognition model based on active learning. Low-confidence frames with a confidence level less than 0.9 were iteratively labeled by experts. Every 10 hours of new video data collected, 5% of the frames with the lowest confidence levels were re-labeled, which enabled the model to achieve an accuracy of over 90%. For gait analysis, we used a mirror-based optical system to record mouse gait. The data from 5 independent tests for each mouse were quantified into a 10×10 grid four-dimensional gait intensity matrix, and three experts independently verified key indicators such as standing time, braking time, and propulsion time.

2.2 Modality-Specific Feature Learning

As shown in Fig 1(a), for ocular imaging modalities, we propose a self-masking-driven three-stage progressive training paradigm, which effectively solves the problem of difficult fine-grained feature extraction. In the first stage, we employed a Transformer encoder-decoder architecture for self-masking pre-training on 10,000 general-domain images, with 40%–60% of the image regions randomly masked. The reconstruction loss was defined as follows:

$$\mathcal{L}_{\text{recon}} = \|\text{Decoder}(\text{Encoder}(x_{\text{masked}})) - x_{\text{unmasked}}\|_2^2 \quad (1)$$

In the second stage, we transferred the pre-trained model to medical image characteristics, performing domain-adaptive fine-tuning using OCTA and CT data. Finally, in the third stage, we split OCTA and CT into independent branches, extracting modality-specific features through the pre-trained encoder from the previous stage, and feeding them into a multilayer perceptron for aging assessment. As shown in Fig 1(b), for behavioral modalities, we implemented

an active learning-enhanced pose estimation framework based on the ResNet152 backbone network. Subsequently, Keypoint-Moseq was adopted for behavioral syllable classification, and continuous trajectories were decomposed into discrete syllables through variational Bayesian inference. For gait modalities, we quantized the footprint distribution into a four-dimensional gait intensity matrix using a mirror-based optical system, and extracted key indicators such as standing time, braking time, and propulsion time through manual statistics. These indicators were aggregated to form high-dimensional structured features.

2.3 Hierarchical Cross-Modal Fusion

As shown in Fig 1(c), our hierarchical fusion architecture performs hierarchical fusion at three levels instead of using simple feature concatenation. First, intra-subject fusion aggregates the feature maps of the corresponding left and right eyes within the same ocular modality. Specifically, the feature maps are normalized and aggregated to reduce acquisition noise and binocular variability:

$$\mathbf{f}_m = \text{Agg}(\text{Norm}(\mathbf{f}_m^L), \text{Norm}(\mathbf{f}_m^R)), \quad m \in \{\text{OCT}, \text{OCTA}\}. \quad (2)$$

Second, intra-ocular-domain fusion integrates OCT and OCTA representations into a shared ophthalmic-domain embedding. Since the OCT and OCTA encoders have been jointly adapted through the self-masked fine-tuning stage, their output embeddings lie in a compatible latent space. The aggregation module $\mathcal{M}$ therefore combines structural and vascular aging cues at the flattened feature-vector level:

$$\mathbf{f}_{\text{ocular}} = \mathcal{M}(\mathbf{f}_{\text{OCT}}, \mathbf{f}_{\text{OCTA}}). \quad (3)$$

The MLP classifiers following this representation further exploit the complementary information between corneal structural changes and vascular remodeling. Third, cross-domain fusion integrates the prediction distributions from ocular, behavioral, and gait modalities using entropy-guided probabilistic weighting. For the i -th modality, the entropy of its prediction distribution is computed as

$$H_i = - \sum_{c=1}^C p_{i,c} \log p_{i,c}, \quad (4)$$

and the modality weight is obtained by

$$w_i = \frac{\exp(-H_i/T)}{\sum_{j=1}^N \exp(-H_j/T)}. \quad (5)$$

We set $T = 1$ to preserve the original confidence distribution. The final probability of age class c is then computed as

$$P_c = \sum_{i=1}^N w_i p_{i,c}, \quad (6)$$

where i denotes OCT, OCTA, behavior, or gait. This design assigns larger weights to more confident modalities and enables adaptive integration of peripheral structural degeneration and central functional decline.

3 Experiments

3.1 Experimental Setup

Dataset. Our experiments are based on the dataset collected in Section 2.1, which contains 166 mice classified into four different age groups, with the training, validation, and test sets split in a 7:1:2 ratio. The dataset includes synchronously acquired data across three modalities: ocular, behavioral, and gait. For the ocular modality, we collected OCTA and OCT images. For the behavioral modality, we recorded 5 minutes of behavioral video for each mouse at a sampling rate of 25 Hz. Pose estimation was performed using the DeepLabCut framework with ResNet152 as the backbone network. This model was trained through active learning, with a confidence threshold of 0.90, and low-confidence frames were selectively relabeled. For the gait modality, after five repeated measurements per mouse, the heatmaps were aggregated into a four-dimensional gait intensity matrix that comprehensively characterizes gait patterns.

Implementation Details. All methods were implemented with the Adam optimizer with an initial learning rate of 5×10^{-5} , layer decay coefficient of 0.65, weight decay coefficient of 0.05, dropout rate of 0.1, and batch size of 16. Images were resized to 224×224 pixels and trained for 500 epochs on four NVIDIA H20 GPUs. For ocular models, we performed 3-fold cross-validation and reported the averaged results. For behavioral and gait modalities, we evaluated more than 10 classical machine-learning models under 5-fold cross-validation and reported the best averaged cross-validation performance as the representative non-image baseline. Ours-B/L/H use ViT-B/L/H-scale ocular encoders and a fixed 8-block, 512-width decoder, all randomly initialized.

Evaluation Metrics. All models were evaluated using standard metrics including Accuracy (ACC), Area Under the Curve (AUC), F1-score (F1), Precision (PRE), and Sensitivity (SEN). Five-fold cross-validation was adopted to ensure statistical robustness of the models in aging prediction across the four age groups.

3.2 Main Results

To verify the effectiveness of the proposed self-masking-driven three-stage training paradigm in characterizing aging in ocular medical images, we compared 17 mainstream backbone network architectures and two advanced pre-training methods on both OCTA and OCT modalities[7,4,2,10,3,6].

As shown in Table 1, CNN-based models exhibited significant performance fluctuations across the two modalities, while Transformer-based architectures showed significantly lower accuracy when directly transferred to the medical imaging domain, reflecting a dual problem of significant domain shift and insufficient fine-grained feature extraction capability. Models such as Twins and DaViT, which introduce specific attention mechanisms, performed well on the OCT modality, but their performance on the OCTA modality was generally below 66%, indicating that simply improving the network architecture is insufficient to solve the problem of extracting fine-grained features unique to medical images, such

as vascular texture and inter-layer structure. Furthermore, pre-trained models from RETFound and VisionFM also exhibited the same problem, indicating that transfer learning cannot effectively address these inherent issues. In contrast, our method, through a three-stage progressive training paradigm, significantly outperformed suboptimal methods in all metrics on both OCTA and OCT modalities. It is noteworthy that the consistent performance across different scale variants indicates that the three-stage training has saturated the model’s performance at the current data scale, demonstrating that domain-adaptive pre-training is more efficient than simply scaling up the model. As shown in Fig 3, the visualization results further reveal that the attention mechanism of our method can precisely focus on aging-sensitive areas such as the retinal vascular network and corneal epithelium, while existing methods tend to focus on background noise areas and localized aging-sensitive areas.

Table 1: Performance on Eye Analysis Task. Best results are **bold**, second-best underlined.

Method	Param(M)	OCTA					OCT				
		ACC↑	AUC↑	F1↑	PRE↑	SEN↑	ACC↑	AUC↑	F1↑	PRE↑	SEN↑
ResNet18[7]	11.18	57.69	79.33	44.05	43.47	47.92	57.69	78.85	42.31	45.24	40.62
ResNet34[7]	21.29	61.54	82.81	61.74	62.61	61.46	69.23	76.19	45.14	44.11	46.88
ResNet50[7]	23.52	61.54	82.66	63.19	65.28	62.50	73.08	83.95	68.30	67.92	72.92
ResNet101[7]	42.51	61.54	87.97	57.99	57.74	61.46	73.08	87.52	66.99	75.00	63.54
ResNet152[7]	58.15	61.54	82.98	62.26	65.14	69.79	80.77	82.90	75.42	73.51	79.17
ViT-B[4]	89.83	61.54	74.04	52.27	53.27	52.08	65.38	68.14	45.00	45.39	45.83
ViT-L[4]	306.46	57.69	78.81	48.81	52.02	51.04	65.38	77.84	45.60	43.93	47.92
Twins-SVT[2]	98.25	61.54	75.69	41.31	43.33	41.67	80.77	95.11	75.42	73.51	79.17
Twins-PCPVT[2]	60.48	65.38	78.60	55.27	55.69	55.21	76.92	87.73	70.91	69.16	77.08
SwinTv2-T[10]	27.58	69.23	89.67	62.95	<u>71.15</u>	60.42	73.08	90.98	63.47	64.49	63.54
SwinTv2-S[10]	48.96	65.38	78.54	43.81	45.89	46.88	73.08	91.18	64.30	62.31	70.83
SwinTv2-B[10]	86.90	69.23	86.95	<u>68.06</u>	68.69	67.71	46.15	54.05	15.79	11.54	25.00
SwinTv2-L[10]	195.21	61.54	75.18	41.08	40.08	42.71	80.77	90.47	75.42	73.51	79.17
DaViT-T[3]	27.59	61.54	74.45	41.00	41.31	41.67	76.92	89.31	72.64	72.73	<u>80.21</u>
DaViT-S[3]	48.98	57.69	84.74	47.37	48.33	48.96	73.08	89.44	63.47	64.49	63.54
DaViT-B[3]	86.93	69.23	86.24	64.00	63.33	<u>70.83</u>	73.08	89.73	70.91	73.77	77.08
EVA[6]	303.12	61.54	82.15	36.31	35.04	38.54	76.92	92.24	73.22	75.30	79.17
RETFound[20]	303.31	49.23	45.90	21.44	20.41	27.71	72.31	89.89	68.93	69.97	68.75
VisionFM[14]	85.80	46.15	50.52	15.79	11.54	25.00	57.69	76.08	53.62	53.08	58.13
Ours-B w/o OCT (OCTA)	85.76	69.23	84.37	62.57	60.65	68.54	<u>83.08</u>	<u>93.53</u>	<u>77.16</u>	76.04	80.00
Ours-L w/o OCT (OCTA)	303.25	67.69	83.62	61.33	59.92	65.42	81.54	91.13	73.92	<u>78.53</u>	75.00
Ours-H w/o OCT (OCTA)	630.69	<u>72.31</u>	<u>90.11</u>	61.41	59.75	65.21	80.00	93.02	69.49	71.35	68.33
Ours-B	85.76	84.62	97.03	81.25	81.25	81.25	84.62	97.03	81.25	81.25	81.25
Ours-L	303.25	84.62	97.03	81.25	81.25	81.25	84.62	97.03	81.25	81.25	81.25
Ours-H	630.69	84.62	97.03	81.25	81.25	81.25	84.62	97.03	81.25	81.25	81.25

3.3 Ablation Study

To quantify the contribution of cross-modal fusion to aging prediction, we compared the complete model with models combining different modalities. As shown

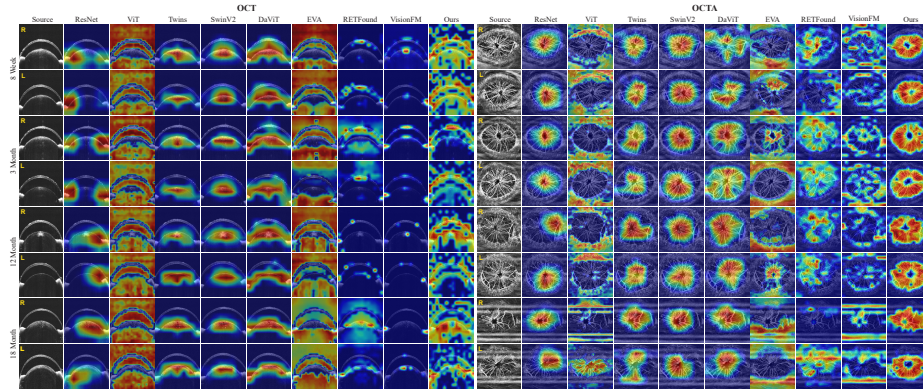


Fig. 3: Comparison of precise extraction of aging-related regions. Ours-B was used to generate heatmaps of Ours. Ours-L/H showed similar localization patterns.

Table 2: Ablation study of different modalities

Method	ACC $\uparrow$	AUC $\uparrow$	F1 $\uparrow$	PRE $\uparrow$	SEN $\uparrow$
Gait+Behavior	69.23	81.85	38.75	35.42	43.75
Ours	84.62	97.03	81.25	81.25	81.25
Ours w/ Gait + Behavior	92.31	97.73	88.89	95.00	87.50

in Table 2, the method using only the ocular modality significantly outperformed the traditional gait-behavior combination method, validating the high sensitivity and robustness of the ocular modality. When hierarchical fusion of behavioral and gait modalities was introduced, the model performance was further improved, indicating that cross-modal fusion effectively reduced false positives and enhanced the detection capability of early aging phenotypes. These results demonstrate a biological complementarity between peripheral structural degeneration in the ocular modality and central functional decline reflected in behavior and gait. The hierarchical fusion architecture, through a two-stage design of image modality alignment and cross-modal feature enhancement, successfully established a collaborative reasoning mechanism for structural-functional dual-dimensional aging characterization, providing a reliable technical path for achieving high-precision, interpretable, and personalized aging assessment.

4 Conclusion

This paper addresses the key challenge of non-invasive aging assessment by proposing a hierarchical cross-modal fusion framework that synergistically integrates three complementary modalities: ocular, behavioral, and gait. We introduce a comprehensive cross-modal dataset covering ocular, behavioral, and gait aspects, the

first cross-modal dataset designed for the joint assessment of aging, and propose modality-specific training paradigms for each modality. Our fusion framework systematically progresses from unimodal encoding and intra-domain OCTA-OCT alignment to inter-domain feature enhancement, employing an entropy-guided probabilistic weighting strategy to effectively bridge peripheral structural alterations with central functional output. Experimental results demonstrate that our method significantly outperforms unimodal baseline methods and state-of-the-art methods, enabling the precise extraction of fine-grained features from aging-sensitive regions, including the retinal vascular network and corneal epithelium. Our work establishes a clinically interpretable non-invasive aging assessment framework, providing a reliable technical pathway for high-precision, personalized gerontological assessment through multi-scale bioinformatics integration.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahadi, S., Wilson, K.A., Babenko, B., McLean, C.Y., Bryant, D., Pritchard, O., Kumar, A., Carrera, E.M., Lamy, R., Stewart, J.M., Varadarajan, A., Berndt, M., Kapahi, P., Bashir, A.: Longitudinal fundus imaging and its genome-wide association analysis provide evidence for a human retinal aging clock. *eLife* **12**, e82364 (Mar 2023). <https://doi.org/10.7554/eLife.82364>, <https://doi.org/10.7554/eLife.82364>
2. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 9355–9366. Curran Associates, Inc. (2021), <https://proceedings.neurips.cc/paper/2021/hash/4e0928de075538c593fbdabb0c5ef2c3-Abstract.html>
3. Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L.: DaViT: Dual attention vision transformers (Apr 2022). <https://doi.org/10.48550/arXiv.2204.03645>, <http://arxiv.org/abs/2204.03645>, arXiv:2204.03645 [cs]
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (Oct 2020), <https://openreview.net/forum?id=YicbFdNTTy>
5. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alpha-Pose: Whole-body regional multi-person pose estimation and tracking in real-time (Nov 2022). <https://doi.org/10.48550/arXiv.2211.03375>, <http://arxiv.org/abs/2211.03375>, arXiv:2211.03375 [cs]
6. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. pp. 19358–19369 (2023), https://openaccess.thecvf.com/content/CVPR2023/html/Fang_EVA_Exploring_the_Limits_of_Masked_Visual_Representation_Learning_at_CVPR_2023_paper.html
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. pp. 770–778 (2016), https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

8. Klusmann, V., Notthoff, N., Beyer, A.K., Blawert, A., Gabrian, M.: The assessment of views on ageing: A review of self-report measures and innovative extensions. *European Journal of Ageing* **17**(4), 403–433 (Feb 2020). <https://doi.org/10.1007/s10433-020-00556-9>, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7752934/>
9. Li, R., Chen, W., Li, M., Wang, R., Zhao, L., Lin, Y., Chen, X., Shang, Y., Tu, X., Lin, D., Wu, X., Lin, Z., Xu, A., Wang, X., Wang, D., Zhang, X., Dongye, M., Huang, Y., Chen, C., Zhu, Y., Liu, C., Hu, Y., Zhao, L., Ouyang, H., Li, M., Li, X., Lin, H.: LensAge index as a deep learning-based biological age for self-monitoring the risks of age-related diseases and mortality. *Nature Communications* **14**(1), 7126 (Nov 2023). <https://doi.org/10.1038/s41467-023-42934-8>, <https://www.nature.com/articles/s41467-023-42934-8>
10. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer V2: Scaling up capacity and resolution. pp. 12009–12019 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.html
11. Luxem, K., Mocellin, P., Fuhrmann, F., Kürsch, J., Miller, S.R., Palop, J.J., Remy, S., Bauer, P.: Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology* **5**(1), 1267 (Nov 2022). <https://doi.org/10.1038/s42003-022-04080-7>, <https://www.nature.com/articles/s42003-022-04080-7>
12. Miao, H., Liu, S., Wang, Z., Ke, Y., Cheng, L., Yu, W., Yu, D., Zhang, K., Gao, Y., Sun, Z.: Artificial intelligence-derived retinal age gap as a marker for reproductive aging in women. *npj Digital Medicine* **8**(1), 367 (Jun 2025). <https://doi.org/10.1038/s41746-025-01699-8>, <https://www.nature.com/articles/s41746-025-01699-8>
13. Nandi, A., Virmani, G., Barve, A., Marathe, S.: DBscorer: An open-source software for automated accurate analysis of rodent behavior in forced swim test and tail suspension test. *eNeuro* **8**(6) (Nov 2021). <https://doi.org/10.1523/ENEURO.0305-21.2021>, <https://www.eneuro.org/content/8/6/ENEURO.0305-21.2021>
14. Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., Zhao, Y., Shi, X., Xian, J., Qu, X., Zhu, S., Pan, L., Chen, X., Zhang, X., Jiang, S., Wang, K., Yang, C., Chen, M., Fan, S., Hu, J., Lv, A., Miao, H., Guo, L., Zhang, S., Pei, C., Fan, X., Lei, J., Wei, T., Duan, J., Liu, C., Xia, X., Xiong, S., Li, J., Lo, B., Tham, Y.C., Wong, T.Y., Wang, N., Yuan, W.: VisionFM: A multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**(12) (Nov 2024). <https://doi.org/10.1056/AIoA2300221>, <http://arxiv.org/abs/2310.04992>, arXiv:2310.04992 [eess]
15. Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., Kennedy, A.: The mouse action recognition system (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* **10**, e63720 (Nov 2021). <https://doi.org/10.7554/eLife.63720>, <https://doi.org/10.7554/eLife.63720>
16. Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. pp. 1653–1660 (2014), https://openaccess.thecvf.com/content_cvpr_2014/html/Toshev_DeepPose_Human_Pose_2014_CVPR_paper.html
17. Weinreb, C., Pearl, J.E., Lin, S., Osman, M.A.M., Zhang, L., Annapragada, S., Conlin, E., Hoffmann, R., Makowska, S., Gillis, W.F., Jay, M., Ye, S., Mathis, A., Mathis, M.W., Pereira, T., Linderman, S.W., Datta, S.R.: Keypoint-MoSeq: Parsing behavior by linking point tracking to pose dynamics. *Nature Methods* **21**(7),

- 1329–1339 (Jul 2024). <https://doi.org/10.1038/s41592-024-02318-2>, <https://www.nature.com/articles/s41592-024-02318-2>
18. Zampino, M., Polidori, M.C., Ferrucci, L., O'Neill, D., Pilotto, A., Gogol, M., Rubenstein, L.: Biomarkers of aging in real life: Three questions on aging and the comprehensive geriatric assessment. *GeroScience* **44**(6), 2611–2622 (Dec 2022). <https://doi.org/10.1007/s11357-022-00613-4>, <https://doi.org/10.1007/s11357-022-00613-4>
19. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **23**(10), 1499–1503 (Oct 2016). <https://doi.org/10.1109/LSP.2016.2603342>, <https://ieeexplore.ieee.org/abstract/document/7553523>
20. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., Keane, P.A.: A foundation model for generalizable disease detection from retinal images. *Nature* **622**(7981), 156–163 (Oct 2023). <https://doi.org/10.1038/s41586-023-06555-x>, <https://www.nature.com/articles/s41586-023-06555-x>