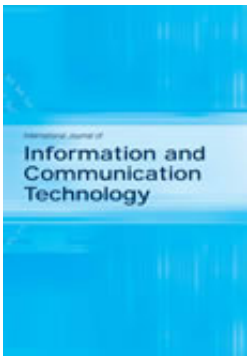
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Deep learning for visual aesthetics: using convolutional vision transformers and HRNet for classifying anime and human selfies

Congli Zhang

DOI: 10.1504/IJICT.2025.10071591

Article History:

Received:	24 March 2025
Last revised:	14 April 2025
Accepted:	14 April 2025
Published online:	18 June 2025

Deep learning for visual aesthetics: using convolutional vision transformers and HRNet for classifying anime and human selfies

Congli Zhang

Zhengzhou Academy of Fine Arts,
Speciality of Fine Arts,
Zhengzhou, Henan, 451452, China
Email: zhangcongli0718@163.com

Abstract: Digital media today plays a vital role in visual aesthetics and bringing them into play can impact user engagement and be crucial for personalised recommendations of content. Using AI, the task to classify and differentiate between human selfies and animated images, which are hard because of the subtle stylistic changes and the complex feature presentations in both categories. In this research study, we proposed an advanced framework that utilises vision transformers (ViT) and high-resolution networks (HRNet) for classification. With the help of an online dataset, the proposed models not only learn high level representations but also representational contextual dependencies well, classifying test data with 99% accuracy for ViT and 97% for HRNet at a level better than 10% of what traditional convolutional neural network (CNN) based models can achieve. The results leading for automatically content moderation, provide a solid base of using advanced vision models into multimedia and digital content processing.

Keywords: vision transformers; ViT; artificial intelligence; deep learning; visual aesthetics; convolutional neural networks; CNNs; feature extraction; classification.

Reference to this paper should be made as follows: Zhang, C. (2025) 'Deep learning for visual aesthetics: using convolutional vision transformers and HRNet for classifying anime and human selfies', *Int. J. Information and Communication Technology*, Vol. 26, No. 20, pp.75–98.

Biographical notes: Congli Zhang is a researcher at the Zhengzhou Academy of Fine Arts, specialising in the intersection of artificial intelligence and visual aesthetics. Her research focuses on applying machine learning and deep learning techniques, particularly in the domains of computer vision and digital image processing.

1 Introduction

Computer vision has experienced a rise in visual content creation and sharing on digital platforms coupled with challenges and opportunities. There are so many visual styles present on these platforms and aesthetic anime and human selfies are the most popular and unique (Hou and Pan, 2023). Anime usually features exaggerated features, bright colours, and non-photorealistic rendering, are often highly stylised pictures (Sardenberg

et al., 2024). However, humans' selfies tend to display real world likeness with differing expressions, different lighting, and different location (Khan et al., 2024). The task of separating these categories into interesting and useful applications in content moderation (Wang, 2021), personality detection (Naz et al., 2024), personalised media collection (Bansal et al., 2024) and recommender systems (Talha et al., 2023). Over recent years with the advances in artificial intelligence (AI), there has been massive progress in image classification tasks. Convolutional neural network (CNN) and transformers became able to effectively identify and distinguish even extremely complicated visual patterns (Wang, 2025). Despite these advances, it is still difficult to tell the difference between aesthetic anime and human selfies because the textures and visual details shared between the two (like shared facial structure or colour palette) are so subtle. This complexity makes the creation of more sophisticated models needed to extract and understand the discriminative features in these visually distinguishable but on occasion converging two categories (Huang and Zheng, 2023). The problem is further compounded by the fact that anime and human selfies are dynamic and changing. New artistic trends in the anime style diversify and are still diversifying while human selfies show great variety in the issues of the cultural, demographic and environmental spectrum (Li and Zhang, 2024). Thus, any robust classification model must also not only excel at static image recognition but be capable of adapting to these evolving patterns. Therefore, global features need to be learned but fine-grained local details are needed to capture the essence of both categories, which requires the incorporation of models capable of learning these features (Lv et al., 2021). Achieving high classification accuracy in real world applications requires addressing these requirements.

In this study a novel framework is proposed for the classification of aesthetic anime and human selfie images based on vision transformers (ViT) and HRNet to solve the problem of distinguishing between human selfies and aesthetic anime images. Combining the self-attention mechanism to capture global dependencies in image, ViTs are well suited to understand complex patterns in anime. However, HRNet aims at high resolution feature maps at every stage of the network, which is necessary for the spatial resolution while humans are distinguishing fine grained difference in their selfies. Through this, intend to take the field of interactive media image classification to higher levels by combining these state-of-the-art architectures. The main contributions of this research are as follows:

- A novel framework is proposed leveraging state-of-the-art transformer models (ViT and HRNet) for classifying aesthetic anime and human selfies.
- Achieved the highest accuracy of 99% using ViT, significantly outperforming the baseline CNN model.
- Demonstrated the effectiveness of HRNet in preserving high-resolution features, enhancing classification performance in complex datasets.
- Conducted comprehensive comparisons and ablation studies, showcasing the superiority of transformer-based architectures over traditional CNN for the task.

The remainder of this paper is organised as follows: Section 2 reviews related work in image classification including human selfie recognition and anime recognition. Then in section 3 describe proposed methods to carry out this composable modelling task, discussing data preparation and model architectures. The experimental results are

presented in Section 4 and discussed. Section 5 concludes the paper with a discussion of key contributions and future direction research.

2 Related work

Recent years have seen rapid growth in approaching the problem of classifying images with deep learning, especially between aesthetic anime images and selfies, as display in Table 1. Li et al. (2021) proposed AniGAN, an unsupervised anime face generation with style guidance style adversarial network. Simultaneously transferring colour and texture styles and local transformations in facial shape to reference anime style, their model approximates portrait photos as anime style images applying GAN based translators in capturing complicated stylistic aspects. Rios et al. (2021) introduced DAF framework based on a large-scale crowd sourced dataset for anime character recognition. They measured the generalisation capabilities of such architectures such as ViT and ResNet's on domain specific datasets to deal with such diverse and imbalanced data distributions really matters. Chen et al. (2023) developed PAniC-3D from portraits of anime characters. Instead, they take on the special problems of anime style domains like hair geometry and non-photorealistic shading. PAniC-3D leaps this gap between 2D and 3D character modelling by reconstructing 3D characters' heads from 2D illustrations through-looking leverage on a line-filling model and volumetric radiance fields.

In their contributions, Li et al. (2022) provided a challenging benchmark for anime style recognition with the collection of a dataset to evaluate model performance, to learn abstract painting styles, instead of discriminative features of individual roles. However, the initial experiments with the techniques of person re-identification for AGW and TransReID demonstrate their bad performance of the semantic gap between different anime styles and highlight the necessity of dedicated approaches in the task of model. Naftali et al. (2022) tried various models such as InceptionV3, InceptionResNetV2, MobileNetV2 and EfficientNet for classifying anime character face. Taking transfer learning as a base, they found that EfficientNetB7 gives the highest top one accuracy and MobileNetV2 at a much faster inference time with a slightly lower accuracy. In Jiang et al. (2023), they proposed Scenimefy, a semi-supervised image to image translation framework to generate high fidelity anime scenes from complex real-world images. To overcome the challenges in scenes complexity and feature of anime style, their approach assumes that learning happens by guiding with structure consistent pseudo paired data. For age and gender classification from in-the-wild facial images, Singh and Singh (2024) derived a hybrid transformer-sequencer model. Finally, based on self-attention mechanism combined with BiLSTM network, they propose a method that exploits both possibility of discovering global and local features, which outperforms existing models. Finding that such architectures could prove to be of use in solving complex tasks such as facial recognition in unconstrained environments, this study serves as a contribution to design systems that can leverage hybrid architecture to deal with complex tasks in this setting. Bekhet et al. (2022) introduce a CNN method for gender recognition from unconstrained selfie images. To address the challenges presented by non-standard selfies from real life environments they achieve an accuracy of 89% on their own dataset. In this work, the performance of deep learning approaches in terms of handling variability, unpredictability in selfie images is stressed. Wang et al. (2023) presented the dual-branch fusion approach for anime character recognition with teaching neural networks to imitate

human habits. The approach studied here utilises the fundamental structure of cartoon and examines other colour and pixel details in a gradual process to get better recognition performance in recognising cartoon character image from the artist's view.

Table 1 Analysis of existing studies

Ref.	*Model*	*Dataset*	*Features*	*# Classes*	*Result (%)*
Li et al. (2021)	GAN	Anime images	Style-guided generation	2	77
Rios et al. (2021)	EfficientNet	Stylised anime	Face recognition	4	94
Chen et al. (2023)	3D hybrid CNN	Anime portraits	Face recognition	6	88
Li et al. (2022)	CNN-based	Custom anime	Anime style recognition	7	82
Naftali et al. (2022)	CNN	Character faces	classification	4	91
Jiang et al. (2023)	VAE	Anime image	Anime scene generation	2	78
Singh and Singh (2024)	GCN-based	Selfie-anime	Multi-modal classification	4	85
Bekhet et al. (2022)	Hybrid transformer-sequencer	In-wild facial images	Age and gender classification	2	96
Wang et al. (2023)	CNN-based	Selfie-anime	Gender recognition	2	94
Lan et al. (2022)	CNN + ResNet	Anime character	Human-like recognition	3	92
Zheng et al. (2020)	CNN	Cartoon face	Face recognition	2	90
Chen and Zwicker (2022)	Transfer learning	Anime character	Pose estimation	2	87
Hasan and Mustafa (2022)	CNN + Genetic algorithm	Selfie images	Gender recognition	2	97

A GCN-based multi modal introduced by Lan et al. (2022) multi label attribute classification for anime illustrations has been given. To better classify the attributes, their approach makes use of domain specific semantic feature to capture relationships between different attributes. To advance research in the domain of cartoon face recognition, Zheng et al. (2020) produced a benchmark dataset. Based on this dataset, they evaluated several types of deep learning models and discussed the challenges and the potential solutions for successful cartoon face recognition. There have been transfer learning techniques explored by Chen and Zwicker (2022) for pose estimation of illustrated characters. However, their study shows that by using pre trained models and fine tune in the domain specific datasets, pose estimation of illustrated characters can achieve higher accuracy. In study of Mustafa (Hasan and Mustafa, 2022) proposed a gender recognition method from selfie images through an integration of the CNN and genetic algorithms (GAs). The hybrid approach presented here seeks to exploit the feature selection abilities and

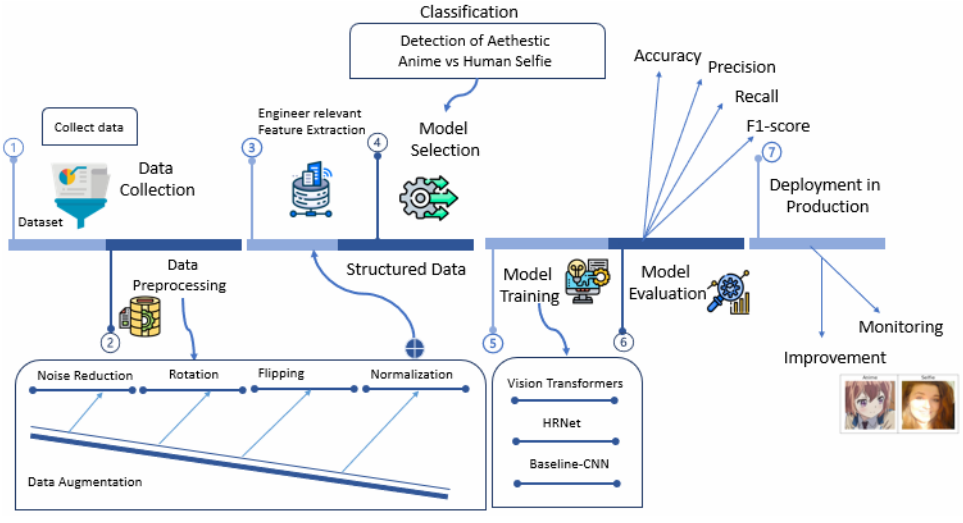
improve the classification accuracy for the variability and unpredictability facial capturing in selfie images.

Collectively these studies contribute to the understanding of the use of deep learning to distinguish between aesthetic anime images from human selfies. Further, they point out how careful selection of model architecture, dataset composition, and training methodology is crucial for accurately modelling the 'special' properties of application categories.

3 Proposed research methodology

This study employs a deep learning-based approach to discriminate between human facial selfies and anime animated images using ViT and high-resolution networks (HRNet). Following a structured pipeline, as shown in Figure 1, data collection, pre-processing are passed into the proposed deep learning architectures using optimised hyperparameter tuning strategy based upon learning rate scheduling and dropout regularisation is used to train the models and prevent overfitting, and then evaluate mode using classification metrics like accuracy, precision, recall, F1-score, and confusion matrices. The experimental setup is implemented using state-of-the-art deep learning frameworks for scalability and computational efficiency.

Figure 1 Framework of applied methodology (see online version for colours)



3.1 Data pre-processing

Data pre-processing is a necessary step in a deep learning pipeline since it is a step to clean up data to be used in training and inference. To make the model robust, data augmentation techniques implement – rotation, flipping, resizing, brightness adjustment and noise removal are applied, as shown in Figures 2 and 3, to classify anime images and human selfies. The details of these techniques are given in Table 2.

Table 2 Applied pre-processing techniques

Technique	*Description*	*Equations*
Rotation	Rotates the image by a specified angle to increase data variety. For each input image *I*, rotation is performed by applying a transformation matrix.	$T_{rotation} = \begin{matrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{matrix}$ where θ is a random angle.
Flipping	Flips the image horizontally or vertically to enhance invariance.	$T_{flip} = \begin{matrix} -1 & 0 & W \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$ where W is the image width
Resizing	Adjust the dimensions of the image to a fixed size while maintaining aspect ratio. For an input image *I* of size $H \times W$ resizing to is performed by interpolation methods.	$T_{resized(x',y')} = \sum_{x=0}^{H}\sum_{y=0}^{W} I(x, y), k(x', x), k(y', y)$ where k is the interpolation kernel
Brightness adjustment	Alters the brightness by scaling pixel intensities.	$T_{brightness} = I.\alpha$ where α is a randomly sample scaling factor
Noise removal	Reduces unwanted noise in the image using filters.	$T_{denoised(x,y)} = \frac{1}{2\pi\sigma^2}\sum_{i=S}^{l}\sum_{j=S}^{W}(I(x-i), y-j))^{2\frac{i^2+j^2}{2\pi\sigma^2}}$ where σ is the standard deviation of Gaussian kernel, s is the size of kernel

3.2 Feature extraction and classification using ViT model

Recently, ViTs were proposed as a powerful alternative to CNNs in the context of image classification tasks, biding on the transformer architecture, developed for image processing. Consequently, ViTs model images as sequences of patches to allow the network to leverage long range dependencies and perspective context.

3.2.1 Patch embedding

This layer processes the images as grids of pixels, ViTs divide an image $I \epsilon \mathbb{R}^{H*W*C}$ into non-overlapping patches size $P * P$. Each patch is flattened into a vector, forming a sequence of patch embeddings, as in equation (1).

$$x_p = Flatten(I[i : i + P, j : j + P])\forall(i, j)\epsilon\{(0, 0), ..., (H - P, W - P)\} \quad (1)$$

where $x_p \epsilon \mathbb{R}^{P^2C}$. These embeddings are then linearly projected into a lower dimensional space, defined in equation (2.)

$$z_o = \left[x_p^1 W_E ; .....; x_p^N W_E \right] + P \quad (2)$$

where $W_E \epsilon \mathbb{R}^{P^2C*D}$ is a trainable matrix, $P \epsilon \mathbb{R}^{N*D}$ is the positional embeddings, and $N = \frac{H.W}{P^2}$ is the number of patches.

Figure 2 Samples of anime after data augmentation (see online version for colours)

3.2.2 Self-attention mechanism

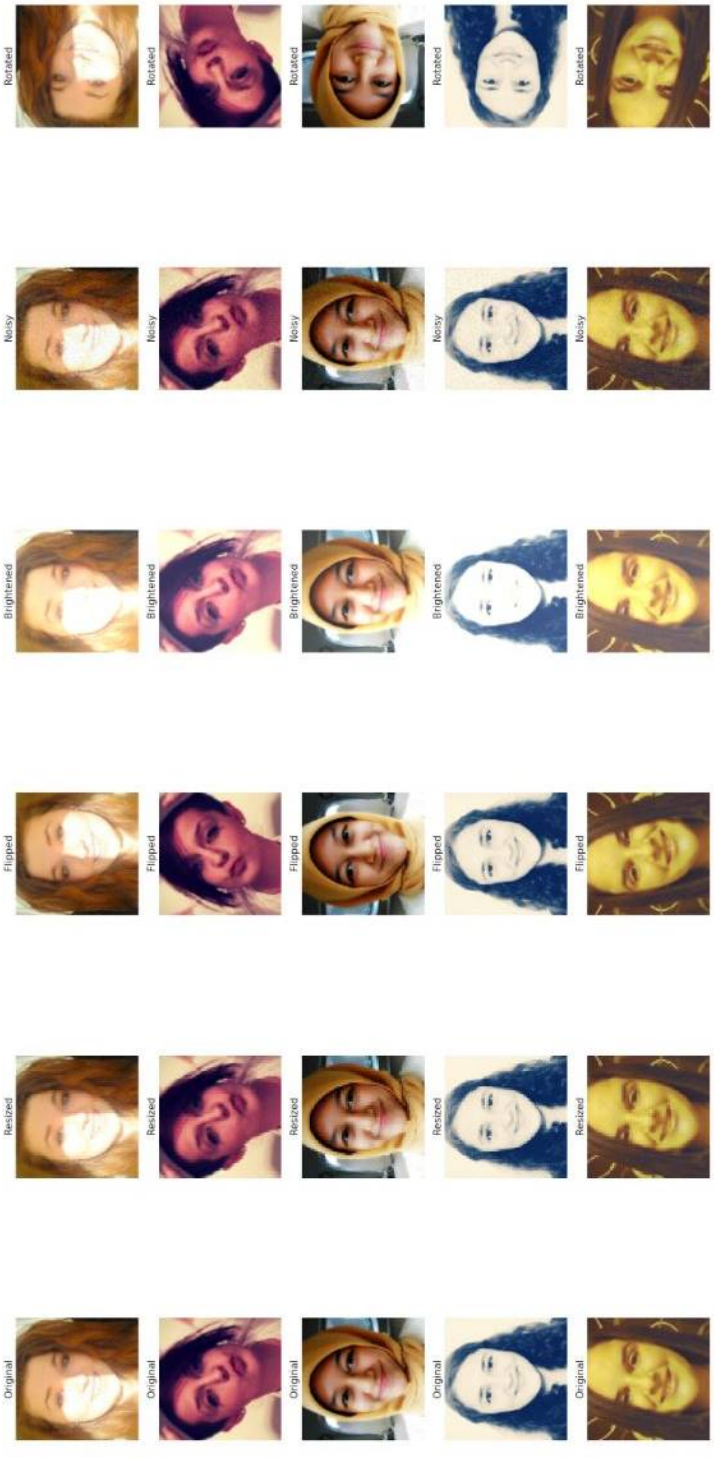
Multi-head self attention (MHSA) mechanism is the core of the ViTs that makes the model able to capture the relationships between all patches. For each patch embedding, three vectors are computed as in equation (3).

$$Q_i = z_i W_Q, K_i = z_i W_K, V_i = z_i W_V \tag{3}$$

where W_Q, W_K, $W_V \epsilon \mathbb{R}^{D*d}$ are trainable weight matrices. The attention score for each pair of patches is computed as in equation (4):

$$Attention(z_i, z_j) = Softmax\left(\frac{Q_i K_j^T}{\sqrt{d}}\right).V_j \tag{4}$$

Figure 3 Samples of human selfie after data augmentation (see online version for colours)



The scaled dot-product attention normalises the similarity score using $\sqrt{d}$ to prevent gradients from becoming excessively large. Multi-head attention aggregated information across *h* attention heads, defined in equation (5):

$$MHSA(z) = concat\left(head_1, ..., head_h\right)W_o \tag{5}$$

where $head_k = Attention(z, z)$ and $W_o \in \mathbb{R}^{hd*D}$.

3.2.3 Feed-forward network (FFN)

Each transformer layer includes a position-wise-FFN applied independently to each patch embeddings.

$$FFN(z) = GELU\left(zW_1 + b_1\right)W_2 + b_2 \tag{6}$$

where $W_1 \in \mathbb{R}^{D*D_f}$, $W_2 \in \mathbb{R}^{D_f*D}$ and $D_f > D$.

3.2.4 Layer normalisation (LN) and residual connections

LN and residual connections stabilise training and improve convergence, in equation (7):

$$z' = LN(z + MHSA(z)),\ z'' = LN\left(z' + FFN\left(z'\right)\right) \tag{7}$$

3.2.5 Classification token and prediction

A learnable classification token $z_{class} \in \mathbb{R}^D$ is prepended to the sequences of patch embeddings. After passing through the layer transformer, the token is used for classification, defined in equation (8). Figure 4 shows the proposed model architecture for this study.

$$y = Softmax\left(Z_{class}W_c\right) \tag{8}$$

where $W_c \in \mathbb{R}^{D*C}$ maps the final embedding to class logits.

3.3 High-resolution network

On the spatial task of semantic segmentation, pose estimation, image classification or other spatial tasks, HRNet is a state-of-the-art architecture. Compared with the traditional architectures shown in Figure 5, HRNet keeps high resolution features at multiple resolutions and fuses the features. By capturing global context and fine-grained details simultaneously, it can enable classification of images.

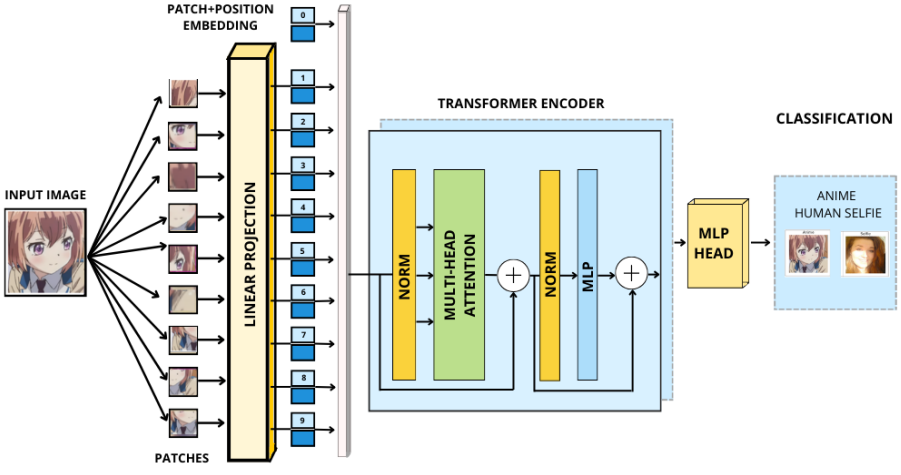
3.3.1 Multi-resolution feature representations

HRNet maintains parallel streams of features maps at different resolutions. For a given input image *I* of dimensions *Height H*, *Width W* and *Channel C*, the network begins with a high-resolution representation X_1^1. Subsequent stages create parallel representation at lower resolutions $X_2^1, X_3^1, ...$ defined in equation (9).

$$X_s^1 = f_s^1(I), s = 1, 2, 3, ... \tag{9}$$

where f_s^1 denotes the convolutional operations for the s^{th} resolution at the first stage, The resolution decreases progressively, typically by a factor of 2.

Figure 4 Proposed model architecture (see online version for colours)

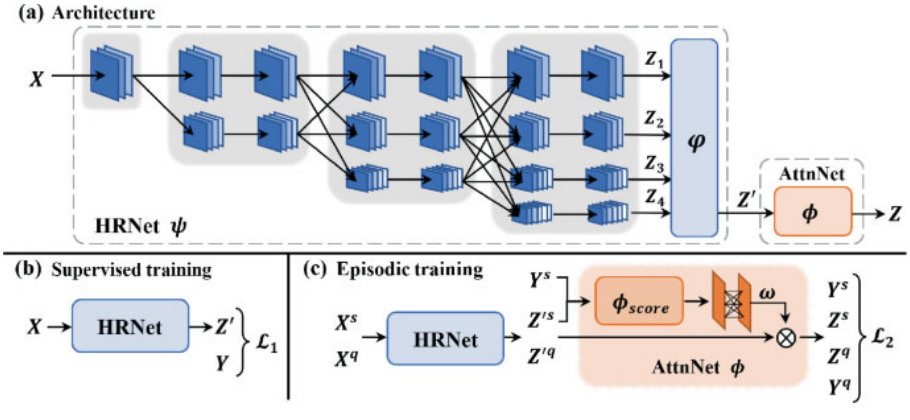


To ensure that high-resolution details are enriched with low-resolution semantic information. HRNet fuses features across resolutions at each stage. Let X_s^t denotes the feature map of resolutions s at stage t. The fused feature map is computed as in equation (10):

$$X_s^t = X_s^{t-1} + \sum_{r \neq s} \mathcal{U}\left(X_r^{t-1}\right) \tag{10}$$

where $\mathcal{U}$ denotes the down sampling or up sampling operations to align resolutions. X_r^{t-1} represents the feature map from a different resolution r at the previous stage.

Figure 5 Working of HRNet model (see online version for colours)



Source: Li and Mu (2023)

3.3.2 Stage wise processing

HRNet refines the feature maps through multiple stages, where each stage consists of repeated bottleneck layers, in equation (11).

$$X_s^t = ReLu\left(X_s^{t-1} + W_s^t * X_s^{t-1}\right) \tag{11}$$

where W_s^t represents the trainable weights of the convolution operation at stage *t*.

3.3.3 Output head

At the final stage, HRNet aggregates multi-resolution features for prediction. The final output Y is computed by concatenating all resolutions and passing them through a classification or regression head, as in equation (12):

$$Y = Softmacx\left(Concat\left(X_1^T, \mathcal{U}\left(X_2^T\right)\right).W_{out}\right) \tag{12}$$

where T is the final stage, W_{out} are the layer weights, and $\mathcal{U}$ ensures all feature maps are aligned to the highest resolution. Table 3 defines both model compatibility and generalisation of model's power.

Table 3 Comparison of vision transformer (ViT) vs. HRNet – advantages

Feature	*Vision transformer (ViT)*	*HRNet*
Global feature extraction	Captures long-range dependencies using self-attention, making it highly effective for complex patterns.	Maintains high-resolution representations throughout, preserving fine-grained spatial details.
Scalability	Scales efficiently with large datasets and benefits from pre-training on massive image corpora.	More efficient for real-time applications with lower computational overhead compared to ViT.
Robustness to occlusions and variations	Processes full image context, making it resilient to occlusions and background variations.	Strong in maintaining spatial consistency, useful for tasks requiring precise localisation.

3.4 Dataset

This study uses a dataset retrieved from a repository and grouped according to anime (animated) and human selfies. The dataset has a tremendous number of fractional variations, poses, lightnings conditions, artistic styles to achieve robustness in classification, as sample shown in Figure 6. To improve model generalisation, the dataset is divided into a standard 70–30 split, where 70% of images are assigned for training and 30% for a test, as distribution display in Table 4.

Table 4 Distribution of images

	Actual images	*Training images*	*Testing images*
Anime	3,500	2,450	1,050
Human selfie	3,500	2,450	1,050
Total	7,000	4,900	2,100

Figure 6 Samples of dataset images (see online version for colours)

3.5 Performance evaluation measures

Classification performance evaluation is a key part of any classification task for evaluating effectiveness of the model. The first set of metrics are accuracy, precision, recall, F1-score, confusion matrix analysis, are defined in Table 5. Together, these metrics demonstrate an understanding of model performance at once based on true positives (TP), false positives (FP), true negative (TN) and false negatives (FN).

Table 5 Model performance measures

Measure	*Description*	*Equation*
Accuracy	Overall proportion of correctly classified samples.	$(TP + TN) / (TP + FP + TN + FN)$
Precision	Proportion of correctly predicted positive samples.	$TP / (TP + FP)$
Recall	Ability to correctly identify all positive samples.	$TP / (TP + FN)$
F1-score	Harmonic mean of precision and recall.	$2.(Precision.Recall) / (Precision + Recall)$

4 Results and discussion

4.1 Results with baseline model

In this section, analysis of baseline CNN model presented that achieves high predictive capability of classifying anime and selfie images to overall accuracy of 93%. The model successfully extracts spatial hierarchies of features based on the convolutional layers, which can simultaneously encode low level textures and high-level semantic patterns. Though the simplicity of CNN makes it less complicated than many advanced deep learning architectures, its feature extraction and classification ability is also robust, which ensures its case of use in complicated image recognition tasks. This potential to generalise well across a wide range of datasets suggests it can be deployed in large scale in real world aesthetic classification applications. It correctly identifies anime images with 93% precision and 94% recall – i.e., it can correctly detect art with low false positives. The selfie classification has a 93% precision and 92% recall, just like the model can tell when a human looks like a human and when a human does not. All these F1

scores of 0.93 indicate CNN's stable performance; indeed, the CNN can learn and generalise well on both categories without overfitting. The Confusion matrix shown in Figure 7 demonstrates the strength of CNN in classifying anime and selfie images. Specifically, 46 out of 708 anime images were misclassified as selfies, as were 33 of 652 selfies. The low misclassification rates here thus confirm CNN's success in discriminating against the subtle textural and structural differences between artistic illustrations and real photographs. The accuracy and loss graphs shown in Figure 8 give more insights into model's learning over multiple training epochs. It shows that the model learns good discriminative features and thereby avoids overfitting and underfitting, while its accuracy curve progressively increases and stabilises at around 93%. Overall, it has been shown that the baseline CNN model has high classification ability within the high accuracy of 93% and balanced metrics in both anime and selfie classes. These results demonstrate that CNN is both a robust and dependable approach to visual recognition, especially in tasks with applications distinguishing between artistic and real imagery.

Figure 7 Confusion matrix of CNN model (see online version for colours)

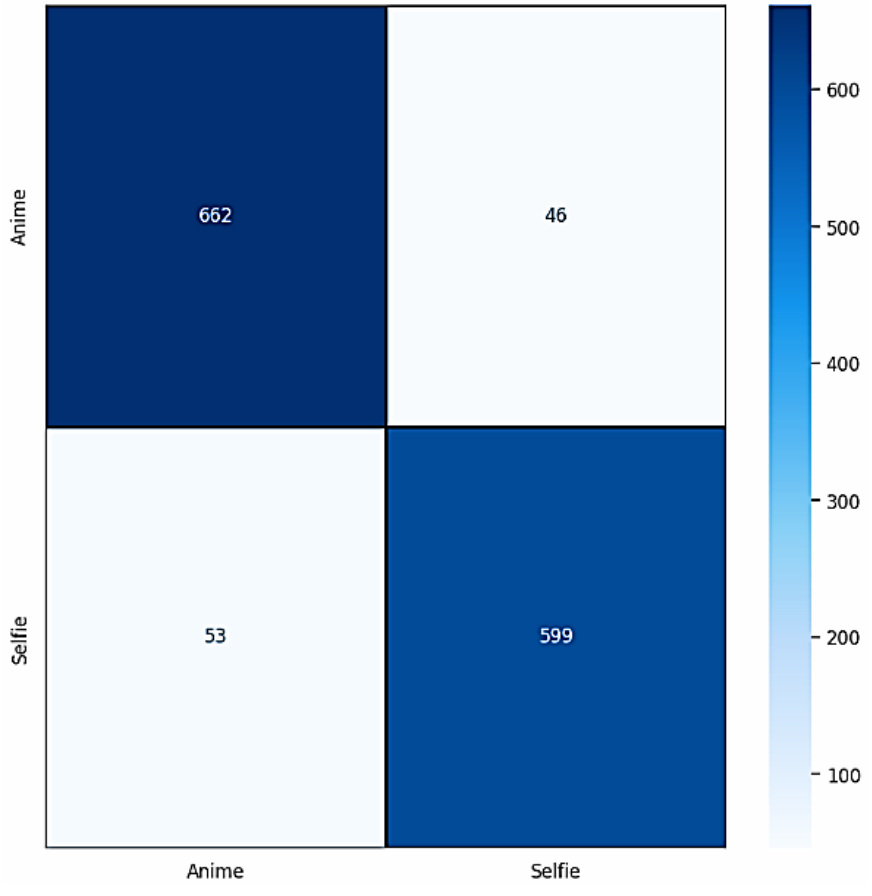
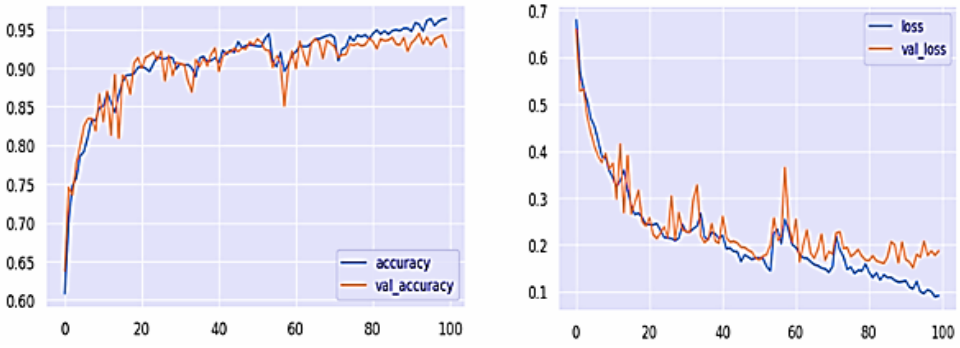


Figure 8 Accuracy and loss graph of model training (see online version for colours)



4.2 Results with proposed models

This work demonstrates the ability of ViT to achieve 99% accuracy in classifying aesthetic anime images and human selfies using the predictive model itself. The global feature extraction via self-attention in ViTs is more powerful than CNN, as it can address intricate spatial relationships, artistic styles, and contextual factors between anime images and human images, as model hyperparameters define in Table 6. Critical configurations of the applied proposed models are the hyperparameters with which the neural networks learn, optimise, and execute in all manners. The different feature extraction strategies are captured using HRNet with multistage modules and ViT with multi-head self-attention on top of the corresponding different model depths. The spatial processing is determined w.r.t. input image size and patch size and ViT proposes patch embeddings to include spatial relationships. The two differ substantially in terms of the number of parameters, i.e., computational demands and memory usage. Specific head architectures suitable for the model types are used to classify it and the choice of activation functions determine the amount of nonlinearity in the network. Learning rate controls size of step forwards that the model's weights updates on, while the optimiser dictate how updates happen with the model's weights during training. Memory usage and gradient estimation quality both depend on batch size: i.e., how many samples will be processed together. In terms of these hyperparameters, we would say that they are finely tuned into a collection of hyperparameters that ensure the best possible model generalisation and convergence. The results shows that both classes have near perfect precision, recall, and F1 score values (0.99), suggesting that the model very rarely misclassifies images. 0.99 shows the macro and weighted averages which further show that the model does not perform in a biased way for any class. Without losing accuracy, the model easily detects small differences in facial structure and colour tones, and stylised artwork. Such a high recall required almost no instances to be undetected, indicating the model's reliability in the real-world classification. Compared to CNN based architectures which typically have trouble with stylistic inconsistencies, and abstract representations commonly found in anime images, this represents a big step forward. Further evidence of the model's robustness comes from the accuracy and loss graphs shown in Figure 9. The study finds that the accuracy curve has consistently high accuracy across both the training sets and the validation sets, indicating that the model learns distinguishing feature efficiently without overfitting. The good convergence rate of the loss function implies that the model runs out relevant

patterns within very short training time to have an efficient computational performance. In contrast to CNN, which rely on localised convolutional filters, ViTs capture global representations using self-attention layers which are noise resilient, lighting stable, and resilient to artistic distortions.

Table 6 Hyperparameter of applied proposed model

Hyper-parameters	*HRNet*	*ViT*	*Description*
Model depth	Multi-stage	Multi-head self-attention	Defines the depth of network architecture and feature extraction approach.
Input image size	120 × 120	120 × 120	The spatial resolution of input images used for training.
Patch size	Not applicable	16 × 16	The size of image patches used in ViT before attention mechanisms.
# Parameter	High	Very high	Indicates the computational complexity and memory footprint.
Classifier	Fully connected	MLP head	Final classification layer used to predict labels.
Activation	ReLU + Final softmax	GELU	Nonlinear function applied to activate neurons.
Optimiser	Adam	AdamW	Optimisation algorithm used for weight updates.
Learning rate	1e-2 to 1e-4 (decay)	1e-2 to 1e-5 (decay)	Determines the step size at each iteration while moving toward a minimum loss.
Batch size	32	32	Number of samples processed before updating model weights.
Dropout	0.2–0.5	0.1–0.3	Regularisation technique to prevent overfitting by randomly deactivating neurons.
Weight initialisation	He initialisation	Xavier initialisation	Strategy for initialising model weights before training.
Training epochs	50–200	100–300	Number of full passes through the training dataset.
Epsilon	1.00E-08	1.00E-08	A small constant to prevent division by zero in optimisation algorithms.
Beta 1	0.9	0.9	Momentum parameter in Adam optimiser controlling moving average of gradients.
Beta 2	0.999	0.999	Second momentum parameter in Adam optimiser.

The result is that the model does not suffer from performance degradation despite being exposed to unseen validation data, which validates the transformer's outstanding learning capability. Confusion matrix in Figure 10 shows that of 708 anime images only three were misclassified as selfies, and of 652 selfies only five were misclassified as anime. The extremely low false positive and false negative confirms that the model is also consistent in its classifications and highly precise. Because localised feature maps of traditional CNN would hinder their ability to recognise abstract concepts, massive

confusion due to style similarities between anime and human appearances when applying CNN to anime and human portrait classification.

Figure 9 Accuracy and loss graph of model training (see online version for colours)

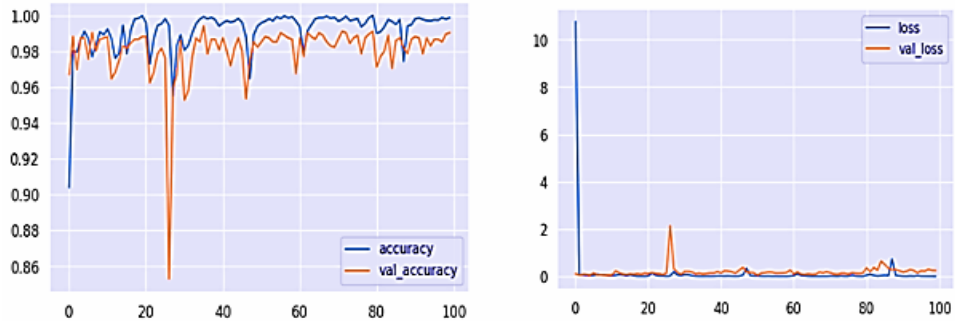
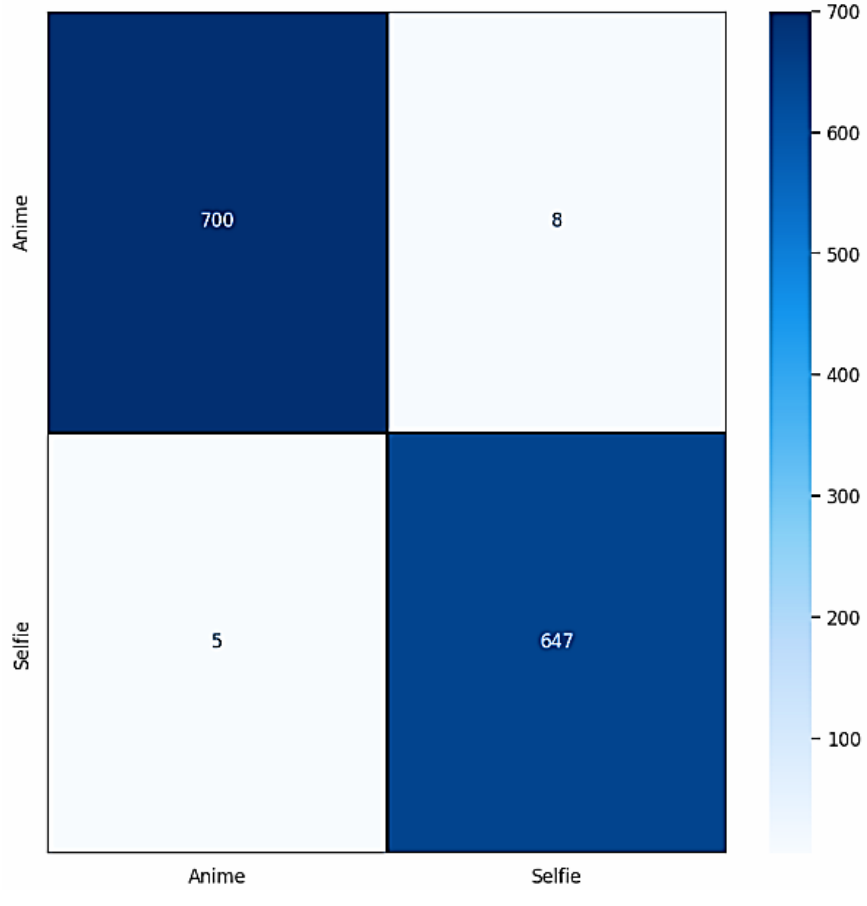


Figure 10 Confusion matrix of proposed ViT model (see online version for colours)



However, ViTs allocate attention to critical image regions dynamically and differentiate accurately. This task is shown to attribute ViT' superior performance to their sensitivity to the manipulation of image patches in sequence, as opposed to fixed receptive fields, as model predictive results based on test data shown in Figure 11. As global dependencies are hard for CNN to capture, ViTs accomplish this thanks to their self-attention mechanism which allows them to compare every pixel with every other pixel. Finally, the CNNs do not perform well in both accuracy and generalisation compared to Vision Transformer, and the model based on Vision Transformer achieves the state-of-the-art results on anime vs. selfie classification. ViTs are shown to combine high accuracy, stable learning curves, and negligible classification errors making them a superior solution for aesthetic classification tasks where stylistic and contextual understanding is critical. For complex image recognition tasks, these results strongly defend the transformer's ability to outperform conventional deep learning models and demonstrate how their self-attention mechanisms and their ability to extract global features make them extraordinarily well suited for this form of image recognition.

4.3 HRNET results

Using the predictive model HRNet has proven to be very proficient in differentiating between anime images and human selfies with 97% accuracy. Instead, HRNet has the strengths in discriminating against fine gangues compared to other conventional models, which benefit quite naturally from feature representations of high resolutions, keeping the spatial information from top to bottom of the network. As a result, this architectural advantage allows the model to learn rich textures, stylistic traits and structural variation, resulting in effective aesthetic classification. The results show HRNet's efficiency with sound arguments. The precision, recall and F1-score values are all consistently high (> 0.97) for both anime and selfie classes, so the source of the model does not appear to weaken its reliability across differing image data distributions.

The anime precision is 0.97, and the selfies is 0.98, indicating that with convoluted details, the model can identify artistic elements distinct from real world properties with extreme precision. Further validating the model's ability to capture extensive, yet important, features, the recall values of 0.98 and 0.97 for anime and selfies, respectively, serve to minimising false negatives and confirming the model's effectiveness on realistic use cases. The model further validates its strong generalisation with accuracy and loss graphs as shown in Figure 12. As all the other curves show a consistently high trajectory, that HRNet is learning complex visual patterns very well in successive epochs. This makes the network stable, as it can still produce high resolution feature maps at many differing scales that accurately capture important image textures and stylisations, meaning they should be stable with different data distributions. At the same time as the loss graph exhibits some fluctuations, the loss graph converges efficiently, implying that HRNet performs error rate minimisation and refining its feature extraction process efficiently. Based on confusion matrix as shown in Figure 13, which helps roughly assess HRNet's decision making accuracy. Over 708 images of anime are deliberately classified as selfies, and successfully correctly misclassify only 15 images of anime as selfies. Indeed, these low false positives and negatives indicate the model's precision. The reason for this strong performance of HRNet is the relative superior multi-scale representation learning provided by its own unique design, able to process global structures and local fine details together. This allows HRNet to maintain high resolution representations

throughout the network, thereby never losing important contextual or stylistically details during learning, which serves as a distinguished ability for HRNet to learn artistic versus real world features. HRNet is an effective model for aesthetic classification in general, and it retains fine grained spatial information, is flexible to image distributions of different classes, and displays stable high accuracy across epochs, as predictive results are defined in Figure 14. HRNet shows its superiority in complex image classification tasks using advanced high resolution feature extraction capabilities which can perform artistic image and real-life image differentiation with high confidence and little misclassification.

Figure 11 ViT classifying model predictive test results (see online version for colours)

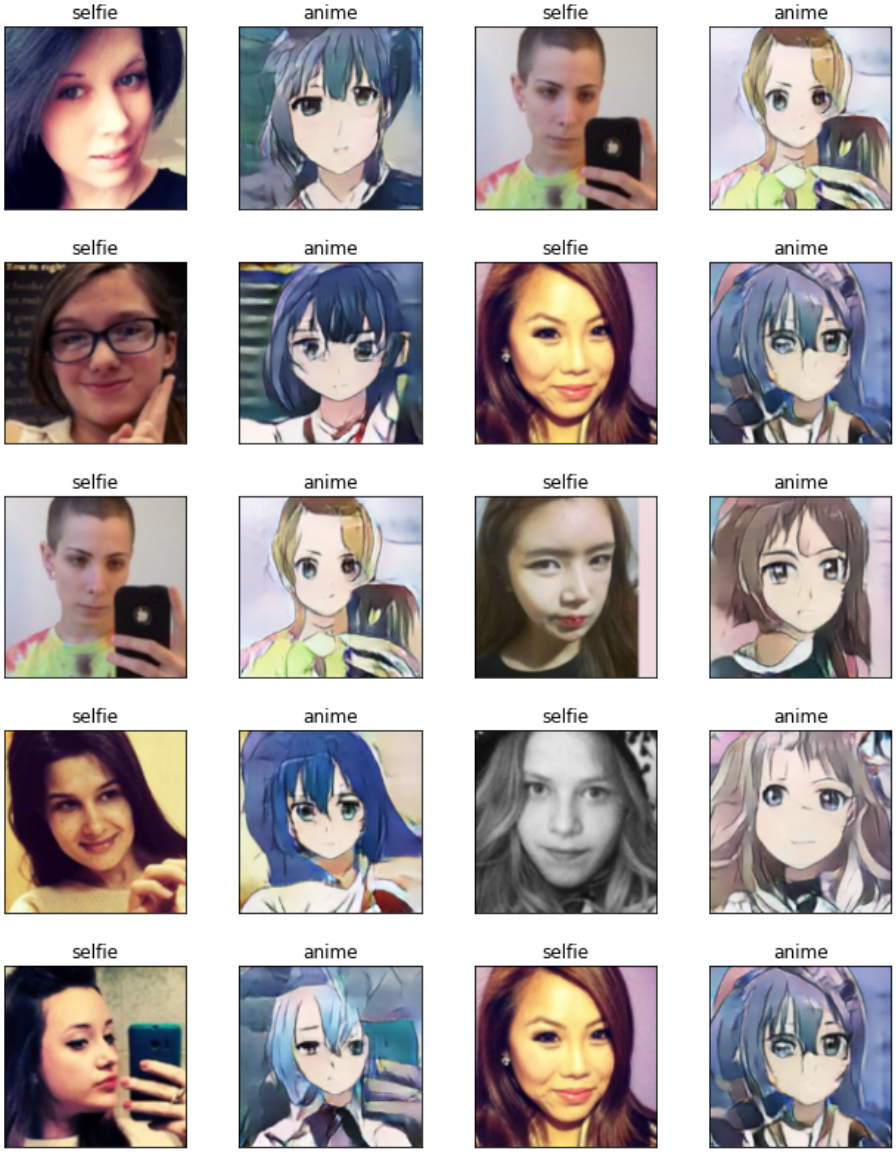


Figure 12 Accuracy and loss graph of model training (see online version for colours)

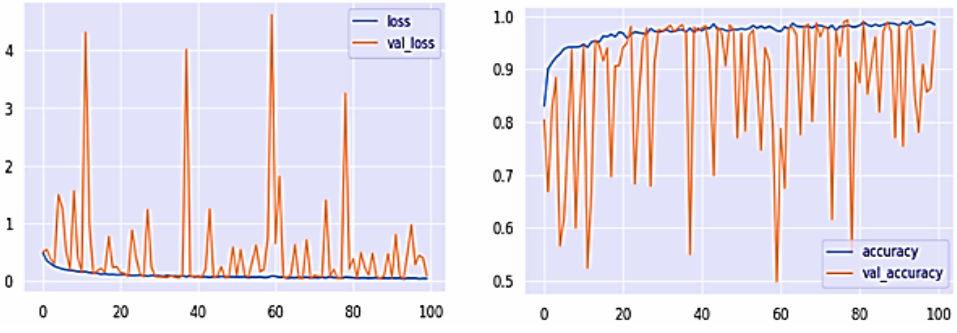


Figure 13 Confusion matrix of proposed HRNet model (see online version for colours)

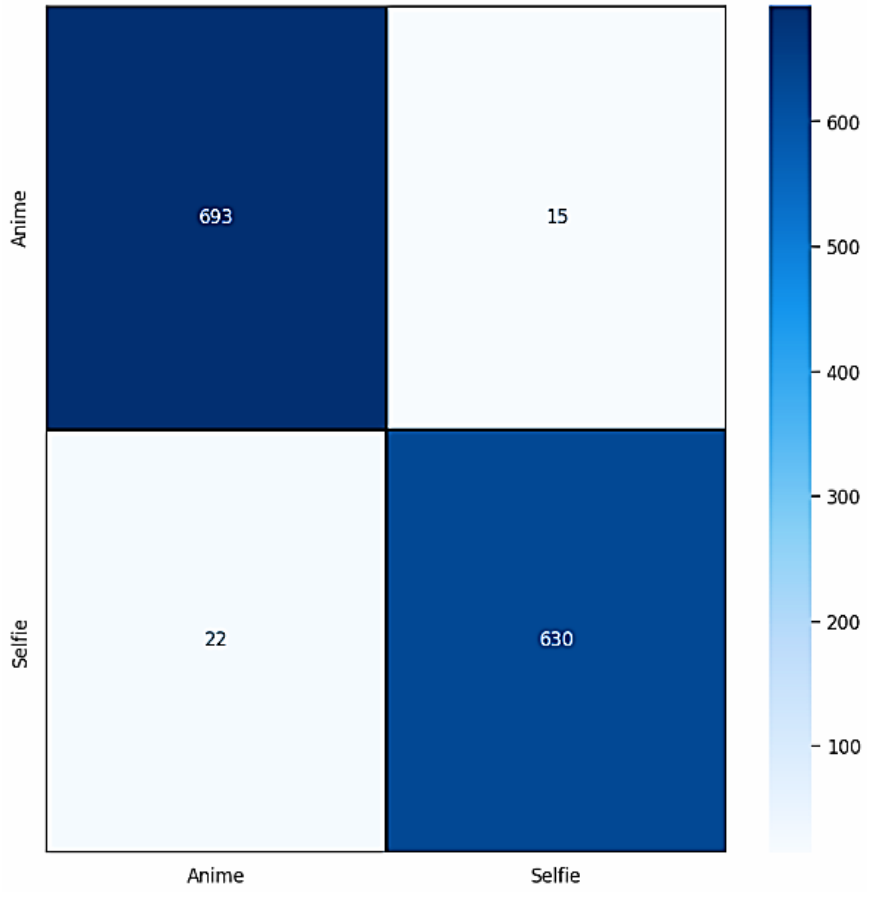
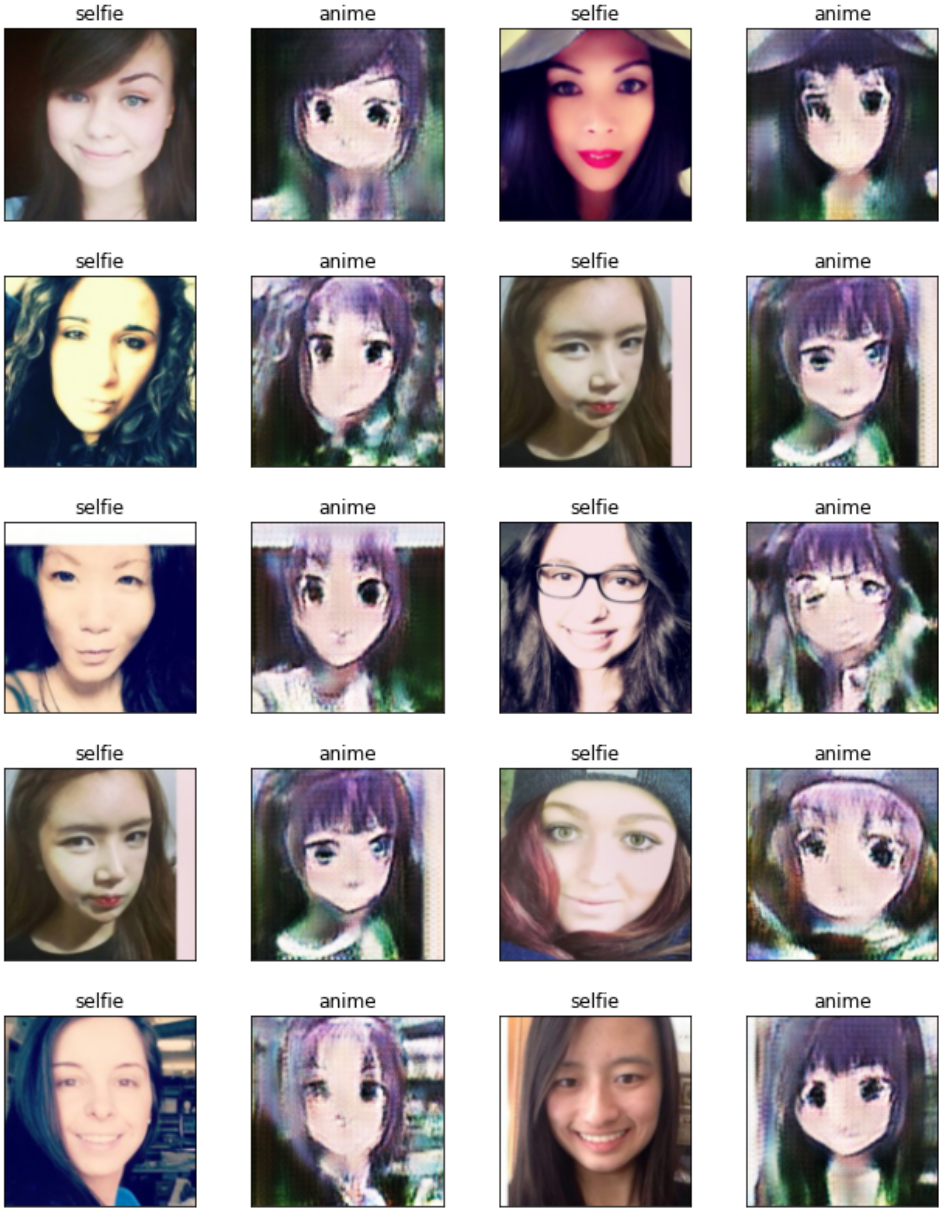


Figure 14 HRNet classifying model predictive test results (see online version for colours)



4.4 Comparison of baseline model with proposed models

As shown by the proposed ViT and HRNet models, the baseline CNN model is significantly outperformed by it in terms of classification accuracy, feature extraction capability, and generalisation performance. CNN achieves a respectable 93% accuracy, but it is hampered by the need to rely on localised receptive fields in order to capture long range dependencies and global contextual information, as comparative analysis shown in

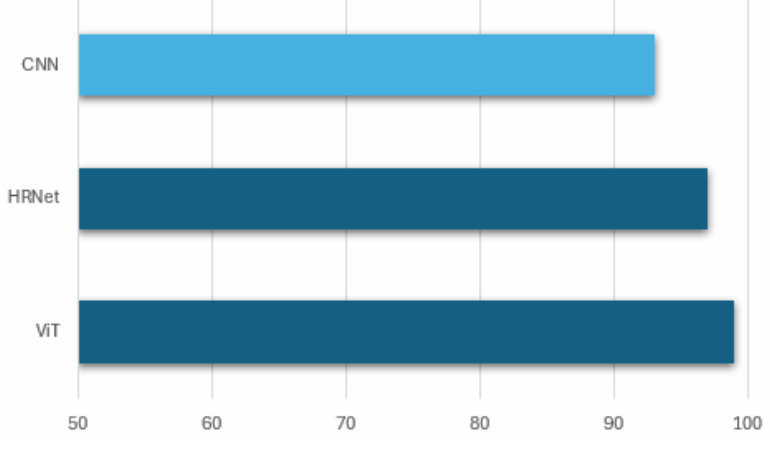
Figure 15. On the other hand, for self-attention of ViT, it learns global relationship with the whole image; it performs better than it with classification accuracy up to 99%, successfully classifying anime from selfie images with almost no misclassification. Similar to CNN, HRNet can preserve finegrained details in high resolution representations throughout the network, yielding a strong classification accuracy of 97%, outperforming CNN's ability to discriminate very subtle artistic and real-world facial attributes. Additionally, the loss and accuracy curves confirm the robustness of the proposed models. Using Table 7, find that the ViT model converges rapidly with little fluctuates, implying efficient learning with superior optimisation stability. HRNet also enjoys a stable convergence, whereas CNN experiences more severe loss and accuracy oscillations indicating less stable feature learning from its hierarchical feature reduction.

Table 7 Comparative results with baseline-model

		Acc	*Pre*	*Re*	*F1*		*Acc*	*Pre*	*Re*	*F1*		*Acc*	*Pre*	*Re*	*F1*
Anime	Baseline-CNN model	93	93	94	93	Proposed-ViT	99	99	99	99	Proposed-HRNet	97	97	98	97
Human selfie		93	92	92	92		99	99	99	99		97	98	97	97

While the baseline CNN is effective, it suffers from highly stylised variations in anime images and complex real world selfie textures, and ViT and HRNet are able to adapt to these varied visual patterns for more reliable classification. Through substantial performance gains, the proposed models very clearly establish themselves as advantageous options for high precision aesthetic image classification tasks, as compared with traditional CNN based approaches.

Figure 15 Analysis of proposed model results with base line model (see online version for colours)



4.5 Comparison with existing studies

The proposed ViT model greatly surpasses state of the art deep learning classifiers for anime and human self selfies. Although effective at generating anime style images, traditional GAN based models like AniGAN attained 77.3 accuracy, an insufficient level

of robustness for recognition. CNN based models, trained on custom anime datasets, achieved an improvement in this measure to 82.9% accuracy; however they are limited at capturing detailed visual patterns because it relies on local feature extraction, as display in Table 8. Unlike such a combination, a more advanced CNN with a GAs was capable of increasing feature optimisation and obtained a 97.2% accuracy on selfie images. But CNN based models tend to lack robustness on diverse dataset and complex spatial relationships. Using the proposed vision Transformer model, achieve an outstanding 99% accuracy, outperforming all previous approaches, which suggests its outstanding ability to learn global contextual dependencies, and high level abstract features. The efficiency of transformer based architectures in visual classification tasks is demonstrated thoroughly by this remarkable performance specifically highlighting the ability of these models to learn the subtle differences between human and anime selfies to an unprecedented precision.

Table 8 Comparison with existing studies

Ref.	*Model*	*Dataset*	*Results (%)*
Li et al. (2021)	AntiGANs	Anime images	77.3
Li et al. (2022)	CNN-based	Custom anime	82.9
Hasan and Mustafa (2022)	CNN + Genetic algorithm	Selfie-anime	97.2
Proposed	*Vision transformers*	*Selfie-anime*	*99*

5 Conclusions and future work

Visual aesthetics classification, especially to differentiate anime images and human selfies, is a challenge in computer vision and deep learning. With continued evolution of digital media comes the need to identify artistic renderings from real images and how they can be used to provide the ability to moderate content, recognise digital art and recommend personalised media. In this paper, a new approach leveraging current state-of-the-art deep learning models such as ViT and HRNet for the classification of aesthetic images with higher accuracy of 99% vs. 97%. The experimental results show that high resolution feature representation and transformer-based architectures allow for better classification performance retaining the global form dependencies while preserving the intricate details. Despite the strong performance, however, this study is limited to image framework and is not yet extended to multimodal fusion, where textual metadata or context can also contribute additional accuracy. Future work will include expanding the dataset to more anime styles and selfie variations and further improving the model's ability to generalise on divergent distributions. In addition, multimodal learning methods will aid in appreciating the aesthetic classifications more introspectively, by combining the image data with textual descriptions. By enabling more interpretable, robust and scalable deep learning models in the imaging classification domain, these improvements will lead the way for more powerful tools.

Declarations

The author declares that he has no conflicts of interest.

References

Bansal, G., Nawal, A., Chamola, V. and Herencsar, N. (2024) 'Revolutionizing visuals: the role of generative AI in modern image generation', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 11, pp.1–22.

Bekhet, S., Alghamdi, A.M. and Taj-Eddin, I. (2022) 'Gender recognition from unconstrained selfie images: a convolutional neural network approach', *International Journal of Electrical & Computer Engineering*, Vol. 12, No. 2, pp.2066–2078.

Chen, S. and Zwicker, M. (2022) 'Transfer learning for pose estimation of illustrated characters', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.793–802.

Chen, S., Zhang, K., Shi, Y., Wang, H., Zhu, Y., Song, G., An, S., Kristjansson, J., Yang, X. and Zwicker, M. (2023) 'Panic-3d: stylized single-view 3D reconstruction from portraits of anime characters', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.21068–21077.

Hasan, B.M.S. and Mustafa, R.J. (2022) 'A study of gender classification techniques based on iris images: a deep survey and analysis', *Science Journal of University of Zakho*, Vol. 10, No. 4, pp.222–234.

Hou, L. and Pan, X. (2023) 'Aesthetics of hotel photos and its impact on consumer engagement: a computer vision approach', *Tourism Management*, Vol. 94, No. 9, p.104653.

Huang, L. and Zheng, P. (2023) 'Human-computer collaborative visual design creation assisted by artificial intelligence', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 9, pp.1–21.

Jiang, Y., Jiang, L., Yang, S. and Loy, C.C. (2023) 'Scenimefy: learning to craft anime scene via semi-supervised image-to-image translation', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.7357–7367.

Khan, U., Khan, H.U., Iqbal, S. and Munir, H. (2024) 'Four decades of image processing: a bibliometric analysis', *Library Hi Tech*, Vol. 42, No. 1, pp.180–202.

Lan, Z., Maeda, K., Ogawa, T. and Haseyama, M. (2022) 'GCN-based multi-modal multi-label attribute classification in anime illustration using domain-specific semantic features', in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, October, pp.2021–2025.

Li, B., Zhu, Y., Wang, Y., Lin, C.W., Ghanem, B. and Shen, L. (2021) 'Anigan: style-guided generative adversarial networks for unsupervised anime face generation', *IEEE Transactions on Multimedia*, Vol. 24, No. 9, pp.4077–4091.

Li, H., Guo, S., Lyu, K., Yang, X., Chen, T., Zhu, J. and Zeng, H. (2022) 'A challenging benchmark of anime style recognition', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4721–4730.

Li, Y. and Zhang, Q. (2024) 'The analysis of aesthetic preferences for cultural and creative design trends under artificial intelligence', *IEEE Access*, Vol. 12, pp.158799–158808, 2024, DOI: 10.1109/ACCESS.2024.3486031.

Li, Z. and Mu, K. (2023) 'Meta-HRNet: a high resolution network for coarse-to-fine few-shot classification', in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer Nature Switzerland, Cham, September, pp.471–487.

Lv, P., Fan, J., Nie, X., Dong, W., Jiang, X., Zhou, B., Xu, M. and Xu, C. (2021) 'User-guided personalized image aesthetic assessment based on deep reinforcement learning', *IEEE Transactions on Multimedia*, Vol. 25, No. 11, pp.736–749.

Naftali, M.G., Sulistyawan, J.S. and Julian, K. (2022) *AniWho: A Quick and Accurate Way to Classify Anime Character Faces in Images*, arXiv preprint arXiv:2208.11012.

Naz, A., Khan, H.U., Alesawi, S., Abouola, O.I., Daud, A. and Ramzan, M. (2024) 'AI knows you: deep learning model for prediction of extroversion personality trait', *IEEE Access*, Vol. 12, pp.159152–159175, DOI: 10.1109/ACCESS.2024.3486578.

Rios, E.A., Cheng, W.H. and Lai, B.C. (2021) *Daf: Re: A Challenging, Crowd-sourced, Large-scale, Long-tailed Dataset for Anime Character Recognition*, arXiv preprint arXiv:2101.08674.

Sardenberg, V., Guatelli, I. and Becker, M. (2024) 'A computational framework for aesthetic preferences in architecture using computer vision and artificial neural networks', *International Journal of Architectural Computing*, Vol. 23, No. 1, p.14780771241279350.

Singh, A. and Singh, V.K. (2024) 'A hybrid transformer-sequencer approach for age and gender classification from in-wild facial images', *Neural Computing and Applications*, Vol. 36, No. 3, pp.1149–1165.

Talha, M.M., Khan, H.U., Iqbal, S., Alghobiri, M., Iqbal, T. and Fayyaz, M. (2023) 'Deep learning in news recommender systems: a comprehensive survey, challenges and future trends', *Neurocomputing*, Vol. 562, No. 12, p.126881.

Wang, H. (2025) 'An investigation into the evaluation and optimisation method of environmental art design based on image processing and computer vision', *Scalable Computing: Practice and Experience*, Vol. 26, No. 1, pp.277–286.

Wang, R. (2021) 'Computer-aided interaction of visual communication technology and art in new media scenes', *Computer-Aided Design and Applications*, Vol. 19, No. S3, pp.75–84.

Wang, Z., Mao, X., Zhang, Z., Zhang, J., Xu, S. and Zhang, X. (2023) 'Teaching neural networks to imitate human habits for recognizing anime characters', in *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, December, pp.100–106.

Zheng, Y., Zhao, Y., Ren, M., Yan, H., Lu, X., Liu, J. and Li, J. (2020) 'Cartoon face recognition: a benchmark dataset', in *Proceedings of the 28th ACM International Conference on Multimedia*, October, pp.2264–2272.