Multi-task Learning for Intelligent Portrait Composition: Deep Residual Networks and Human Pose Estimation

Di Shao
School of Journalism and Communication, Anhui Broadcasting Movie and Television College
Hefei 230000, China
Email: amtccc@163.com

Keywords: multi-task learning, human pose, portrait, composition, deep residual network, adaptive point set

Received: January 23, 2025

Traditional photography composition is difficult to meet the current needs and aesthetic preferences of portrait photography. Therefore, to improve the composition quality and efficiency of portrait photography, an intelligent composition technology combining multi-task learning framework and deep learning technology is proposed. Firstly, taking the deep residual network as the basic skeleton, the lightweight classification network MobileNet V2 is introduced to improve the portrait detection performance. Afterwards, image quality is improved through the multi-scale feature fusion and multi-task learning, and the scale feature extraction is performed on the input image. Moreover, the human pose estimation network is used to detect human keypoints, dynamically adapting to different human poses and scales. The relay node is used to associate the relationship between human instances and keypoints, improving the intelligent composition effect. The effectiveness of the research method is analyzed from three aspects: intelligent cropping for portrait detection, pose recognition analysis, and synthesized image quality. The MS COCO and MuPoTS-3D datasets are selected for evaluation, including recall rate, F1-value, average Intersection over Union ratio (Avg IoU), mismatch, Area under Curve (AUC), and precision-recall curve (PCK). The results showed that the proposed MobileNet V2-ResNet50 achieved an accuracy of 94.51% in extracting facial image information, with an Avg IoU of 0.69, while the Avg IoU of the other three comparison methods was less than 0.65. The proposed fusion algorithm had information entropy, structural similarity, average structural similarity, peak signal-to-noise ratio, brightness relationship factor, and mutual information of 7.596, 1.129, 7.081, 1.828, 1.078, and 8.826, respectively. The overall image quality fusion effect was significantly better than other algorithms. The MobileNet V2-ResNet50 network significantly outperforms the UNet-Transformer-CBAM and MCTN models in terms of computational efficiency, with the floating-point operations of only 2.5B, parameter count of 5M, and inference time of 15ms. The research method achieved a human pose detection accuracy of 88.94% on the dataset, with AUC and PCK scores of 44.99 and 83.24, respectively. The detection accuracy of RSC-MS, OP-GAN, and PGF-HPE models did not exceed 88.5%, and their AUC and PCK scores differed significantly from the research method. This method can effectively identify and optimize portrait composition, enhancing the visual effect and artistic expression of photos.

Povzetek: Opisano je večopravkovno učenje z ResNet in MobileNet V2 za inteligentno kompozicijo portretov, ki združuje zaznavo poze, obrezovanje in estetsko optimizacijo z visoko natančnostjo.

1 Introduction

With the widespread popularity of electronic devices and changes in information dissemination media, the image, as an information carrier, presents an explosive growth trend. Higher requirements have been put forward for the photography and composition skills of works [1]. A good composition not only presents complete and interesting image information, but also brings strong visual impact and clear thematic sense to people, endowing ordinary things with unique beauty [2]. Photography needs and aesthetic preferences make people more inclined to crop images to improve aesthetic quality. Portrait photography not only requires capturing the image of a person, but also conveying emotions and imagery through clever composition. Traditional photography composition relies on the photographer's experience and aesthetics. However, intelligent composition uses artificial intelligence technology to automatically identify and analyze shooting scenes, optimize image layout, and make photos more visually impactful and aesthetically pleasing [3]. In previous studies, intelligent composition in portrait photography mainly relied on a single image processing technique, such as edge detection, feature matching, etc. Although these methods have achieved certain results, they have obvious limitations. Farhat et al. used a photography composition assistance framework to improve composition effects. A fork join framework was used to achieve photography composition, which included designing from aspects

such as object detection, photo classification, and pose clustering analysis. The results indicated that this method had good photographic application effects [4]. Li et al. used lightweight networks to preserve fine-grained details in images for portrait details in complex backgrounds. The experimental results showed that this method had good visual effects and performed well in evaluation indicators [5]. Considering that existing automatic image cropping methods often use image boundaries as boundaries, it is difficult to improve the aesthetic appeal of image cropping. Therefore, Zhong et al. used the field of view evaluation model to determine the extension of image content and implement cropping based on image quality. The results indicated that this method could effectively improve the visual image quality [6]. Wang et al. proposed a content aware image scaling mechanism based on composition detection and composition rules, which optimized the visual effect of images by calculating composition rules. The results indicated that this method could preserve the image content structure and overall aesthetic appeal [7]. Li et al. used attribute assisted multimodal memory network to improve the quality of image aesthetic evaluation, inferred aesthetic quality by associating attribute perception semantic information. The results indicated that this method had good discriminative power on the dataset [8]. The relevant work situation is summarized, as shown in Table 1.

Table 1: Summary of related work

Literature	Method	Method Details	Performance results	Limitation
[4]	Photography composition auxiliary framework	Design photography composition from the aspects of object detection, photo classification, pose clustering analysis, etc	The photography composition effect is good and can meet personal photography preferences	Insufficient attention to image scale features
[5]	Lightweight network	Preserve fine-grained details of the image	The visual accuracy exceeds 90%	The accuracy of posture details is slightly lower
[6]	Field of View Evaluation Model	Crop based on image quality	Can preserve and preserve the content	Not considering the cropping relationship
			structure and overall aesthetic of the image	between images, scenes, and characters
[7]	Content aware image scaling mechanism	Calculate composition rules	Balancing image content structure and aesthetics	Strong dependence on rules, strong subjectivity, and limited adaptability to complex scenarios
[8]	Attribute assisted multimodal memory network	Association attribute perception semantic information to infer aesthetic quality	The difference in aesthetic quality rating is relatively small	Semantic information extraction has limitations and cross generalization ability

With the explosive growth in the number of images, portrait photography puts higher requirements on composition techniques. Traditional intelligent composition methods rely on a single image processing technique (e.g., edge detection, feature matching, and key point detection) and ignore task correlation, leading to poor results in complex scenes (e.g., multiple people, and occlusion). Existing methods have obvious deficiencies in speed, accuracy, and aesthetic optimization, making it difficult to meet users' needs for high-quality composition. In addition, traditional aesthetic evaluation methods mostly rely on manual rules and lack in-depth understanding of image content (such as human posture, emotional expression, etc.). Therefore, to construct an efficient and robust intelligent composition model, solve the composition optimization in complex scenes, achieve real-time and aesthetic quality improvement, and promote the practical application and popularisation of intelligent composition technology, the study proposes to improve it while using neural network processing. Specifically, to improve the quality of image, this study introduces a multi-task learning framework that integrates portrait detection, pose estimation, and intelligent cropping, optimizing image composition while adapting to diverse human poses. The proposed method enhances accuracy and artistic quality compared with conventional approaches. The implementation of multitask learning mainly focuses on regression tasks, adding classification and comparison tasks to calculate image quality score labels and

regression losses, respectively. Afterwards, the input image is extracted with different scales of features, and then dimensionality reduction, feature node generation, feature fusion and other steps are carried out. The corresponding quality score is used as a label to calculate the comparative loss composed of quality ranking loss and quality relative deviation loss. The quality score and label score of the image are calculated to obtain its coarse-grained and fine-grained features, thereby improving the image quality.

The innovation of the research lies in two aspects. One of them is the multi-task learning framework, which can improve image quality scores while achieving basic facial feature detection. The second is to jointly process human pose recognition and photography composition optimization as two closely related tasks, that is, based on the structure of human pose, further predicting and querying pose keypoints to grasp the semantic and positional information of each keypoint. The proposed method breaks the overly single learning focus of early research, strengthens the integration and collaboration between human pose recognition and photographic composition, and can provide more reference guidance for intelligent composition while considering the diversity of character poses and the artistic nature of photographic composition.

2 Methods

To achieve the design and real-time optimization of multi-task learning networks in the research objectives, this study first adopts a lightweight design strategy to construct an efficient multi-task detection model, which uses ResNet50 as the basic network architecture and MobileNet V2 as the support in the structural design part. The improvement approach of MobileNet V2 can enrich the output features of the receptive field and perform convolution operations on the reduced dimensional features to refine semantic information. The multi-task learning approach is used to calculate the regression task, classification task, and comparison task of the image separately, and calculate the image quality score label and regression loss separately. Different scale features are extracted from the input image, and the corresponding quality scores are used as labels to calculate the quality scores and label scores of the image, and the coarse-grained and fine-grained features of the image are obtained. By fusing feature representations, the computational complexity is reduced and the operational performance of the model is improved to meet application requirements. Secondly, to improve the accuracy of human pose analysis and optimize composition aesthetics, a multi-person pose estimation algorithm is proposed, which enables the network structure to dynamically adapt to different human poses and scales, and achieves intelligent portrait cropping and aesthetic preferences through keypoint detection and positioning, pose analysis, and adjustment.

2.1 Multi-task detection of portrait photography images based on lightweight design

The essence of image composition is the conscious arrangement of elements to convey specific emotions, information, and aesthetics. Intelligent composition, as a common image processing method, can effectively improve composition effects and enhance image quality. The Residual Network 50 (ResNet50) performs well in tasks such as image classification and object detection due to its powerful feature learning ability. Therefore, the research takes the ResNet50 as the basic network architecture. The MobileNet V2 is taken as the support in the structural design. A multi-task-based image composition optimization method and a lightweight design concept for portrait detection are designed to improve the quality of image evaluation. The point by point convolution processing of MobileNet V2 can form an inverted residual structure, which reduces the number of parameters while ensuring channel information extraction. The improvement ideas of convolutional networks are reflected in two aspects. One is to introduce width multipliers and resolution parameters into the original number of input layers, convolution kernel size, and output layer channel for processing. The second is to improve the accuracy of object detection by superimposing dilated convolutions in the output part of feature extraction. The proven performance of ResNet50 in image classification and object detection tasks makes it a reliable choice for multi-task learning. The lightweight MobileNet V2 architecture effectively optimizes computational efficiency and requires less deployment environment. Although EfficientNet provides a balance between accuracy and efficiency through composite scaling, it may not provide the same feature richness as ResNet50 for multi-task scenarios and consumes more computational resources. Visual transformers require large-scale datasets and computational resources, making them less suitable for real-time applications on mobile devices. ResNet50 provides the feature richness necessary for accurate multi-task detection, while MobileNet V2 ensures that the model remains lightweight and efficient. This combination achieves a balance between high performance and real-time applicability, which is crucial for intelligent image synthesis in portraiture. Figure 1 displays the detection network structure.

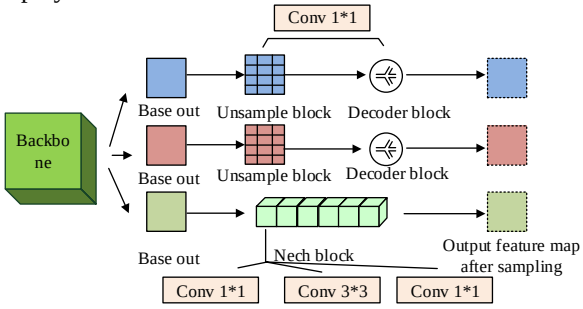


Figure 1: Schematic diagram of detection network structure

In Figure 1, the Backbone denotes MobileNetV2, which is the backbone part of the network and is responsible for extracting base features from the input image. Deep separable convolution reduces the number of parameters and computation and improves the efficiency of the model. Base out refers to the intermediate feature extraction output layer, which provides the initial feature representation as the subsequent input. The feature maps in this layer usually have high semantic information, but the sensory field may be limited. Upsample Block refers to the upsampling module, which is used to upsample the low-resolution feature maps to a high resolution. Decoder Block is used to fuse high-level semantic features with low-level detail features to improve the model's ability to capture target location and details. Neck Block is the Expansion Convolution Block, which generates features with multiple receptive fields through different expansion rates for enhanced feature extraction. The stacked Expansion Convolution Blocks can expand the receptive fields. The 1×1 convolution in this layer is to downscale the channel and reduce the computation. 3×3 convolution is to refine the semantic context information. The core of the improvement lies in the Neck Block design, which improves the sensory field and multi-scale detection ability of the model by expanding convolution and feature fusion. Afterwards, to improve the intelligent composition effect, the multi-task learning is used to enhance image quality, with regression tasks as the main focus, and additional classification and comparison tasks introduced to calculate quality score labels and regression losses. Feature maps extracted at different scales are processed through dimensionality reduction, node generation, and feature fusion to enhance image representation. Figure 2 is a schematic diagram of the multi-task learning process.

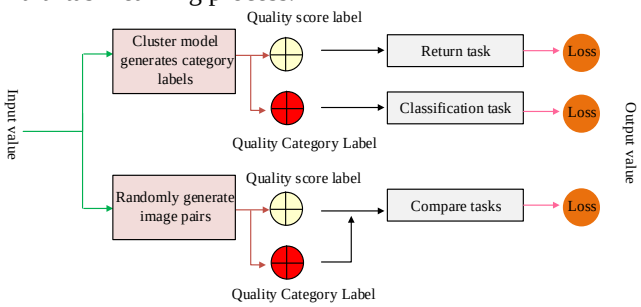


Figure 2: Schematic diagram of multi-task learning process

In Figure 2, the input images are subjected to category label generation and random image pair generation. Regression, classification, and comparison tasks are performed based on the quality score and label score of the image. The classification task first classifies the quality scores into two categories (good/bad) through a clustering model and then calculates the classification loss using a single image as input. The quality score of the image can be calculated based on the comparison loss under the combination of classification loss, regression loss, and other results. All quality scores are clustered to obtain image quality categories and confidence levels. The goal of classification task is to assign images to predefined categories to reduce subjective bias. It can be used as a supplement to regression tasks to reduce the subjective bias associated with a single regression task. The comparison task randomly selects image pairs from the image set, uses their quality scores as labels, and calculates the comparison loss composed of quality ranking loss and quality relative deviation loss. The corresponding quality score is used as the label to calculate the comparison loss consisting of quality ranking loss and relative quality deviation loss. Equation (1) can be used to represent the weighted sum of the loss function [9].

$$Loss = \alpha L_R + \beta L_C + \lambda L_B \tag{1}$$

In equation (1), L_R represents the regression loss. L_C represents the classification loss. L_B represents the comparison loss. α, β, λ are weights of the corresponding loss function. The introduced classification loss and comparative loss reduces subjective bias in quality scoring and enables the model to obtain coarse-grained and fine-grained features, improving the generalization ability of model sharing [10]. The study takes L1 norm loss to calculate the regression task of a single image, as shown in equation (2).

$$L_R = \frac{1}{N}\sum_{i=1}^{N}\left|S_i - \hat{S}_i\right| \tag{2}$$

In equation (2), N represents the number of images. S_i is the true score of the i-th image. $\hat{S}_i$ is the predicted score for the i-th image [11]. In portrait photography, some pixel values may be abnormal due to lighting or occlusion problems, the data may have noise or outliers, and L1 loss can better handle these situations. L1 loss can directly measure the pixel level error, which is more suitable for the continuity and local characteristics of image data. It is more inclined to produce sparse solutions, better preserving image details and edge information [12]. In image clustering processing, the Gaussian mixture model (GMM) is used to obtain its quality category. Equation (3) is the probability density function of the GMM.

$$p(X) = \sum_{k=1}^{K} w_k \cdot p(X \mid \mu_k, \sigma_k) \tag{3}$$

In equation (3), μ_k represents the mean. σ_k is the variance. w_k is the weight of the k-th Gaussian distribution. $p(X \mid \mu_k, \sigma_k)$ is the probability density function. GMMs model data can fit any complex data distribution through the linear combination of multiple Gaussian distributions. In image clustering tasks, image features (such as color and texture) often have complex distribution characteristics, which can be well captured by GMMs. Unlike K-means clustering methods, GMM can effectively evaluate the probability distribution of image quality categories, more flexibly define the boundaries between quality categories and adjust the

number of categories, so as to improve the clustering results [13]. By clustering all quality scores, image quality categories and confidence levels can be obtained. The prediction results of classified images can be obtained through the cross entropy loss function, which is shown in equation (4).

$$L_c = -[C_i \times \log(p_i) + (1 - C_i) \times \log(1 - p_i)) \quad (4)$$

In equation (4), C_i represents the quality classification of the image. p_i is the prediction probability of the image. The main task of comparing images is to determine the quality results of two images. Previous methods often encounter sorting errors when faced with inconsistent prediction ranking results [14]. To improve the performance of the model in comparing information, a loss function based on the concept of relative deviation is proposed for definition, as shown in equation (5).

$$L_{RD} = \left| \frac{S_1 - S_2}{S_1 + S_2} - \frac{\hat{S}_1 - \hat{S}_2}{\hat{S}_1 + \hat{S}_2} \right| \quad (5)$$

In equation (5), S_1 and S_2 represent the true scores of the image. $\hat{S}_1$ and $\hat{S}_2$ are the predicted scores for the image. To ensure the intelligent composition performance and scene applicability of portrait images, the above quality evaluation model has been light-weighted to improve the inference speed of the model. Firstly, the label data is expanded through knowledge distillation. The ResNet50 skeleton network is trained with lightweight sub networks and module fusion is performed. Considering that traditional knowledge distillation mostly focuses on discrete labels, the study investigates corresponding processing of image quality, as shown in Figure 3.

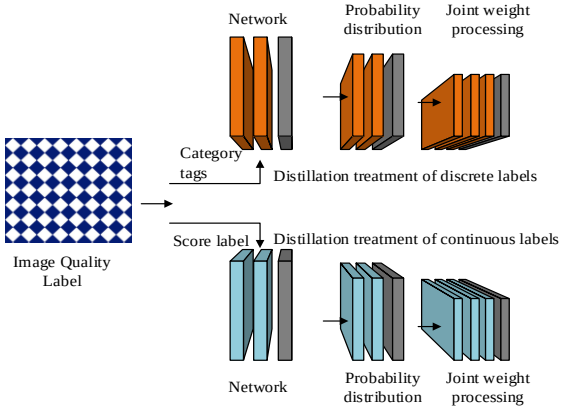


Figure 3: Image quality processing content

On discrete data, knowledge distillation is still used to process quality category labels, which can be expressed as equation (6).

$$L_L = \frac{1}{N} \sum_{i=1}^{N} \omega \cdot Loss(\varsigma(Z), C_L)_i + (1 - \omega) \cdot Loss(\varsigma(Z)))_i \quad (6)$$

In equation (6), the real image C_L and the logistic regression network Z are normalized by ς to obtain the loss value under its loss weight ω. T is the distillation temperature. N represents the number of samples.

2.2 Intelligent portrait composition based on human pose analysis

Intelligent composition systems for portrait photography often need to operate in real-time or near real-time conditions, especially in mobile devices or embedded systems. Lightweight design can significantly reduce computational complexity, memory footprint, and computational load, enabling systems to operate in resource-constrained environments. This improves the response speed of the system. In Section 2.1, the network with lightweight design can reduce the number of parameters and maintain high feature extraction ability. Therefore, multiple tasks can be calculated in parallel, which can provide computational support for multi-person recognition of human posture analysis and complete multi-task detection. Intelligent composition needs to understand the human body's pose, position, scale, and other information, which relies on the multi-task detection results in Section 2.1. The feature information output from section 2.1 is an important basis for structured portrait data analysis. The lightweight design of multi-task detection for portrait photography images provides an efficient and accurate preprocessing method for human pose analysis. The human pose analysis can guide the implementation of intelligent composition to achieve autonomous composition based on differences in human pose, thereby better meeting aesthetic requirements [15]. The research mainly analyzes the estimation algorithm for multi-person poses and proposes a single-stage differentiable multi-person pose estimation approach, which introduces additional features at adaptive human body local correlation points and uses relay nodes to associate the relationship between human body instances and key points. The features of different related points can be regressed to the key point positions within the corresponding regions, thus establishing position point connections in a unidirectional process [16]. Figure 4 shows the structure of the adaptive pose network.

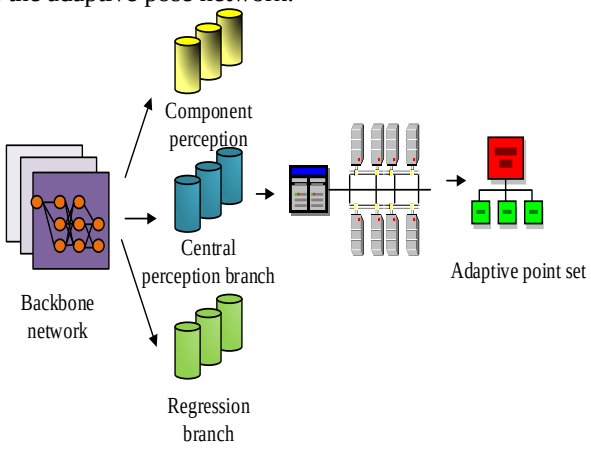


Figure 4: Adaptive pose network structure

In the human body representation section, the study utilizes a pixel by pixel keypoint regression framework, which represents human instances using a center point and points related to human body parts (adaptive points). The adaptive approach is used to represent local correlation points. This process can be explained, as displayed in equation (7).

$$EC \to \{P_h, P_s, P_a, P_r, P_l\} \tag{7}$$

In equation (7), P_h, P_s, P_a, P_r, and P_l respectively represent the head, shoulders, arms, buttocks, and legs. Unlike analyzing the offset vectors of reference points and keypoints, the study represents the human body as an adaptive point set, which can dynamically adapt to different human poses and scales through network learning. Figure 5 shows the overall network architecture.

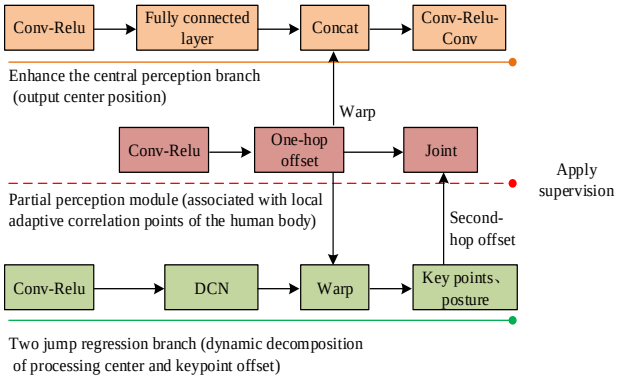


Figure 5: Schematic diagram of overall network architecture design

In Figure 5, the network architecture consists of three parts: component perception, center perception branch, and regression branch. The backbone network extracts visual features, and the component perception module associates local adaptive related points of the human body. The receptive field uses aggregation to predict the center of the human body at different scales and poses. In the keypoint regression section, the two-hop localization method is a phased strategy to locate key points, which aims to improve the accuracy and robustness of key point localization. Specifically, the two-hop method includes a first hop offset and a second hop offset. The first hop is the offset predicted from the center point of the human body to the adaptive part related points, which uses variable convolution operations to enhance feature extraction capabilities and locate adaptive part related points. The second step is to predict the offset from the adaptive part related points to specific keypoints, and then use variable convolution operations to capture local details and locate specific keypoints. Based on two-hop method, the key point localization problem is decomposed into two sub-tasks, reducing the difficulty of each task. Staged positioning can dynamically adjust the receptive field, adapt to the diversity of human postures and local deformations, and improve the robustness and accuracy of the model. The adaptive receptive field can predict the center localization of a single channel heatmap. The loss of the center perception branch is represented by a pixel by pixel function, as shown in equation (8) [17-18].

$$L_{ct} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}(1-\overline{P}_c)^{\alpha}\ln(\overline{P}_c) & , P_c = 1\\ (1-P_c)^{\beta}\ln(1-P_c) & , P_c \neq 1\end{cases} \tag{8}$$

In equation (8), $\overline{P}_c$ represents the predicted value. P_c represents the label value. α and β are hyperparameters. In the keypoint regression section, the two-hop method is used to implement regression branches, which involves performing variable convolution operations on features to determine their corresponding positions and locate keypoint positions. Two-hop includes the first offset $\overline{off}_1$ and the second offset $\overline{off}_2$, as shown in equation (9).

$$\overline{off} = \overline{off}_1 + \overline{off}_2 \tag{9}$$

The first hop is from the center of the human body to the relevant point of the adaptive part, and the second hop is from the relevant point of the adaptive part to a specific key point. The study uses bone perception regression loss to represent the bone links within human body regions to improve localization accuracy. Meanwhile, parallel scores are added to the backbone network during multi-task training to generate keypoint heatmap labels. The center perception branch and the two-hop regression branch output the center position and offset position, respectively. Finally, the candidate pose center positions are selected using a pooling kernel. The pose decoding results can be obtained by combining the predicted center and keypoint offset [19]. In response to the quality estimation problem in multi-person pose regression analysis, this study uses consistency instance scores and query encoding modules to represent human instance position scores and pose scores. The regression quality is measured by the similarity of key points between predicted pose and real pose, as shown in equation (10).

$$KPS_p = \frac{\sum_i \exp(\frac{-d_{p,i}^2}{2s_p^2 k_i^2})\lambda(v_{p,i} > 0)}{\sum_i \lambda(v_{p,i} > 0)} \tag{10}$$

In equation (10), $d_{p,i}$ is the distance between the predicted position of key point i and the label position. S_p is the scale corresponding to pose k. k_i is the annotation error of key points. By solving the confidence level in equation (10), the consistency instance score can be obtained. Query encoding includes keypoint queries and pose queries. When processing model tasks or features in parallel, multi-branch structures perform separate prediction and regression processing on keypoints of different types or parts. Bilinear interpolation can be used to extract and aggregate the semantic and positional information of each keypoint [20]. The transformed keypoint regression features can be expressed as equation (11).

$$R_k(c)' = \sum_{n=1}^{k} \{R_k(c)(c + \overline{D}_{c \to q} + \overline{D}_{q \to q'_n})\} \quad (11)$$

In equation (11), q' represents the semantic point. q is the key point query. $R_k(c)$ represents the regression feature of the original keypoints at the prediction center c. $\overline{D}_{c \to q}$ is the offset from the center to the key point. n represents the number of semantic points, and $\overline{D}_{q \to q'_n}$ is the offset of the semantic meaning of the key points queried. After the key point query position is processed by bilinear interpolation and connected through the channel dimension, new pose instance features can be obtained, as expressed in equation (12).

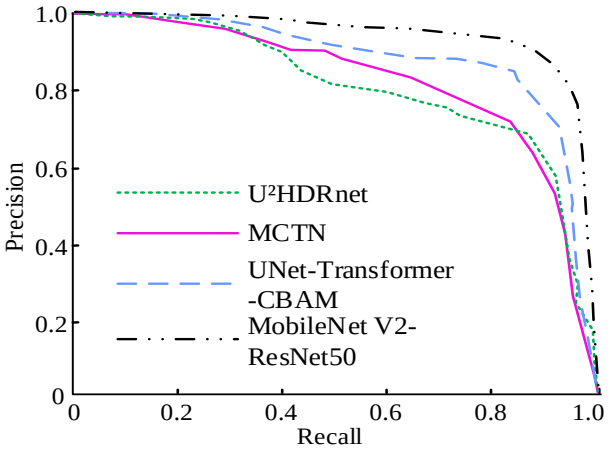
$$R_I(c)' = Concat(\{R_I(c)(c + \overline{D}^{i}_{c \to q})\}_{i=1}^{U}) \quad (12)$$

In equation (12), U represents the number of keypoint queries. $\overline{D}^{i}_{c \to q}$ is the offset from the center to the key point query. $R_I(c)$ represents the posture instance feature. In pose query encoding, it is possible to incorporate structural pose information into instance feature representations and perform consistency processing on their instance score labels. Through key point detection and positioning, posture analysis and adjustment, the position and posture of the portrait in the image can be adjusted. It can realize the human pose estimation, which is more in line with the aesthetic principles and user preferences and meets the needs of personalized settings. Hyperparameters are optimized using grid search, ensuring that model performance is balanced between computational efficiency and accuracy. The batch size is selected to optimize processing speed without compromising feature extraction.

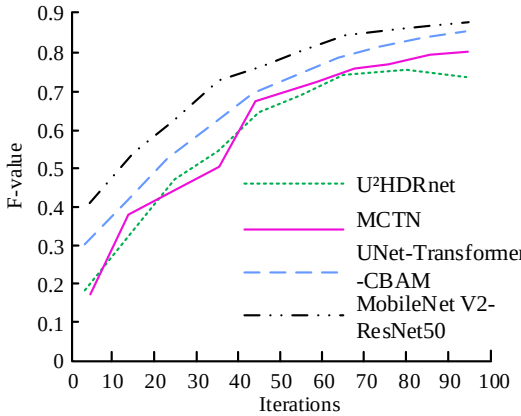
3 Results

3.1 Portrait image cropping effects

To analyze the effectiveness and applicability of the proposed intelligent composition method, this study examines it from three aspects: intelligent cropping for portrait detection, pose recognition analysis, and composition image quality. The MS COCO and MuPoTS-3D datasets are used for pose recognition and intelligent composition analysis of portraits. 30% of the images in the above dataset are randomly selected to form the photographic image composition test dataset. MS COCO is a widely used computer vision dataset with approximately 118,000 training images. It contains rich natural scene images (indoor, outdoor, urban, rural, etc.) and portrait images of individuals (single person, multiple people, different poses, and different lighting conditions). The annotation information includes object detection, instance segmentation, and keypoint detection. The MuPoTS-3D dataset is a dataset for multi-person 3D pose estimation, suitable for pose analysis and composition research in complex scenes. The image type includes scenes of multi person interaction and multiple role instances. The annotation information includes human keypoint annotations, character bounding boxes, and approximately 8,000 training images. The experimental environment is designed as follows. The central processing unit is Intel core i5-6500, equipped with NVIDIA RTX series graphics processors. The deep learning framework is PyTorch, the image processing library is OpenCV, the computer memory size is 12GB, and the optimizer is Adam. The learning rate of the network is set to 0.001, the batch size is 32, and the training epoch is 100. Firstly, the lightweight improved MobileNet V2-ResNet50 network proposed in the study is analyzed for its portrait cropping effect. It is compared with the improved deep learning network (UNet-Transformer-CBAM), lightweight detection network (Center-Net), multi-feature fusion image processing algorithm (U²HDRnet), and Multi-branch convolution-Transformer network (MCTN). The research utilizes Precision, Recall, F1-value, Average Intersection over Union ratio (Avg IoU), Mismatch, Area under Curve (AUC), Precision-Recall Curve (PCK), etc. to conduct experimental result analysis. Precision refers to the proportion of samples predicted by the model to be positive classes that are actually positive classes. The high precision indicates that there are fewer false classification samples with positive prediction. Recall refers to the proportion of samples that are actually positive and correctly predicted as positive by the model. The higher the recall, the more positive samples the model can find. The F1-value is the harmonic mean of precision and recall, used to balance precision and recall. The higher the F1-value, the better the balance between precision and recall achieved by the model. IoU is an indicator used to measure the overlap between predicted boxes and real boxes, commonly used in object detection tasks. The higher the average IoU value, the greater the overlap between the predicted box and the real box. AUC is the area under the ROC curve, which is used to measure the performance of a classification model. The ROC curve is plotted with False Positive Rate (FPR) as the horizontal axis and True Positive Rate (TPR) as the vertical axis. The closer the AUC value is to 1, the better the classification performance of the model. PCK is an accuracy recall curve used to measure the precision and recall of a model at different thresholds. The larger the area under the PCK curve, the better the performance of the model at different thresholds. The area under PCK can be used to evaluate the proportion of correctly estimated keypoints. Figure 6 shows the Precision-Recall (PR) results of extracting portrait image information.



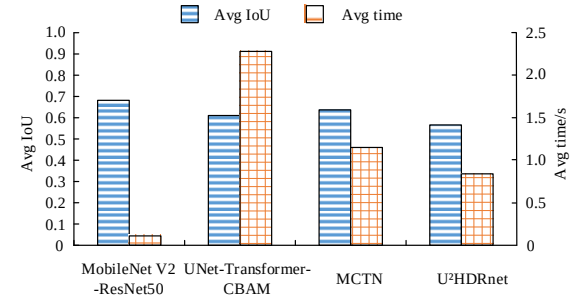
(a) PR value



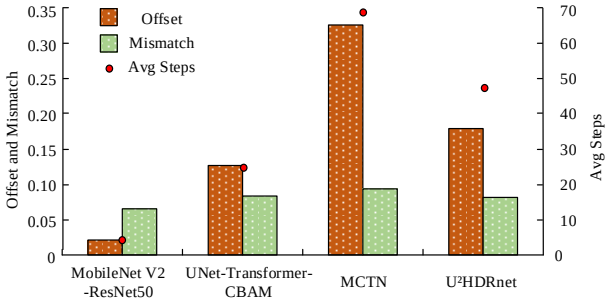
(b) F-value

Figure 6: Comparison of PR curves and F 1 value of different models

In Figure 6 (a), the PR curves of the MobileNet V2-ResNet50 model and the UNet-Transformer-CBAM model were closer to the bottom right corner. The precision of the MCTN and U²HDRnet algorithms was slightly lower, not exceeding 88% and 86.78%, respectively. However, the MobileNet V2-ResNet50 proposed in the study achieved a precision of 94.51% in extracting portrait image information, with good cropping accuracy. In Figure 6 (b), the MobileNet V2-ResNet50 model achieved good portrait image cropping, with small fluctuations in the curve nodes and a smooth direction. Its overall F1-value on the Open Images dataset exceeded 85%. Afterwards, the computational complexity of the above algorithm is analyzed using the Avg IoU, Average Steps (Avg Steps), Average Time/s (Avg time), offset, and Aspect Ratio Error (Mismatch). The results are shown in Figure 7.



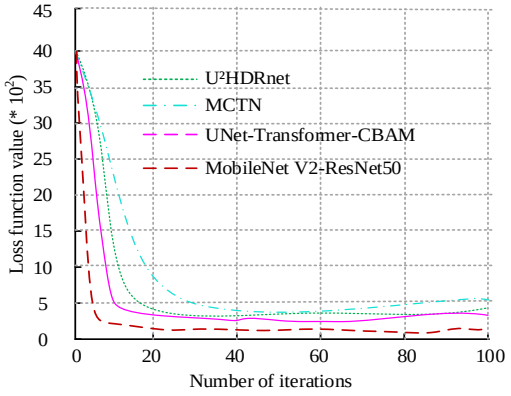
(a) Avg IoU and Avg time



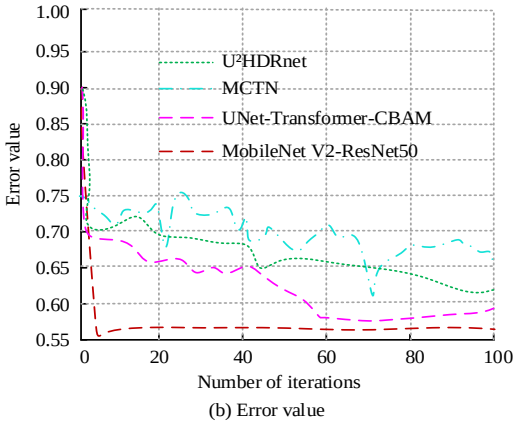
(b) Offset, Mismatch and Avg Steps

Figure 7: Computational complexity of different portrait detection algorithms

In Figure 7 (a), the Avg IoU exhibited by the MobileNet V2-ResNet50 model was 0.69, while the Avg IoU of the other three comparison methods was all less than 0.65. The Avg time of the MobileNet V2-ResNet50 model was significantly lower than other algorithms. Offset refers to the distance between the optimal cropping window and the predicted cropping window. A large value indicates a larger deviation in image processing. In Figure 7 (b), the offsets of MobileNet V2-ResNet50, UNet-Transformer-CBAM, MCTN, and U²HDRnet were 0.021, 0.126, 0.324, and 0.178, respectively, and the Mismatch was 0.065, 0.083, 0.093, and 0.081, respectively. The offset and Mismatch of MCTN were 0.324 and 0.093, respectively, making the calculation more complex and requiring more time and steps. The above results indicate that the proposed portrait detection model has good image cropping performance. Further analysis is conducted on the loss function and extraction error results during the portrait image cropping process, as shown in Figure 8.



(a) Magnitude of the loss



(b) Error value

Figure 8: Facial image feature extraction results of different algorithms

In Figure 8 (a), when the number of iterations was less than 20, the decreasing slopes of the loss curve from large to small were MobileNet V2-ResNet50>KUNet-Transformer-CBAM>U²HDRnet>LBP>MCTN. The loss curve of MobileNet V2-ResNet50 model converged quickly, with an average loss value of less than 5. The average loss values of other algorithms were between 5 and 10. In Figure 8 (b), the classification error curves of the four algorithms showed a decreasing trend with the increase of iterations, with the proposed algorithm having a smaller error value and the MCTN algorithm having the highest loss value. To further analyze the performance of the above models in image processing, ablation experiments are conducted to verify and test the processing speed of different models. The results are shown in Table 2.

Table 2: Ablation results and processing speed of different models

Model	Ablation results			Processing speed		
	Recall/%	Precision/%	F1-value/%	FLOPs/B	Inference time/ms	Parameter measurement/M
UNet-Transformer-CBAM	88.36	87.25	89.99	15	50	25
Center-Net	87.25	87.53	86.98	3	20	8
U²HDRnet	86.14	85.11	86.98	10	35	15
MCTN	86.08	85.93	87.12	18	60	30
MobileNet V2-ResNet50	88.74	89.95	90.14	2.5	5	15

In Table 2, the MobileNet V2-ResNet50 model exhibited recall, precision, and F1-value of 88.74, 89.95, and 90.14, respectively, in the ablation results, which were much higher than other models. The UNet-Transformer-CBAM model, which performed well, had an F1-value of 89.99%. The Floating-Point Operations (FLOPs) of the MobileNet V2-ResNet50 model and the Center-Net model were less than 5B, and the parameter metric of the Center-Net model was the smallest (8M), but its inference speed (20ms) was greater than that of the MobileNet V2-ResNet50 model (5ms). Overall, MobileNet V2-ResNet50 balanced precision, recall, and computational efficiency, making it suitable for real-time portrait photography composition tasks. UNet-Transformer-CBAM and MCTN performe well in complex scenarios, but have high computational costs and are suitable for offline high-precision tasks. Center-Net has high computational efficiency, but slightly lower accuracy in composition tasks, making it suitable for lightweight real-time detection tasks. U²HDRnet performs well in improving image quality, but its performance in composition tasks is not as good as MobileNet V2-ResNet50, making it suitable for post-processing tasks.

3.2 Intelligent portrait composition effect

With the development of mobile devices, photography has gradually become an important tool for people to record their lives. However, the different ideas and intentions of the photographer result in differences in the visual effects presented in the image. The relationships between elements in the images and the overall style of the character all reflect the ingenuity of image construction [21-22]. The application of composition rules and forms can present different aesthetics and artistic qualities in portrait photography. For example, a central composition will attract the viewer's attention to the center of the picture, thereby highlighting the subjectivity of the characters. The three-point composition arranges character themes horizontally or vertically, presenting a more balanced aesthetic style and a more narrative visual effect [23-24]. The portrait composition quality of above models is compared, and the results are shown in Figure 9.

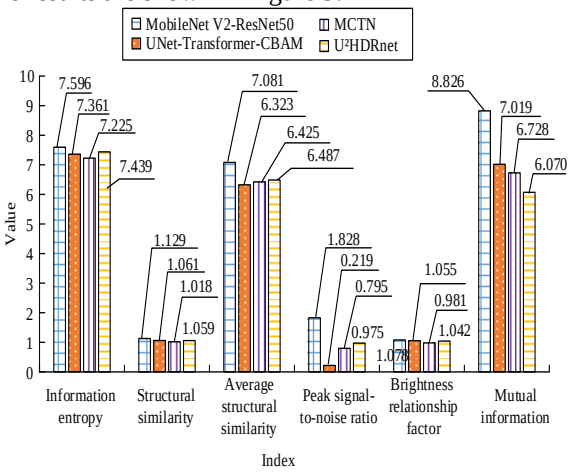


Figure 9: Image quality evaluation effects under different algorithms

In Figure 9, the information entropy, structural similarity, average structural similarity, peak signal-to-noise ratio, brightness relationship factor, and mutual information of the proposed fusion algorithm were 7.596, 1.129, 7.081, 1.828, 1.078, and 8.826, respectively. The overall image quality fusion effect was significantly better than other algorithms. The next was UNet-Transformer-CBAM and U²HDRnet, with information entropy values exceeding 7.35. Afterwards, the proposed Adaptive-Human pose estimation (AHPE) algorithm was tested on the MS COCO and MuPoTS-3D datasets with a convolution size of 3×3. It is compared with Random Sample Consensus-Mean-Shift (RSC-MS) algorithm, Open Pose-Generative adversarial network (OP-GAN), and Progressive Gaussian Filtering-Human pose estimation (PGF-HPE). The recall, precision and F1-value are shown in Table 3.

Table 3: Comparison results of portrait image detection

Dataset	Algorithm	Evaluation		
		Recall/%	Precision/%	F1-value/%
MS COCO	AHPE	88.55	89.94	90.05
	RSC-MS	87.05	87.17	88.41
	OP-GAN	88.07	88.25	87.11
	PGF-HPE	86.93	87.31	88.16
MuPoTS-3D	AHPE	89.78	91.36	90.20
	RSC-MS	86.87	87.53	86.98
	OP-GAN	86.25	87.73	87.12
	PGF-HPE	88.43	88.94	89.56

In Table 3, on the MS COCO dataset, AHPE achieved a precision of 88.94% for human pose detection, while the detection precision of RSC-MS, OP-GAN, and PGF-HPE was 87.17%, 88.25%, and 87.31%, slightly worse than the algorithm proposed in the study, and their corresponding F1-values were basically no more than 90%. On the MuPoTS-3D dataset, the performance of human pose detection from high to low was as follows: AHPE algorithm>PGF-HPE algorithm>OP-GAN algorithm>RSC-MS algorithm. The precision, recall, and F1-value of the proposed adaptive human pose recognition algorithm reached 91.36%, 89.78%, and 90.20%, respectively. Afterwards, further analysis is conducted on the key point prediction results of human nodes using different algorithms, and classification evaluation is performed using AUC and PCK metrics. The higher the value, the higher the prediction accuracy of the model. The predicted key points of human body nodes are shown in Table 4.

Table 4: Prediction scores of key points in human body nodes

Dataset	Algorithm	Evaluation	
		AUC	PCK
MS COCO	AHPE	44.99	83.24
	RSC-MS	35.94	76.19
	OP-GAN	42.26	82.51
	PGF-HPE	43.22	82.87
MuPoTS-3D	AHPE	47.87	87.12
	RSC-MS	43.86	84.11
	OP-GAN	46.09	86.34
	PGF-HPE	45.52	83.77

In Table 4, the AUC and PCK of AHPE algorithm on the MS COCO dataset were 44.99 and 83.24, respectively, which were much higher than those of RSC-MS algorithm (35.94 and 76.19), OP-GAN algorithm (42.26 and 82.51), PGF-HPE algorithm (43.22 and 82.87), and the maximum difference in AUC and PCK scores exceeded 5. On the MuPoTS-3D dataset, AHPE algorithm and OP-GAN algorithm could better capture key nodes of the human body, and the score deviation was relatively small, with RSC-MS algorithm performing the worst. Specifically, the AUC and PCK of AHPE were 47.87 and 87.12, respectively, while the AUC and PCK of RSC-MS algorithm were 43.86 and 84.11, respectively. The reason for this result is that the AHPE algorithm has good adaptive performance and can effectively capture pose key nodes. However, the RSC-MS algorithm ignores the relationship between the central pose and key points during pose analysis, which makes it more prone to erroneous results. The MobileNet V2-ResNet50 model is combined with the AHPE model to obtain a hybrid intelligent composition method. Two types of portrait images are selected for intelligent composition analysis, and the results are shown in Figure 10.

(a) Original image (b) Optimized composition of the image

Figure 10: Intelligent portrait composition presentation effects

In Figure 10, when there were characters and scenery present, this method utilized intelligent cropping to better consider the relationship between characters and background, reducing unnecessary scene whitespace. When it was a multi-person image, this method highlighted the story expression of the characters and had a better image presentation effect than the original image.

4 Discussion and conclusion

A portrait photography intelligent composition system based on multi-task learning and human pose analysis was designed, which could automatically optimize portrait composition. The results showed that the PR curves of MobileNet V2-ResNet50 model and UNet-Transformer-CBAM model were closer to the bottom right corner, while the precision of MCTN and U²HDRnet algorithms was slightly lower, not exceeding 88% and 86.78%, respectively. When the training data is insufficient or diverse, MCTN model may be difficult to fully learn effective feature representation, introducing high computational complexity and over-fitting risk. The multi-branch structure may lead to feature redundancy, reducing the model's ability to focus on key features. The design goals of U²HDRnet are not completely consistent with the accuracy requirements of composition tasks, resulting in poor performance in composition tasks. The proposed method achieves an accuracy of 94.51% in portrait detection, outperforming existing approaches. The multi-task learning framework enhances image quality while adapting to various human poses. The Avg IoU exhibited by the MobileNet V2-ResNet50 model was 0.69, while the Avg IoU of the other three comparison methods was all less than 0.65. The Avg Steps and Avg time of the MobileNet V2-ResNet50 model were significantly lower than other algorithms. The offsets of MobileNet V2-ResNet50, UNet-Transformer-CBAM, MCTN, and U²HDRnet were 0.021, 0.126, 0.324, and 0.178, respectively, and the Mismatch was 0.065, 0.083, 0.093, and 0.081, respectively. The loss curve of MobileNet V2-ResNet50 model converged quickly, with an average loss value of less than 5, while the average loss values of other algorithms were between 5 and 10. The improved lightweight MobileNet V2-ResNet50 network significantly outperformed the UNet-Transformer-CBAM and MCTN models in terms of computational efficiency, with FLOPs of only 2.5B, parameter count of 5M, and inference time of 15ms, making it suitable for real-time applications. The UNet-Transformer-CBAM model and MCTN model had high computational costs and were difficult to deploy on resource constrained devices. Compared with the lightweight network proposed by Li et al. [5], the research method further optimized computational efficiency while preserving fine-grained details, making it suitable for portrait detail processing in complex backgrounds. Human pose estimation is an important part of pose analysis, which can abstract important information representing human pose from input images. The adaptive human pose recognition algorithm AHPE proposed in the study achieved a precision of 88.94% for human pose detection on the dataset, while the detection accuracy of the RSC-MS, OP-GAN, and PGF-HPE comparison models did not exceed 88.5%. The prediction results of key points in human body nodes showed that the AHPE algorithm had AUC and PCK scores of 44.99 and 83.24 on the MS COCO dataset, which were much higher than the RSC-MS algorithm's 35.94 and 76.19, the OP-GAN algorithm's 42.26 and 82.51, and the PGF-HPE algorithm's 43.22 and 82.87, respectively. The maximum difference in AUC and PCK scores exceeded 5 points. The MobileNet V2-ResNet50 model proposed in the study was combined with the AHPE model to obtain a hybrid intelligent composition method. Compared with the photography composition assistance framework proposed by Farhat et al. [4], the research method was more efficient in object detection and composition analysis, and did not require complex multitasking frameworks. Compared with the field of view evaluation model proposed by Zhong et al. [6], the research method did not rely on image boundaries as cropping boundaries, which better enhanced the aesthetic appeal of images. Two types of portrait images were selected for intelligent composition analysis. The results showed that the whitespace in the scene content was reduced, highlighting the relationship between characters and the background or the sense of storytelling of objects.

The research method combines the lightweight of MobileNet V2 and the high-precision characteristics of ResNet50, which can simultaneously adapt to multi-task requirements such as object detection, image classification, and composition analysis. This research method has good intelligent composition effect and can effectively improve the intelligent cropping effect of portrait photography. Compared with the attribute assisted multi-modal memory network proposed by Li et al. [8], the research method performs more efficiently in image aesthetic evaluation and does not require complex multi-modal information fusion. It solves the detail loss problem of lightweight networks proposed by Li et al. [5] in complex backgrounds. This research method simplifies the multi-task processing framework by improving the composition rules and visual field evaluation, and solves the problem that the existing cutting methods proposed by Zhong et al. [6] are difficult to improve the aesthetic feeling. However, the performance of the research method in cluttered backgrounds and their dependence on hardware resources still need further improvement. Stronger attention mechanism can be introduced to enhance the target detection and synthesis ability of the model in the clutter background. Combined with multi-scale feature fusion technology, the ability of the model to deal with complex texture and dense targets is enhanced. While the proposed approach improves composition accuracy, its computational cost remains a consideration for real-time applications. Future optimizations could involve model pruning or quantization to enhance efficiency for mobile and embedded devices. Meanwhile, considering that the portrait in the actual photography process is also affected by factors such as lighting, occlusion, and complex

background, there may be a certain degree of error in pose recognition. Therefore, in the future, it is possible to strengthen the behavior recognition scheme and pose estimation method for skeleton sequences, while considering user interaction and personalized settings, and attempt to provide personalized intelligent composition functions by introducing user feedback mechanisms.

References

[1] Jiang N, Sheng B, Li P, Lee T Y. Photohelper: portrait photographing guidance via deep feature retrieval and fusion. IEEE Transactions on Multimedia, 2022, 25: 2226-2238. DOI: 10.1109/TMM.2022.3144890.

[2] Bao W. The application of intelligent algorithms in the animation design of 3D graphics engines. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 2021, 13(2): 1-12. DOI: 10.4018/IJGCMS.2021040103.

[3] Ather D, Madan S, Nayak M, Tripathi R, Kant R, Kshatri S S, Jain R. Selection of smart manure composition for smart farming using artificial intelligence technique. Journal of Food Quality, 2022, 1: 4351825. DOI: 10.1155/2022/4351825.

[4] Farhat F, Kamani M M, Wang J Z. CAPTAIN: Comprehensive composition assistance for photo taking. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(1): 1-24. DOI: 10.1145/3462762.

[5] Li R, Zhang D, Geng SL, Zhou MQ. Matting Algorithm with Improved Portrait Details for Images with Complex Backgrounds. Applied Sciences. 2024,27;14(5):1942. DOI: 10.3390/app14051942.

[6] Zhong L, Li F H, Huang H Z, Zhang Y, Lu S P, Wang J. Aesthetic-guided outward image cropping. ACM Transactions on Graphics (TOG), 2021, 40(6): 1-13. DOI: 10.1145/3478513.3480566.

[7] Wang B, Si H, Fu H, Gao R, Zhan M, Jiang H, Wang A. Content-Aware Image Resizing Technology Based on Composition Detection and Composition Rules. Electronics, 2023, 12(14): 3096. DOI: 10.3390/electronics12143096.

[8] Li L, Zhu T, Chen P, Yang Y, Li Y, Lin W. Image aesthetics assessment with attribute-assisted multimodal memory network. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 7413-7424. DOI: 10.1109/TCSVT.2023.3272984.

[9] Baraheem S S, Le T N, Nguyen T V. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. Artificial Intelligence Review, 2023, 56(10): 10813-10865. DOI: 10.1007/s10462-023-10434-2.

[10] Wang Q, Ma Y, Zhao K, Tian Y. A comprehensive survey of loss functions in machine learning. Annals of Data Science, 2022, 9(2), 187-212. DOI: 10.1007/s40745-020-00253-5.

[11] Zhang K, Ren W, Luo W, Lai W S, Stenger B, Yang M H, Li H. Deep image deblurring: A survey. International Journal of Computer Vision, 2022, 130(9), 2103-2130. DOI: 10.1007/s11263-022-01633-5.

[12] Ghasemi-Naraghi Z, Nickabadi A, Safabakhsh R. LogSE: An Uncertainty-Based Multi-Task Loss Function for Learning Two Regression Tasks. Journal of Universal Computer Science, 2022, 28(2):141-159. DOI: 10.3897/jucs.70549.

[13] Zhang H, Peng Y. Image clustering: An unsupervised approach to categorize visual data in social science research. Sociological Methods & Research, 2024, 53(3): 1534-1587. DOI: 10.1177/00491241221082603.

[14] Chikarkova M. Artificial intelligence and digital art: current state and development prospects. Skhid. 2023,4(3):9-13. DOI: 10.21847/2411-3093.2023.4(3).294658.

[15] Chen R, Ghavidel Aghdam MR, Khishe M. Utilization of Artificial Intelligence for the automated recognition of fine arts. PloS one. 2024,19(11):0312739. DOI: 10.1371/journal.pone.0312739.

[16] Gross EC. Artificial Intelligence Generated Art Imitation and the Art World: Implications and Further Questions. Boletín de Arte. 2023,28(44):313-316. DOI: 10.24310/ba.44.2023.15972.

[17] Xu J, Liu W, Xing W, and Wei X. MSPENet: multi-scale adaptive fusion and position enhancement network for human pose estimation. The Visual Computer, 2022, 39(5):2005-2019. DOI: 10.1007/s00371-022-02460-y.

[18] Kumar P, Chauhan S, Awasthi L. Human pose estimation using deep learning: review, methodologies, progress and future research directions. International Journal of Multimedia Information Retrieval, 2022, 11(4):489-521. DOI: 10.1007/s13735-022-00261-6.

[19] Huang L, Liu G. Functional motion detection based on artificial intelligence. The Journal of Supercomputing, 2022, 78(3): 4290-4329. DOI: 10.1007/s11227-021-04037-3.

[20] Wang X, Wu Y, Zhu L, and Yang Y. Symbiotic Attention with Privileged Information for Egocentric Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12249-12256. DOI: 10.1609/aaai.v34i07.6907.

[21] Zhou K, Wu T, Wang C, Wang J, and Li C. Skeleton Based Abnormal Behavior Recognition Using Spatio-Temporal Convolution and Attention-Based LSTM. Procedia Computer Science, 2020, 174(8):424-432. DOI: 10.1016/j.procs.2020.06.110.

[22] Zhang H, Luo Y, Zhang L, Wu Y, Wang M, Shen Z. Considering three elements of aesthetics: Multi-task self-supervised feature learning for image style classification. Neurocomputing, 2023, 520(1): 262-273. DOI: 10.1016/j.neucom.2022.10.076.

[23] Yang G Y, Zhou W Y, Cai Y, Zhang S H, Zhang F L. Focusing on your subject: Deep subject-aware

image composition recommendation networks. Computational Visual Media, 2023, 9(1): 87-107. DOI: 10.1007/s41095-021-0263-3.

[24] Zhang M, Li M, Yu J, Chen L. Aesthetic photo collage with deep reinforcement learning. IEEE Transactions on Multimedia, 2022, 25(1): 4653-4664. DOI: 10.1109/TMM.2022.3180217.