# Pet Age Automatic Recognition

**靳宇凡 23020241154347 计算机 2 班**

**张述锐 23020241154473 计算机 2 班**

**邱俊 23020241154431 计算机 2 班**

**王意然 36920241153253 AI**

**魏咏晨 36920241153255 AI**

## Abstract

With the rapid development of the pet industry in China, pet health insurance has become a key sector. As pets live longer, accurately identifying their age is crucial for insurance claims. This paper proposes a study on pet age prediction using image recognition technology. By building models based on ResNet, SSRNet, PVT_tiny, and PVTV2, we focus on extracting facial features of pets and treat the task as a regression problem. using Huber Loss to enhance model robustness. The study plans to conduct data cleaning, data augmentation, and model design, and evaluate the model's performance using Mean Absolute Error (MAE).

## Introduction

With the continuous rise in income levels in China, the number of pet owners has been steadily increasing, and the variety of pet-related products and services has become increasingly diverse. In recent years, the domestic pet industry has experienced rapid expansion, particularly between 2017 and 2022, with the market size exceeding one trillion yuan. As more people become pet owners and pet care concepts evolve, the future potential of the pet market remains immense, with sustained growth on the horizon.

Among the various sectors of the pet economy, pet health insurance has emerged as a key component, with the premium scale rising year by year. However, as pets' average lifespans increase, their risk of illness and insurance claims also rises, posing significant challenges for insurance companies. Accurately identifying the actual age of a pet during the insurance process is crucial, as older pets present higher health risks. Insurance companies must evaluate and mitigate these risks promptly to maintain healthy claim rates. Efficient and accurate identification of a pet's age not only helps insurers manage claim risks but also improves the user experience during the insurance process, reducing potential disputes over claims.

Given this background, we propose the research topic of "Pet Age Recognition," aiming to use advanced image recognition technology to accurately predict a pet's age. The goal of this research is to build an intelligent pet age recognition system based on photos of pets, utilizing physiological traits and visual changes over time. This system would provide insurance companies with a reliable and precise method for evaluating pet age during underwriting. This not only contributes to the digital transformation of the pet insurance industry but also supports pet health management by offering valuable insights.

## Related Work

Early works on age estimation are mainly focused on designing robust aging features and selecting learning algorithms. Some features are specifically designed for the age estimation problem, such as the facial features and wrinkles, the learned AGES (Aging pattern Subspace) features, as well as the biologically inspired features (BIF).

More recently, CNN-based methods have been widely adopted for age estimation due to its superior performance over existing methods. Zakariya Qawaqneh et al. used Deep CNN for age estimation based on the VGG-Face model, and they proved that a CNN model can be utilized for age estimation to improve performance[1]. Hlaing Htake Khaung Tin used Principal Component Analysis (PCA) to predict age of face images[2]. Haibin Liao et al. used CNN and Divide-and-Rule strategy for age estimation of face images. They used CNN to extract robust features from the images, then age-based and sequential study of rank-based age estimation learning methods is utilized and then a divide-and-rule face age estimator is propose[3]d. They proved that the performance of divide-and-rule estimators is much better than classical SVM and SVR. Olatunbosun Agbo-Ajala and Serestina Viriri proposed a novel CNN model to extract features from unconstrained real-life face images and classified them to age and gender groups. They achieved classification accuracy of 84.8% on age group and 89.7% on gender[4]. Zhang et al.[5] proposed that individual aging patterns are influenced by internal and external factors, and they utilized two deep learning models, CNN and LSTM, to learn these aging patterns. The CNN is used to extract features from the face, while the LSTM is employed to learn the aging patterns of individuals from the time series images of the face. Mei et al.[6] and Chen et al.[7] adopted CNN architectures with multiple different convolutional kernels and layers to extract facial features, combining the outputs of multiple CNNs to achieve final age recognition. Compared to a single CNN framework, the combination of multiple CNNs significantly improved the accuracy of age recognition.

While age estimation using neural networks is an active research area for humans, this problem for dogs has been so far overlooked. Yet dogs' ageing processes are in many aspects similar to human, which may lead to cross-fertilization between these areas.

# Proposed Solution

## ResNet50

ResNet50 (Residual Network with 50 layers) is a deep convolutional neural network model proposed by Kaiming He et al in 2015 to solve the problem of gradient disappearance in deep neural network training. The key feature of ResNet is the Residual Block, which allows deeper networks to be trained effectively by introducing skip connections that let the network learn the Residual directly rather than fitting the output directly. ResNet50 structure features: ResNet50 contains 50 layers, including convolutional layer, pooling layer and full connection layer. ResNet50 uses multiple residual blocks, each of which passes input directly to subsequent layers via a skip connection.

The main components used by ResNet50 include: 1. Convolution layer is used to extract features. 2. Batch Normalization is used to accelerate training and stabilize the network. 3.ReLU activation function is used to introduce nonlinearity. 4. The global average pooling layer is used to reduce the number of parameters. 5. The Softmax layer classifies. The structure of ResNet50 is shown in Figure 1.

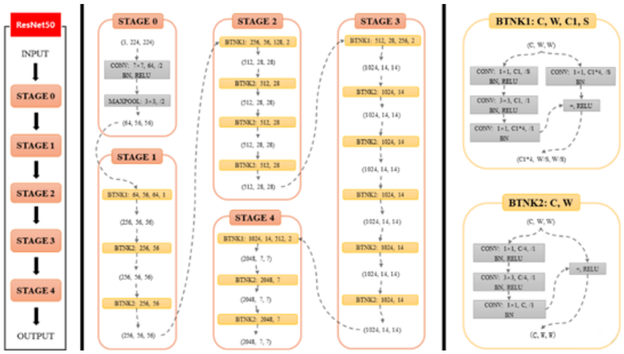


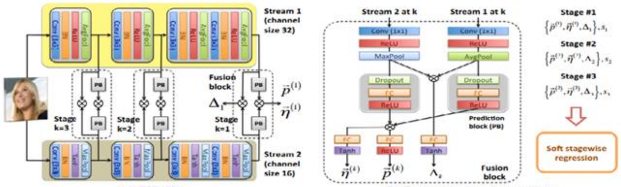
Figure 1.ResNet50 network structure

To use ResNet50 for pet age recognition, transfer learning and task adaptation of the model are required. We extend the data by rotating, cropping, flipping and other techniques to enhance the generalization ability of the model. The image size is resized to 256x256 pixels, and the pixel value is normalized. Also using the pre-trained ResNet50 model, the last fully connected classification layer of ResNet50 was removed and replaced with an output layer suitable for pet age classification. Cross-Entropy Loss is used as a loss function and a learning rate decay strategy is used to optimize the training process. Finally, a separate test set is used to evaluate performance.

## SSRNet

SSRNet (Soft Stagewise Regression Network) is a deep learning model designed for age estimation. It predicts age values in a phased manner, is suitable for small sample scenarios, and is lightweight, fast and efficient. The core idea of SSRNet is that SSRNet breaks down the age prediction task into several stages, each of which is responsible for predicting part of the age information, and finally combines the results to obtain a complete age prediction. The weighted average method is used to extract the final result from the predicted value to reduce the noise effect. Using a multi-branch structure, the model uses multiple sub-branches, each responsible for different feature extraction or age interval prediction.

The output of each branch is consolidated in a distributed manner. SSRNet is designed for small sample and limited computing resources, the model structure is simple, the number of parameters is small, easy to train and deploy. The SSRNet model breaks down the age estimation task into a combination of classification (rough age interval) and regression (fine age value), balancing the advantages of both methods. The structure of SSRNet is shown in Figure 2.



(a) SSR-Net

(b) Fusion block and regression

Figure 2.SSRNet network structure

In the data preparation stage, we cropped the image size to 64x64 pixels, and used lightweight CNN as the feature extraction module of SSRNet. Use the smoothL1 loss function (smoothL1Loss) as the loss function. In the verification set, the distribution of the predicted values was observed to be close to the true age. Model performance was evaluated using MAE. Finally, a separate test set is used to evaluate performance.

## PVT_tiny

ViT replaces the CNN backbone network with a pure Transformer model without convolutions and achieves good results in image classification tasks. Although ViT is suitable for image classification, it is difficult to directly use it for pixel-intensive prediction (such as object detection and segmentation) for two main reasons: the feature map output by ViT is single-scale and low-resolution. Even for common input image sizes, ViT has relatively high computational and memory costs. To solve the above problems, Wenhai Wang proposed a pure Transformer backbone network, Pyramid Vision Transformer(PVT), which can be used as a replacement for CNN in many downstream tasks, including image-level prediction and pixel-level dense prediction. As shown in Figure 3, PVT does this by using more fine-grained image blocks (4×4 pixels) as input to learn high-resolution features, which is critical for intensive prediction tasks. A progressively shrinking feature pyramid is introduced to reduce the sequence length of Transformer as the network deepens, significantly reducing the computing cost. The introduction of a spatial-reduction attention (SRA) layer further reduces resource consumption when learning high-resolution features.

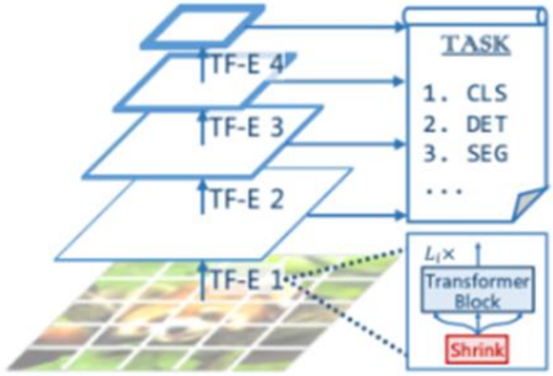


Figure 3.Pyramid Vision Transformer

The PVT-TINY (Pyramid Vision Transformer Tiny) is a lightweight version of the Pyramid Vision Transformer (PVT) family, designed for visual tasks. The PVT-Tiny is a powerful lightweight Transformer model with pyramidal multi-scale features and an efficient self-attention mechanism. When applied to pet age recognition, its global modeling capabilities can be used to capture age-related subtle features while taking into account efficiency and performance, making it ideal for real-world scenarios with limited resources. The data preparation phase is the same as ResNet50. We use PVT-Tiny weights that are pre-trained on large-scale data sets. Replace the last few layers of the model to accommodate pet age identification. Freeze the first few layers of PVT-Tiny and train only the upper and

## PVTv2

Pyramid Vision Transformer v2 (PVTv2) is an upgraded version of Pyramid Vision Transformer (PVT), which improves computing efficiency and performance through a series of improvements. PVTv2 inherits the pyramid structure and multi-scale feature extraction capabilities of PVT, while optimizing the attention computation method and network structure to make it more efficient and accurate in visual tasks. Compared to PVT, the main improvement of PVTv2 is the improved Attention mechanism, PVTv2 introduces Linear Attention in the attention calculation of each layer, reducing the complexity from square to linear, significantly reducing the amount of computation. The use of packet convolution reduces computational costs, while retaining feature extraction capabilities and obtaining more efficient Token embedding modules. The convolution operation is added at each stage to improve the capture ability of local context and obtain a larger receptive field. Reduce model parameters and FLOPs while improving feature representation, suitable for resource-constrained devices. PVTv2 performs well in tasks such as classification, object detection, and instance segmentation.

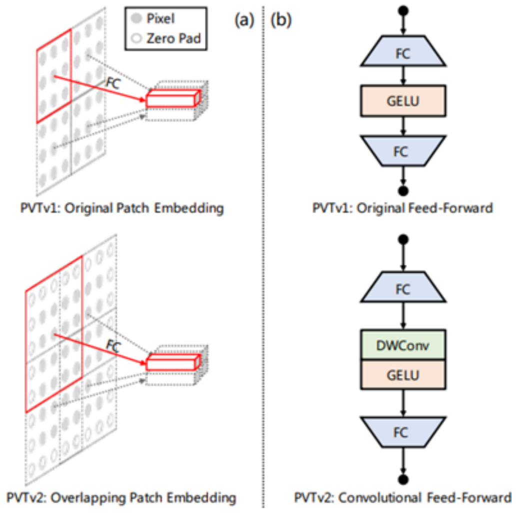

Pixel
Zero Pad
(a)
(b)
FC
PVTv1: Original Patch Embedding
FC
GELU
FC
PVTv1: Original Feed-Forward
FC
PVTv2: Overlapping Patch Embedding
FC
DWConv
GELU
FC
PVTv2: Convolutional Feed-Forward

Figure 4: Two improvements in PVT v2. (1) Overlapping Patch Embedding.(2)Convolutional Feed Forward Network.

PVTv2 can be well applied to pet age recognition tasks, especially for feature extraction and analysis of diverse pet data (different breeds, posture, light, etc.). The data preparation phase is the same as ResNet50. Make a pre-trained

PVTv2 model. The loss function is trained using both MSE and smooth L1 loss functions. Finally, a separate test set is used to evaluate performance.

## Eexperiment

### Dataset

We selected the dataset provided by the Kaggle competition for pet age estimation: https://www.kaggle.com/datasets/marquis03/afac2023-pet-age-identification.The dataset primarily consists of pet dogs and is divided into training, validation, and test sets. Both the training and validation sets are noisy datasets, with approximately 6% of the data containing incorrect labels. The labels are provided in months, ranging from [0, 192]. The test set does not include any noise. The structure of the dataset and sample examples are shown in Table 1.

Table 1. Dataset Composition

| DATASET | SIZE | LABEL |
|---|---|---|
| trainset | 20000 | train.txt |
| valset | 3000 | val.txt |
| testset | 3000 | None |

### Model Training

First, we preprocess the images in the dataset. For SSRNet, the input images are resized to 64x64 pixels, while for the other three models, the images are resized to 256x256 pixels and randomly horizontally flipped with a probability of 50%. This operation enhances the diversity of the dataset, especially when the target position is not affected by flipping. Finally, each channel of the images is normalized to accelerate the convergence of the neural network, enabling faster and more stable training.

For training SSRNet and ResNet models, the models are trained for 100 epochs. During each epoch, the dataset is divided into multiple batches for training and validation, with a batch size of 64. The optimizer parameters are set as follows: learning rate = 0.002, weight decay = 1e-4, step size for learning rate adjustment = 10, and decay factor gamma = 0.5.

ResACM uses a pretrained ResNet50 model based on the ImageNet dataset, while SSRNet is trained from scratch on the dataset provided in the referenced paper.

For training the PVT_tiny and PVTv2 models, the models are trained for 60 epochs with a batch size of 128. The learning rate is set to 0.001, and a learning rate adjustment function is implemented, which reduces the learning rate by a factor of 10 if the training loss does not decrease within 5 epochs. Both models use pretrained parameters based on the ImageNet dataset.

For the ResNet and SSRNet models, different loss functions are used for the classification and regression structures. The classification model ResACM adopts the cross-entropy loss function (CrossEntropyLoss), while the regression model achieves optimal results using the smooth L1 loss function (smoothL1Loss) after small-scale testing.

For the PVT and PVTv2 models, both MSE and smooth L1 loss functions are used for training, with the latter performing slightly better than the former.

### Performance metrics

We ultimately use the best Mean Absolute Error (MAE) on the validation set as the evaluation metric. MAE represents the difference between the predicted values and the actual values, and it is calculated using the following formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i} - y_i|$$

where $\widehat{y_i}$ represents the predicted value of the iii-th sample,$y_i$represents the actual value of the iii-th sample, and NNN is the total number of samples.

The MAE data for all the attempted models is shown in the table 2.

Table 2. MAE Data

| MODEL | MAE |
|---|---|
| ResNet | 29.1557 |
| SSRNet | 25.9381 |
| PVT_tiny | 26.5188 |
| PVTV2 | 26.8344 |

Additionally, for the best-performing model, we calculated the accuracy of the absolute differences between the predicted and actual values on the validation set within 3 months and 5 months.

The performance metrics of the optimal SSRNet model are shown in the figure 5~7.

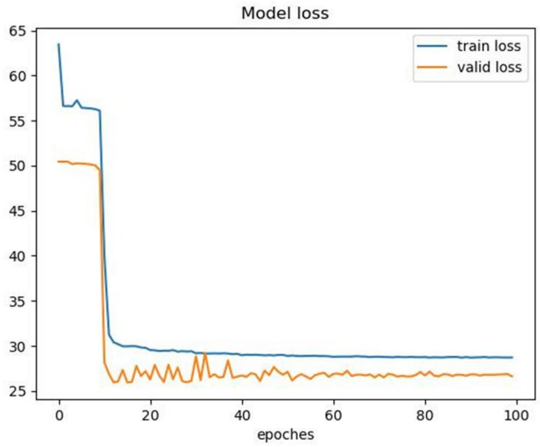


Figure 5. The loss function changes with iterations.

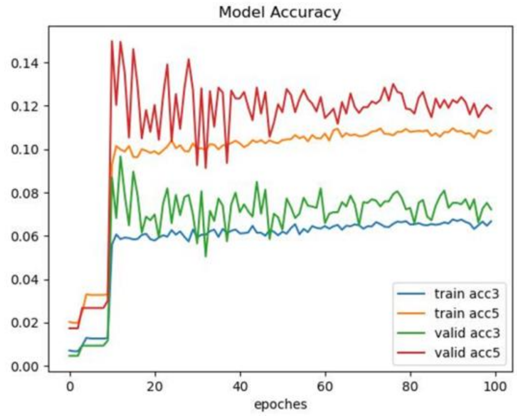


Figure 6. The Model Accuracychanges with iterations.

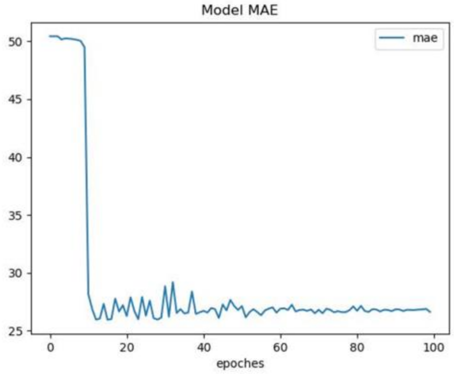


Figure 7. The Model MAE with iterations.

## Prediction

Based on the above metrics, the best-performing model is the SSRNet model. We then input the test set data into the trained model and obtained the following prediction results show in figure 8.

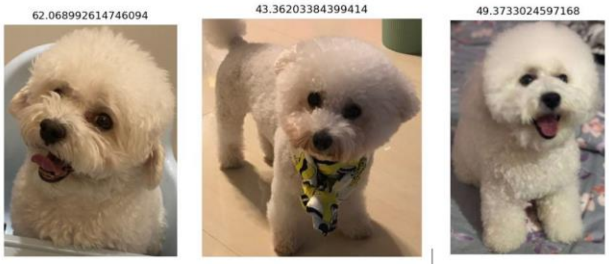


Figure 8. Prediction result

## Conclusion

This study used four different models—ResNet, SSRNet, PVT_tiny, and PVTV2—to predict the age of dogs based on their facial features. We trained and evaluated the models on a dog image dataset, focusing on extracting key facial features related to age. The experimental results show that SSRNet achieved the best performance.

# References

[1] Qawaqneh, Zakariya, Arafat Abu Mallouh, and Buket D. Barkana. "Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model." arXiv preprint arXiv:1709.01664 (2017).

[2] N. Chauhan, "Predict Age and Gender using Convolutional Neural Network and openCV", Medium, 2020. [Online].convolutional-neural-network-and-opencv-fd90390e3ce6. [Accessed: 05- Jun- 2020].

[3] Liao, Haibin, et al. "Age estimation of face images based on CNN and divide-and-rule strategy." Mathematical Problems in Engineering, 2018.

[4] O. Agbo-Ajala and S. Viriri, "Face-Based Age and Gender Classification Using Deep Learning Model", Image and Video Technology, pp. 125-137, 2020.

[5] ZHANG H Y，ZHANGY，GENG X.Recurrent age estimation[J].Pattern Recognition Letters(PRL),2019(125):271-277.

[6] Mei S，Geng Y，Hou J，et al. Learning hyperspectral images from RGB images via a coarse-to-fine CNN[J]. Sciece China. Information Sciences,2022,65(5):152-162.

[7] Chen S，Zhang C，Dong M. Deep Age Estimation：From Classification to Ranking[J]. IEEE Transactions on Multimedia, 2022(99):1-12.