Evaluation of Image Colourisation Algorithms

HAREEHARAN ELANGOVAN

24/10/2022

School of Mathematics,
Cardiff University

A dissertation submitted in partial fulfilment of the
requirements for MSc data science and analytics
by taught programme, supervised by Paul Rosin

Executive summary

Humans have always seen the world in colour. In the last few decades, there has been a rapid and enormous transition from black-and-white images to colour ones. Image colourisation aims to produce a natural-looking colour image from a given grey-scale image, which remains a challenging problem. As the colourisation of the black & white images evolves, the metrics to evaluate the quality of the colourised images have to be evolved one step ahead.

Well-known objective evaluation metrics for measuring image quality metrics will measure the error between original and processed images. One of these objective measures' standard and significant limitations is that they require reference images to evaluate the quality of grayscale images by fidelity to the original image. This paper proposes a novel approach based on the clear combination of mathematical formulae and deep learning to evaluate and provide an overall score for the colourised image by getting the object's colour distribution. That will be learned and transformed into a database to act as the reference data, which has plausible colours for each label. The colourised image's evaluation follows the same process of segmenting objects in the picture. The colour distribution of objects was extracted to check the plausibility of each pixel colour by referring to the colour distribution of the same object in the generated train data. The colourfulness and colour cast are used to measure and evaluate the unsegmented pixels to produce the overall score for the image by considering both segmented and unsegmented pixels.

Acknowledgement

Thank you to my friends and family for supporting me throughout my degree and I would like to express my deepest gratitude to my supervisor professor Paul rosin for guiding me throughout my dissertation in a direction so that my experience on this project has way more helped me to acquire in-depth knowledge about deep neural networks and statistical modeling.

Table of Contents

List of Figures

List of tables

Acronyms

FR-IQA	**-**	Full reference image quality assessment
NR-IQA	-	No reference image quality assessment
RR-IQA	-	reduced reference image quality assessment
MSE	-	Mean Squared Error
PSNR	-	Peak Signal-To-Noise Ratio
SSIM	-	Structural Similarity Index
UIQM	**-**	Underwater Image Quality Measure
QSSIM	-	Quaternion Structural Similarity Index
RGB	-	Red Green Blue
HSV	-	Hue Saturation Value
CIELAB	**-**	Commission Internationale d'Eclairage
CNN	**-**	Convolutional neutral network
CSF	**-**	Contrast sensitivity function
HVS	**-**	Human visual system
COCO	-	Common Objects in Context

1. Introduction

Before we go into the technical aspects, images are influential in everyone's life since they link to our history and remind of people, sites, memories and sentiments. Colour in a picture, in particular, may effectively convey stories visually and can be used to express them emotionally. Furthermore, historical black-and-white photographs are recognized as precious works of art with extraordinary artistic merit. Looking at them makes it hard to visualise the real scenario properly. As a result, digital colourization evolved in the 1970s, adding colour to digital grayscale photos. Enhancing the picture more convincing, bridged the spatial gap seen between the present and the past. The ultimate objective of colourization is to create full-colour pictures that look convincing to a human viewer, yet human assessment is expensive and time-consuming. Automatic colourization algorithms began to emerge. When the colourization algorithm was developed gradually, the metrics used to evaluate the quality of colourized photos began to play an essential part in enhancing the algorithm's ability to produce realistic colours.

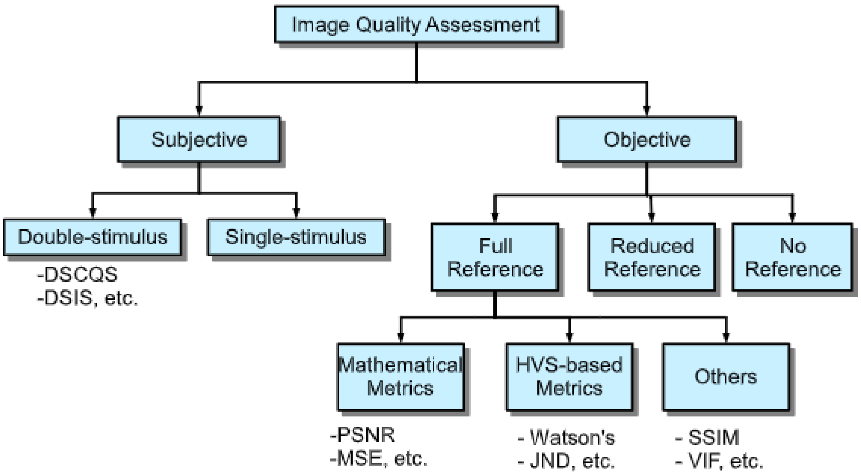


Fig. 1 Image Quality Assessment(Thung and Raveendran 2010)

Image quality may be measured in two ways: subjectively and objectively (Thung and Raveendran 2010). Quality is rated subjectively by a group of human spectators, which is usually uncomfortable, time-consuming, and costly. On the other hand, objective measures are quantifiable characteristics that can automatically predict perceived visual quality. Objective image quality metrics are categorised as follows: FR-IQA requires the availability of a whole reference image. NR-IQA indicates that the reference image is not available. The reference image in the third category is only partially accessible. It takes the form of a set of extracted features known as evaluating the RR-IQA. Many popular objective image quality assessment techniques exist, including MSE, PSNR, SSIM, UIQM, and QSSIM, which fall under FR. The NR-IQA issue is much more complicated than the previous two FR-IQA and RR-IQA methods.

1.1 Problem statement

One of the standards and significant limitations of these objective measures is that they only evaluate the quality of grayscale images and predict the perceived difference between a distorted image and a Ground truth image. So it is difficult to measure the quality of colourized grey scale images taken back in the 18th century since the Ground truth image, which will act as a reference image, will not be available; hence the vast majority of objective assessments are so-called image-difference metrics(Preiss 2015). Also, Since there is no comparison image, the statistics of the reference image, the nature of the HVS, and the impact of distortions on image features must be addressed unsupervised. Evaluating the efficiency of a quality measure in the absence of a reference picture with a particular distorted image is similarly problematic. (Kamble and Bhurchandi 2015) Provides a thorough analysis of the various NR-IQA algorithms developed so far. Image quality evaluation is difficult and remains an active study subject because of a lack of knowledge of the HVS.

1.2. Thesis goal

The thesis's objective is to develop an evaluation metric to evaluate the quality of an image with no reference image using deep learning, computer vision, and a mathematical formula by analysing and extracting the plausible colour of various objects from real-world photos using segmentation. As a result, each label will have its colour distribution. Finally, utilise those colour distributions to assess the colour plausibility of objects in the colourized test picture to produce an overall score.

1.3. Outcome

The research outcome will be to develop a no-reference IQA algorithm and determine whether the algorithm-generated score coincides with the human visual system score on the same set of images.

2. Background

2.1. Convolutional neural networks

Convolutional neural networks, a subset of multilayer neural networks and deep learning techniques, are now widely used in a variety of computer vision applications. A CNN is made up of many key components, including convolution layers, pooling layers, including fully connected layers, and is intended to automatically and interactively learn spatial hierarchies of features by going from low-level to high-level patterns using a backpropagation technique.

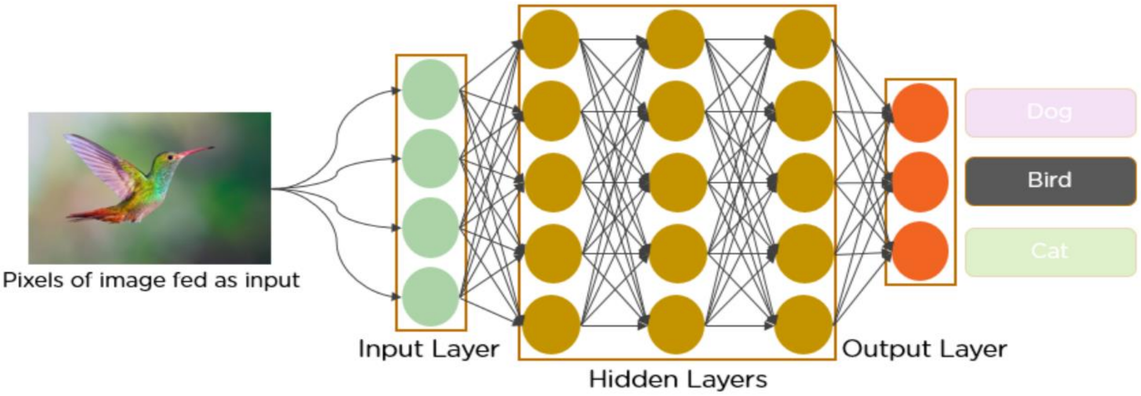


Fig. 2 Architecture of CNN

(Mandal 2021) CNN is made up of several layers of artificial neurons. Artificial neurons are mathematical functions that compute a weighted sum of many inputs and produce an activation value whenever an image is fed into a ConvNet; once an image is an input into a ConvNet, each layer creates various activation functions passed into the next layer.

2.2. Colour space

Colour terms like white, red, yellow, and brown, as well as adjectives like brilliant, dark, dull, and saturated, are used naturally to describe colours(Berns and Reiman 2002)**.** Every item in the colour vector corresponds to a colour space coordinate representing the colour as a spot in the colour space (Reinhard et al. 2008). There are several colour models. Some of them are RGB, CMYK, YIQ, HSV, HLS, Etc.

2.2.1. RGB

The RGB colour model consists of three colours: red, green, and blue, which are combined to generate a particular colour and are individually coded on a 256-point scale ranging from 0 to 255, all three colours merge to generate white, represented as RGB(255, 255, 255), and black at zero intensity (0, 0, 0). Although the RGB model is a primary method of describing colours, it does not correspond to how the human eye sees colours.

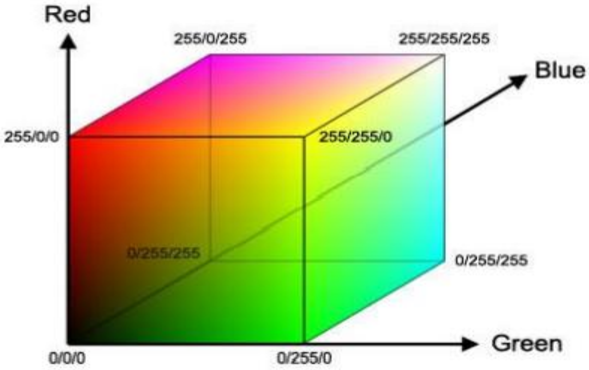


Fig. 2 RGB colour model(H and K 2018)

2.2.2. HSV

HSV (hue, saturation, value) colour models were developed to replicate how people see and interpret colour. The hue is what distinguishes a colour. The hue axis values vary from 0 to 360, commencing and terminating with red and going via green, blue, & other intermediate colours. Saturation indicates how much the colour deviates from grey. The values range from 0 to 1, representing the maximum saturation of a particular hue under specified lighting conditions. The level of lighting is represented by the value (in HSV) ranging from 0 (no light, black) at the bottom to 1 (white) at the top, with total saturation reached at V=1 in HSV.

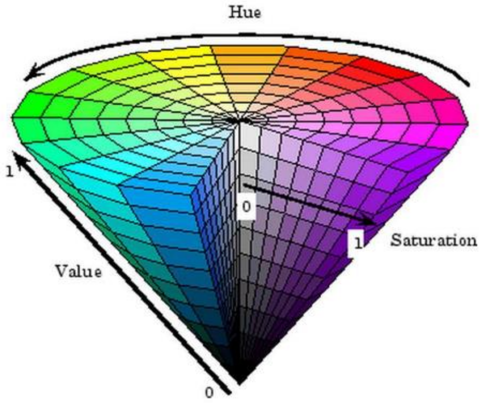


Fig. 3 HSV colour space (Erdoğan and Yılmaz 2014)

2.2.3. CIELAB

(Ly et al. 2020)The CIE in CIELAB is an abbreviation of “Commission Internationale de l'Eclairage for the French name of the International Commission on Illumination”, The 3D colour space CIELAB, sometimes referred to as CIE L*a*b*, is independent of devices and allows for accurate measurement and comparison of all viewable colours using three different colour values.

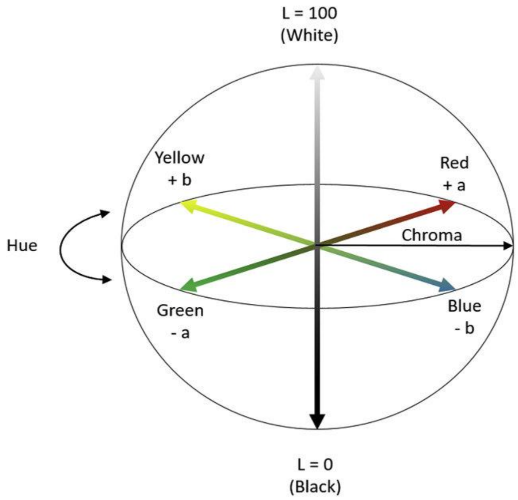


Fig. 4 CIELAB colour space diagram(Ly et al. 2020)

The CIELAB colour space uses each of the three variables to measure the correct colour and compute colour differences represented by the letters L*, a*, and b*. On a scale of 0 to 100, L* represents brightness, while a* and b* indicate chromaticity with no numerical limitations. Negative a* stands for green, positive a* stands for red, negative b* stands for blue, and positive b* stands for yellow. The centre of this plane is neutral or grey.

2.3. Colour cast

A colour cast is just a tint of a specific colour that is typically unpleasant and affects the whole or a portion of a photographic picture. Colour cast prevents the image from recovering the subject's actual colour, decreasing its aesthetic and fundamental element. The amount of colour cast is further characterised as reddish, blue, greenish, and yellowish.

2.4. Colorfulness

Colourfulness Rather than achieving brightness and colour fidelity, new pictorial imaging systems seek the best-looking picture. So, colourfulness is assessing and measuring the image's pixel colours to produce a score based on the image's colourfulness.

2.5. Metrics to evaluate Colourization

2.5.1. RMSE

(Teng et al. 2021) The RMSE between the produced colour image and ground truth picture. The bigger the RMSE number, the greater the disparity between colourized pictures produced by the colourization algorithm and the ground truth picture. To calculate RMSE the RGB values are extracted from the image. Equations (1) and (2) define RMSE, where H denotes the image's height, W denote the image's width in pixel value, and where x can be considered as ground truth, and y can be represented as a generated colour image, respectively. The RGB values extracted from the image were used to calculate.

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (x - y)^2 \qquad (1)$$

$$RMSE = \sqrt{MSE} \qquad (2)$$

2.5.2. PSNR

(Teng et al. 2021) Peak Signal-To-Noise Ratio is a popular image assessment metric that measures the difference among corresponding pixels. Greater the Peak Signal-To-Noise Ratio number, the lesser the difference, also vice versa. Equation (3) defines the computation of metric using equation(1) MSE is derived, where n denotes the image bit depth.

$$PSNR = 10\ log_{10}\left(\frac{(2^n - 1)^2}{MSE}\right) \qquad (3)$$

2.5.3. SSIM

(Teng et al. 2021) SSIM focuses on image structure similarity and assesses image similarity based on factors such as structure (s), contrast (c), and luminance (l). A measure of this kind analyses structural similarity and detail consistency among a reference, for which value ranges from 0 to 1 for colour image generated by the model. Regarding PSNR, a more significant SSIM number denotes better picture quality and conversely. The SSIM equations are presented below, where Y represents the generated colour image, and X is the ground truth picture.

μx - mean of X, μy - mean of Y

σ_x^2 - variance of X σ_y^2 - variance of Y

σxy - covariance of X and Y

c_1, c_2, c_3 - constants

$$SSIM(X,Y) = l(X,Y)^{\alpha} \times c(X,Y)^{\beta} \times s(X,Y)^{\gamma} \qquad (4)$$

where

$$l(X,Y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2+\mu_y^2+c_1}$$

$$c(X,Y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2+\sigma_y^2+c_2} \qquad (5)$$

$$s(X,Y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y+c_1}$$

In general to avoid the denominator to tend to zero where γ, β, and α are set to 1
c1 = (K1 * L)2, c2 = (K2 * L)2, C3 = 0.5 * C2, K1 = 0.01,
L = 255.19 and K2 = 0.03

Equation (4) can be rewritten as,

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \qquad (6)$$

3. Literature review

(Žeger et al. 2021) This research includes an overview and review of grayscale picture colourisation techniques and categorising the currently available colourisation methods. This paper uses several measures for evaluating the quality of coloured images. A detailed analysis of current colourisation procedures and a quantitative assessment of the results were carried out. The author used five different algorithms have been used for evaluation and analysis, and Colourization has been done on five distinctive photographs made by the author. In this paper, five different images were used with notable changes in colour. For image quality evaluation, PSNR, SSIM, Etc were used. Also, methods that try to measure the image's visual quality, and not only fidelity to the original, have been used. (Hasler and Suesstrunk 2003), UIQM (Panetta et al. 2016), and UCIQE (Yang and Sowmya 2015) were used to measure colourfulness. Various user-guided approaches, such as scribble-based systems, in which technicians manually choose suitable colours for each item represented in an image. Although intricate and time-consuming, each section was allocated a proper colour, much as in a colouring book, and was also applied. Example-based methods process works by transferring colour information from a coloured source picture to the corresponding portions of a grayscale destination image. This approach decreases but does not eliminate human intervention in the colourization process. Simple feedforward CNNs that predict the distribution of potential colours for each pixel. User-guided colourisation neural networks need user engagement such as scribbles, sketches, points, strikes, or textual phrases at the neural network input in the delayed distribution or real-time. Diverse colourisation neural networks use a two-part generator that generates colour information while classifying semantic material. The discriminator learns to distinguish between true and false input. The colourisation technique uses the CIELAB colour space. The primary reason for categorising a technique as a multi-path colourisation neural network is that it learns characteristics from multiple routes. Exemplar-based colourization neural networks are a subset of the field of example-based techniques in which one or more reference images transmit the colour to the target image. The user-guided colourization neural network gives good results since it combines both the user and neural network.

(Thung and Raveendran 2010)Historically, picture quality has been defined by the appearance of visual distortions such as blockiness, Gaussian noise, blurriness, and colour shifts. As a result, measuring the visibility of these distortions is the most popular technique for modelling an image quality metric. The error signal is normalised according to its visibility, as assessed by human perception psychophysics. Some of the HVS aspects that are often employed in IQM include the contrasting sensitivity function (CSF), which states that human perception is much more sensitive towards lower spatial frequencies than higher ones. Human eyes respond to brightness contrast something beyond the absolute luminance value. Contrast masking is just the process of reducing the visibility of one visual component due to the presence of another. The structural similarity approach assumes that HVS is well adaptable to highly organised natural scene information. The observed visual distortion should thus be roughly approximated by a measure of structural information change. An overview of the most recent, cutting-edge image quality evaluation research is given in this paper. The subjective IQA approach and the publicly accessible image database are discussed briefly. The full-reference IQM is the centre of this discussion on objective quality metrics. HVS-based metrics have a greater connection with human perception but have significant drawbacks, such as design complexity and difficulty determining visibility thresholds. Contrarily, the IQA has been represented by structural similarity as a change in structural features in the picture. It has a minimal computational complexity while correlating well with subjective evaluation.

4. Methodology

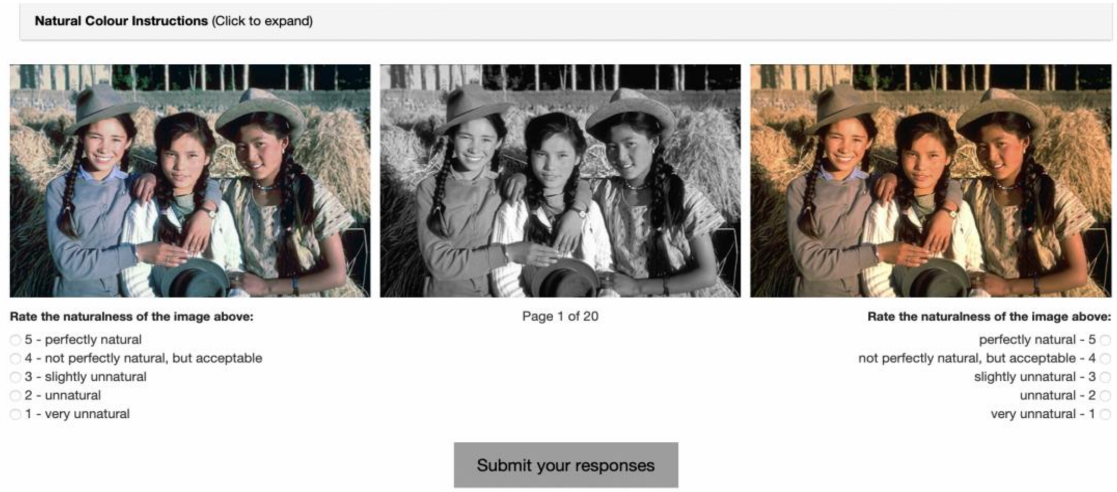
4.1. Dataset

4.1.1. Train dataset

To generate the colour distribution for each item, the coco val2017 dataset is utilised as the training dataset. The COCO dataset, which stands for Common Objects in Context, is intended to represent a wide variety of 80 classes that we encounter in everyday life.

4.1.2. Test dataset

(Mullery & Whelan, 2022) The Human Evaluated Colourisation Dataset, derived from 20 pictures from the Berkeley Segmentation Dataset, was utilised for testing. The original picture will be called the ground truth. The Ground truth image is then used to produce the 65 distinct images for each of the 20 photos in Photoshop [Adobe, 2021] by keeping the L*channel intact and making modifications to a*b* to make them perceptually consistent.



Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

○ 5 - perfectly natural
○ 4 - not perfectly natural, but acceptable
○ 3 - slightly unnatural
○ 2 - unnatural
○ 1 - very unnatural

Page 1 of 20

Rate the naturalness of the image above:

perfectly natural - 5 ○
not perfectly natural, but acceptable - 4 ○
slightly unnatural - 3 ○
unnatural - 2 ○
very unnatural - 1 ○

Fig. 5 Front window of the user(Mullery & Whelan, 2022)

The user is shown a window with one picture as the ground truth, which may be on either side and one as the decolourization image, which is pseudo-random, so that the user does not develop acclimated to the sort of colour shift. Because there are 65 recolor variations, 65 surveys of 20 comparisons are made, totallingg 1300 replies.

4.2 Detectron2

Detectron2 is a Facebook AI Research (FAIR) framework for enhanced performance and scalability in object detection and image segmentation. Detectron2 was created using the open-source PyTorch deep learning framework. Mask R-CNN, RetinaNet, Faster R-CNN, TensorMask, DensePose, and other object detection algorithms are supported. There are three primary forms of segmentation supported:

(1) panoptic segmentation,
(2) instance segmentation, and
(3) semantic segmentation.

Common Objects in Context (COCO) format

(Yagüe et al. 2022) To make Detectron2 highly efficient, a JSON-based image file system was utilised, which enables objects of interest to be labelled. This document is divided into three sections: (1) Images, (2) Categories, and (3) Annotations. All this information was derived through binary masks or ground truth pictures, which depicted the actual image in black and the indicated faults in white.

4.2.1. Instance segmentation

i. Mask R-CNN

The Mask R-CNN (He et al. 2017), created by FAIR, is considered one of the most effective instance segmentation frameworks. CNN such as Mask R-CNN image segmentation of state-of-the-art. Faster R-CNN was used to build Mask RCNN. Image segmentation must be explained for a complete comprehension of Mask R-CNN. Image segmentation is the technique of finding objects and their boundaries. Faster R-CNN receives the object mask from a third element added by Mask R-CNN, which provides two results for a class label as well as a bounding box for each object. The supplementary mask output varies from the box and class outputs, demanding the retrieval of a more complete spatial configuration of an item that contains, in addition to the present branch for bounding box detection, a branch of predicting a mask of the object (Region of Interest).

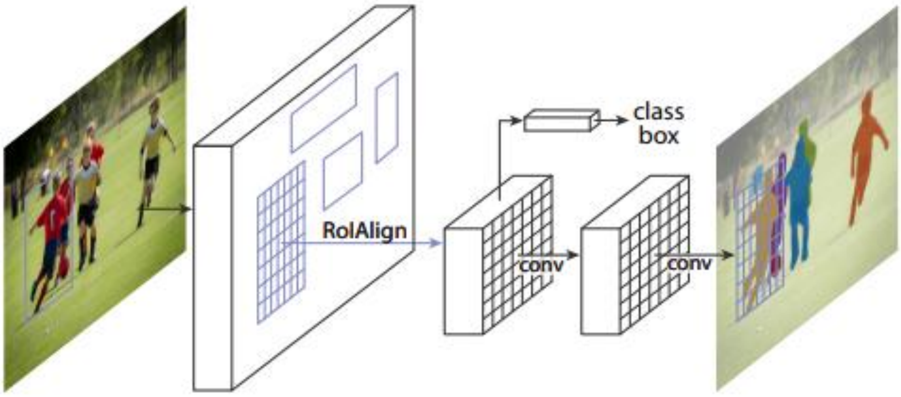


Fig. 6 Instance segmentation R-CNN framework(He et al. 2017)

Semantic segmentation and object detection are integrated, an evolution of the RCNN (Girshick et al. 2014), Fast RCNN (Girshick 2015), and Faster RCNN (Ren et al. 2017) methods. This framework works in two stages: (a) generating region proposals and (b) categorising each created proposal.

Many convolutional backbone architectures are utilized for feature extraction throughout an entire image. A bottom-up and top-down pathway is used in the Mask-RCNN design. The bottom-up component is in charge of convolutions and feature map development.

We also analyse a Feature Pyramid Network(FPN), a more effective backbone previously introduced by (Lin et al. 2017). Using a ResNet-FPN backbone for feature extraction with Mask RCNN yields high accuracy and performance benefits.

ii. Region Proposal Network (RPN) & Region of Interest (ROI) Align

(Huang et al. 2019) Feature maps are the result of the backbone's first step, which is utilised in the RPN. The R-CNN stage gathers features from each proposal by RoIAlign and then conducts proposal classification, mask prediction, and bounding box regression.

4.3 Training dataset generation

4.3.1 Architecture

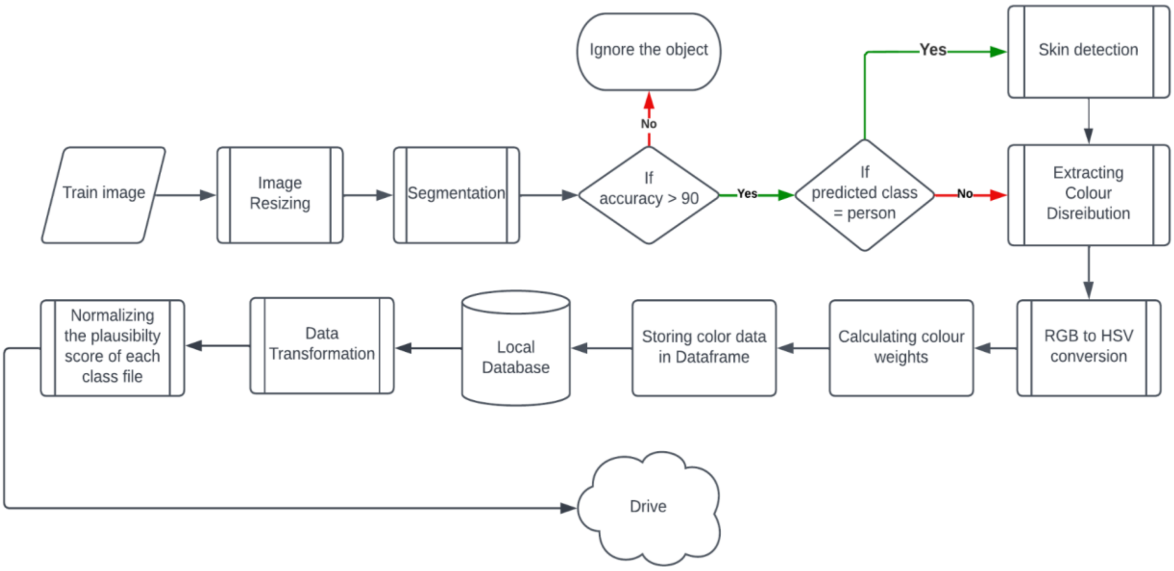

Ignore the object
Yes
Skin detection
No
Train image
Image Resizing
Segmentation
If accuracy > 90
Yes
If predicted class = person
No
Extracting Colour Disreibution
Normalizing the plausibilty score of each class file
Data Transformation
Local Database
Storing color data in Dataframe
Calculating colour weights
RGB to HSV conversion
Drive

Fig. 7 Architecture of generating train data

a. Resizing the image

Images from the COCO dataset and actual images from real-world scenarios provide training data. Consequently, each image may vary in height and width, and most images were too large, requiring more processing time and power to analyze the image, prolonging the training process. So the images have been scaled by modifying their dimensions.

b. Segmentation

The backbone network is the structure used to extract features in deep learning networks for object detection tasks. Detectron2 is used for image segmentation throughout this research, with pre-trained ResNet101 (ResNet with 101 layers) serving as the backbone. However, two popular models for Faster R-CNN are X101-FPN & R101-FPN (Lee and Park 2020). Though X101- FPN exceeds ImageNet in terms of box AP, it consumes more time to train as well as to predict and maybe sometimes overfitting. Furthermore, the R101-FPN baseline results are comparable to the ensemble result of the second-place winner(Dai et al. 2016)

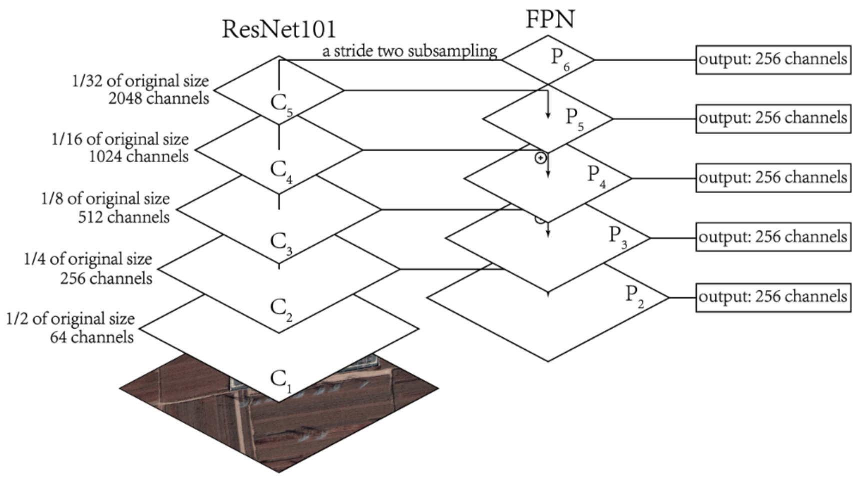


Fig. 8 Structure of ResNet101 FPN(Qiao et al. 2020)

(Qiao et al. 2020) The construction of ResNet101, which follows a bottom-up approach, can be seen on the left. ResNet is a framework for residual learning which uses shortcut connections to deepen neural networks. ResNet generates feature maps in five phases, C1-C5, with varying feature sizes. On the left side, the sizes and channels of C1-C5 have represented. FPN's top-down structure is on the right side. FPN is a feature acquisition and blending component. That corresponding bottom-up map integrates with upsampled coarser-resolution feature maps. FPN manufactures P2-P6, which has 256 channels imported into numerous detectors. The model will deliver the output of instance attributes.

Boxes - boundaries of segmented class(ed, xmin,ymin,xmax,ymax)
Mask - a binary picture with segmented class values of 0 and 1
Labels - predicted class name
Scores - predicted class score of metrics from 0 to 100 %

c. Skin Detection

The skin detection function is utilised to extract skin from a human, using a maximum colour range of [0,133,77] and a minimum colour range of [235,173,127]. The images were converted from BGR to YCR format, and the skin region was identified by specifying maximum and minimum colour values to constrain the YCrCb values.

d. Extracting colour distribution

Extcolours is a Python module that takes an image file location as input. It extracts the colours from the image pixel by pixel in RGB and groups them based on their colour codes to determine how often the same colour pixels are repeated in an image and outputs the number of pixels.

Syntax - `extcolors.extract_from_path(img_location)`

e. RGB to HSV

The RGB colour space is device dependent, according to []. So, the same signal or image might seem different on various devices. Because the chrominance and luminance components in RGB colour space are intermingled, it is unsuitable for colour analysis or colour-based segmentation algorithms. The HSV colour space approximates human eyesight and can be used to extract additional information. As a result, the RGB colour codes collected from the segmented picture were converted using the RGB to HSV conversion function.

f. Calculating colour weights

Each HSV colour value reflects the frequency of occurrence of the colour (the number of times the same pixel colour occurred in the image). To achieve a precise colour impact in a segmented image, the number of occurrences of the same colour in a segmented object is divided by the total number of segmented object pixels.

g. Data Transformation

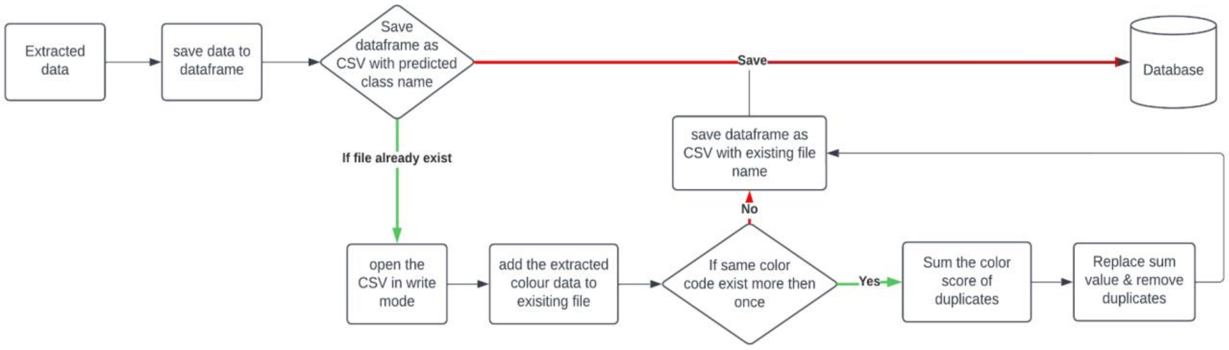

Extracted data
save data to dataframe
Save dataframe as CSV with predicted class name
Save
Database
If file already exist
save dataframe as CSV with existing file name
No
open the CSV in write mode
add the extracted colour data to exisiting file
If same color code exist more then once
Yes
Sum the color score of duplicates
Replace sum value & remove duplicates

Fig. 9 Data transformation

The extracted colour data from the image are converted into a dataframe and stored as CSV with the class name as the file name. The colour data, which includes colour values and weights, will be stored across each row of the data frame, and the data frame stored under the Predicted class name (e.g.-The colour extracted from the predicted boat segment has been saved in the CSV file boat.csv). When the same class is detected in consecutive photos or on the same image, the most recent data would be added to the existing file, resulting in a final dataset which contains each CSV file as generated train data for each class. Also, if the same colour code appears more than once in the data frame, the colour weights value of repeated colour values has been summed together. Remove the duplicate colour values, leaving only one entry with the total colour weights.

- RGB - RGB colour values
- HSV - HSV colour code of the object
- Pixels - weights calculated by dividing the occurrence of pixels by the total pixels of the object

h. Normalising the colour weights

Data preparation in machine learning usually use the normalisation technique. Without distorting variances in value ranges or losing information, normalisation seeks to change the values of numerical columns in a database to use a comparable scale. The rank-based normalisation method is utilised in this study to sort and rank each value of each predicted class-generated train data based on pixel weights.

1. Create empty dictionary A
2. Assign rank as 1
3. Sort the received data
4. Iterate through sorted data in a variable called num
    a. If num is not in A then append num as key and rank as value to A
    b. Increment rank by 1
5. Returns rank of each rows

The rank will be in the int with the scale of 1 to the length of the data. The value scaler function is used to normalise the value from 0 to 1.

4.3.2. Workflow

The training images will be saved in the folder and fed into Detectron2. That segments one or more objects in the picture. Then EXTCOLOURS analyses each object in the image and extracts the colours from each object, and the colour weights are computed by their respective values. The retrieved colour values are then transformed to HSV using the RGB to HSV function. Before storing the extracted data in CSV, the converted colour values and colour weights for their corresponding object from the picture will undergo data transformation. This procedure has repeated for each object in each image of the training dataset. Finally, there will be a set of CSV files with the predicted name as the file name. The ranks are allocated to each row by considering the colour weights value ranging from 1 to the number of colour values available in the file. Then the ranks are normalised on a scale of 0 to 1 for each data file.

4.4 Test Image

4.4.1. Architecture

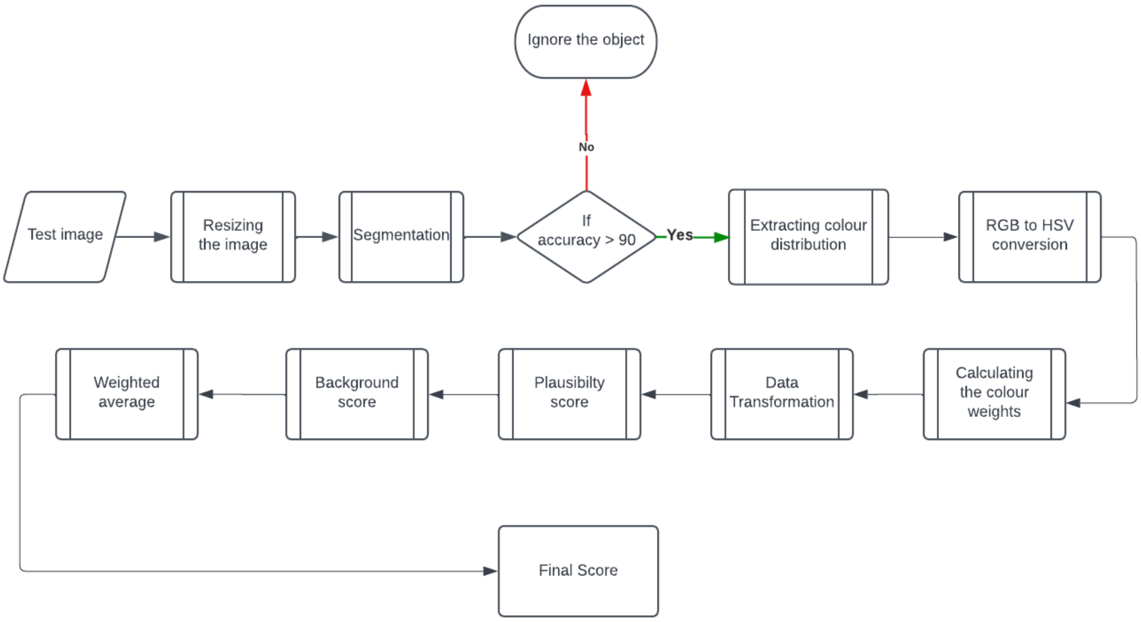


Fig. 10 Testing architecture

a. Data Transformation

The test picture's data transformation is the same as the training procedure, but the occurrence of the colour and the total number of pixels of an item from which the colours has extracted are included in the dataframe for scoring purposes.

- RGB - RGB colour values
- HSV - HSV colour code of the object
- Class_name - predicted class object
- Freq of colour - total count of particular colour occurred in object
- Total no pixels- object’s total number of pixels
- Pixels - occurrence of colour divided by the total pixels of the object to compute weights

The colour values from the test image are stored in a dataframe named Final dataset with the columns shown above.

b. Plausibility score

In the plausibility score step, a score was assigned to each colour existing in the test image's object using the Generated train dataset. It is been accomplished by iterating each row in the Final dataset to obtain the Class name, utilised to navigate and retrieve the training dataset of the specific object to discover the nearest or exact plausible colour code for the test image object HSV colour value.

The closest colour for the test data HSV in the Generated trained dataset was obtained utilising the Sphere Euclidean distance.

i) Sphere Euclidean distance

The Euclidean distance calculates the shortest path between two or more points by examining the root of square differences in point coordinates. By modifying the euclidean distance to determine the difference between two points in the sphere, a new Euclidean distance for the curved surface was proposed.

$$\sqrt{(|(360\ -\ Max)| + Min)^2\ +\ (S_2 - S_1)^2\ +\ (V_2 - V_1)^2} \quad \textit{(1)}$$

$$Point\ 1 = (H_1\ S_1\ V_1) \qquad Point\ 2 = (H_2\ S_2\ V_2)$$

$$Maximum = Max(H_1\ , H_2) \qquad (2)$$

$$Minimum = Min(H_1\ , H_2) \qquad (3)$$

Substitute (2) & (3) in (1)

Since the Hue in the HSV is a sphere that is measured by degrees from 0 to 360. The proposed Euclidean distance is to find the nearest point in the sphere form of HSV. Considering the hue point(H1 & H2), the maximum and minimum are obtained among them. Then proceed to substitute the maximum and minimum of H1 & H2 in the sphere Euclidean distance equation. As a result, the nearest point for the sphere can be obtained. As well as Calculate the shortest distance from the clockwise direction using the standard Euclidean equation. So, Finally, the minimum distance from clockwise and anti-clockwise is to be mentioned as the shortest distance.

The colour value of the test image is considered as point 1, and the sphere Euclidean distance with all colour values is calculated using colour values from the generated train dataset as point 2. The colour value with the shortest distance for point 1 in the produced train dataset has been picked, and the Nor rank value of the picked colour value has been assigned as the score for the test image colour value. Only if the Euclidean distance is equal to 0; if the Euclidean distance is not equal to 0, then Nor rank used as the plausibility score is divided by the distance between point 1 and point 2.

c. Background Score

The mask values of the detected objects in the test image are stored as an array in the list, and masks are inverted to crop out the detected objects, Which were scored using the generated train dataset. Finally, unscored background pixels were left in the image.

i) Colourfulness

(Hasler and Suesstrunk 2003) Requested 20 novice participants to rank the colour vibrancy of the photographs on a scale of 1 to 7. 84 photos made up the set of photographs used in this survey. the following scale values:

1. Not colourful
2. Slightly colourful
3. Moderately colourful
4. Averagely colourful
5. Quite colourful
6. Highly colourful
7. Extremely colourful

They concluded at a straightforward measure after a series of experimental computations that was connected with the outcomes of the participants.

Derivation for image colourfulness metric

$$rg = R - G$$

$$Yb = \frac{1}{2}(R + G) - B$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^{2} + \sigma_{yb}^{2}}$$

$$\mu_{rgyb} = \sqrt{\mu_{rg}^{2} + \mu_{yb}^{2}}$$

$$C = \sigma_{rgyb} + 0.3 * \mu_{rgyb}$$

In equation 1, rg represents the difference between the Red and Green channels, where R stands for Red, G for Green, and B for Blue. In the following equation, yb is half the total of the Red and Green channels less the Blue channel. Before obtaining the final colourfulness measure, the standard deviation (σ_{rgyb}) and mean (μ_{rgyb}) are computed.

ii) Colour cast

The colour cast can be confirmed, if the chromaticity distribution is primarily a single peak in the ab chromaticity coordinate plane of the histogram, perhaps if the distribution is relatively concentrated as well as the chromaticity average D gradually increase, the larger the chromaticity average, the more severe the colour cast. However, if the chromaticity distribution in the histogram on the ab chromaticity coordinate plane contains noticeable numerous peaks distributed relatively. Then the degree of a colour shift in the picture will be substantially decreased, also maybe no colour shift.

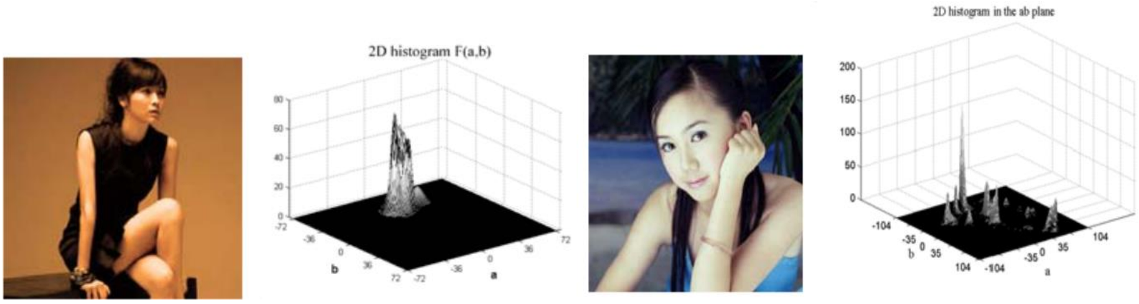


Fig. 10 colour cast and normal image

The image has a noticeable colour cast, and its [a,b] histogram has just one peak, as seen in the figure above. Whereas the second image is highly symmetrical, and its [a,b] histogram has similarly relatively distributed uniformly.

$$d_a = \frac{\sum_{i=1}^{M}\sum_{i=1}^{N} a}{MN} \quad d_b = \frac{\sum_{i=1}^{M}\sum_{i=1}^{N} b}{MN} \qquad (1)$$

$$D = \sqrt{d_a^2 + d_b^2} \qquad (2)$$

$$M_a = \frac{\sum_{i=1}^{M}\sum_{i=1}^{N} (a - d_a)^2}{MN} \quad M_b = \frac{\sum_{i=1}^{M}\sum_{i=1}^{N} (b - d_b)^2}{MN} \qquad (3)$$

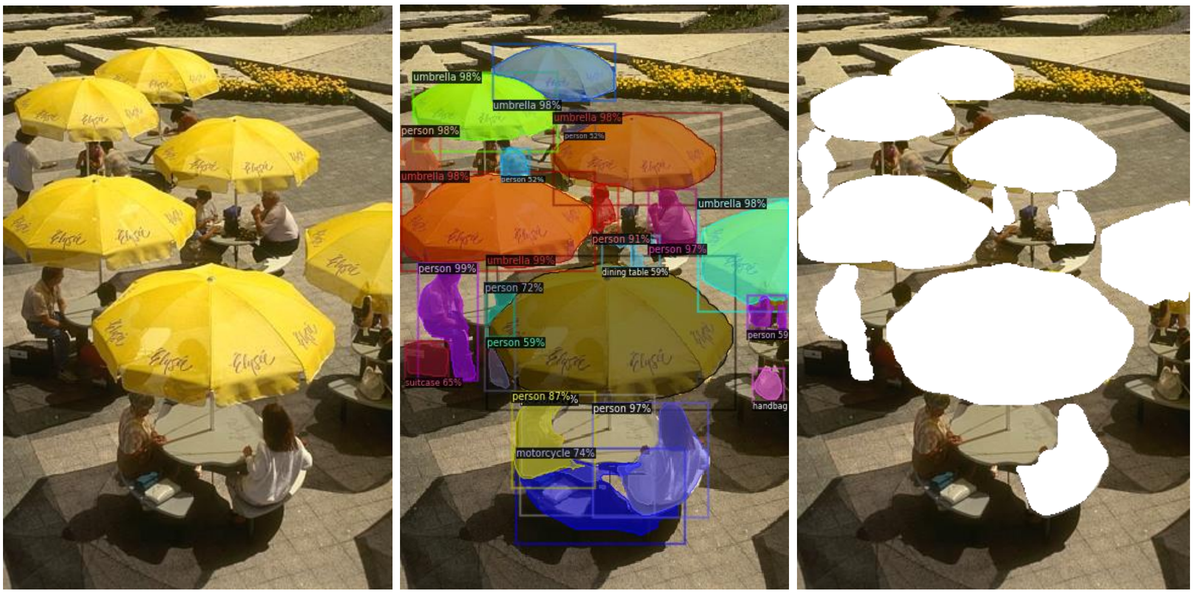
$$M = \sqrt{M_a^2 + M_b^2} \qquad (4)$$

$$K = D/M \qquad (5)$$

where M denotes the image's width and N denotes the image's height in pixels. The comparable circle's centre coordinates are (da, db) and its radius is M on the a-b chromaticity plane. A-b chromaticity plane neutral axis origin and centre of the matching circle are separated by a distance D of (a = 0, b = 0). Thus, the actual position of the matching circle on the a-b chromaticity plane will define the colour cast of the entire picture. If da > 0, the colour is reddish; otherwise, it is greenish. If db > 0, the colour will be yellowish; otherwise, it will be blue. The colour cast factor K is added, with a higher K value indicating a more severe colour cast.

Combining colour cast and colourfulness

As a result, the colour cast and colourfulness of the background image can be analysed to provide a score to pixels in the background based on the colourfulness value. First, the colour cast for the entire image is determined. Then the background image colourfulness is computed and assigned as the score for background pixels.



(A) (B) (C)

Fig. 11 The given test image and processed background

If the colourfulness value is above 100 and the colour cast value has surpassed the colour cast threshold of 1.5, then the background score will be divided in half. In the above figure, (A) is the test picture fed into the detecron2 for segmentation, (B) the segmented image, and (C)the background image, also known as unsegmented pixels, which will be evaluated based on the background's colourfulness and colour cast.

d. Weighted average

Rather than measuring the score evenly. Each object in the image obtains a score from the weighted average depending on how much it contributes to or influences the image. This is accomplished by multiplying the object's total plausibility score by the significance of the object. Whereas the object's plausibility score may be calculated by adding the plausibility scores of all the object's colour values. While its weights could be calculated by dividing the total number of pixels in the object by the sum of all segmented objects in the picture.

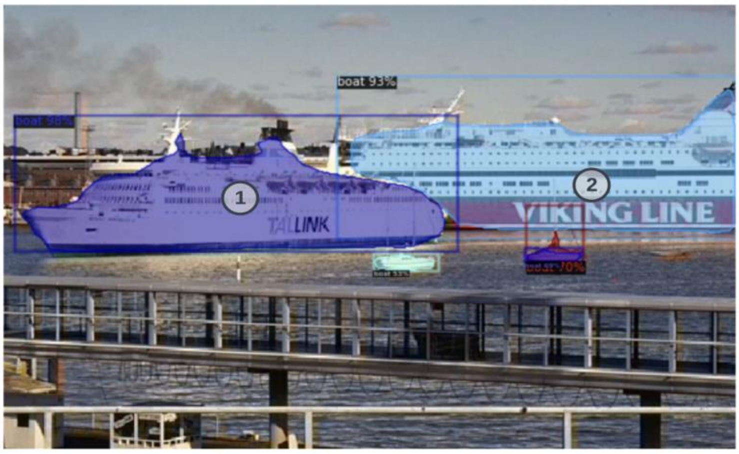


Fig. 12 Segmented image

The contribution for ship 1 in the final score is allocated less when compared to ship 2 since ship 1 occupies fewer pixels in the images than ship 2. The plausibility scores of all the objects are summed together and scaled from 0 to 100 using the Value scalar function.

e. Final score

The score of segmented object pixels and unsegmented pixels, also known as background pixels, are merged on a ratio basis in the final score by adding up all the plausibility scores of the object and the colourfulness of the background.

4.4.2. Workflow

The test image will be transmitted to the model initially, with each object in the image passing through the entire flow after being segmented using detectron2. If the object's accuracy is more than 90%, then the object will proceed to the further processing step. If the segmented object prediction class label is a person, then the skin extraction function is used to segregate the skin from the rest of the segmented image. The colours in the object are then extracted using EXTCOLOURS. The extracted colours are converted to HSV from RGB colour space using RGB_TO_HSV, and colour weights are determined before all the calculated data are transformed into a dataframe called the Final_dataset. Each row in the Final dataset dataframe has been iterated to obtain the closest colour value for point 1 in a specified class of generated train data as point 2 using sphere Euclidean distance and the colour value with the shortest distance for point 1 in the generated train data Nor rank is chosen and calculated based on the distance between the colour points and determined as the plausibility score. The background score evaluates the colourfulness and colour cast of the background using the image's un-segmented pixels. The weighted average has been used to assign a score to each object based on their contribution to the image.

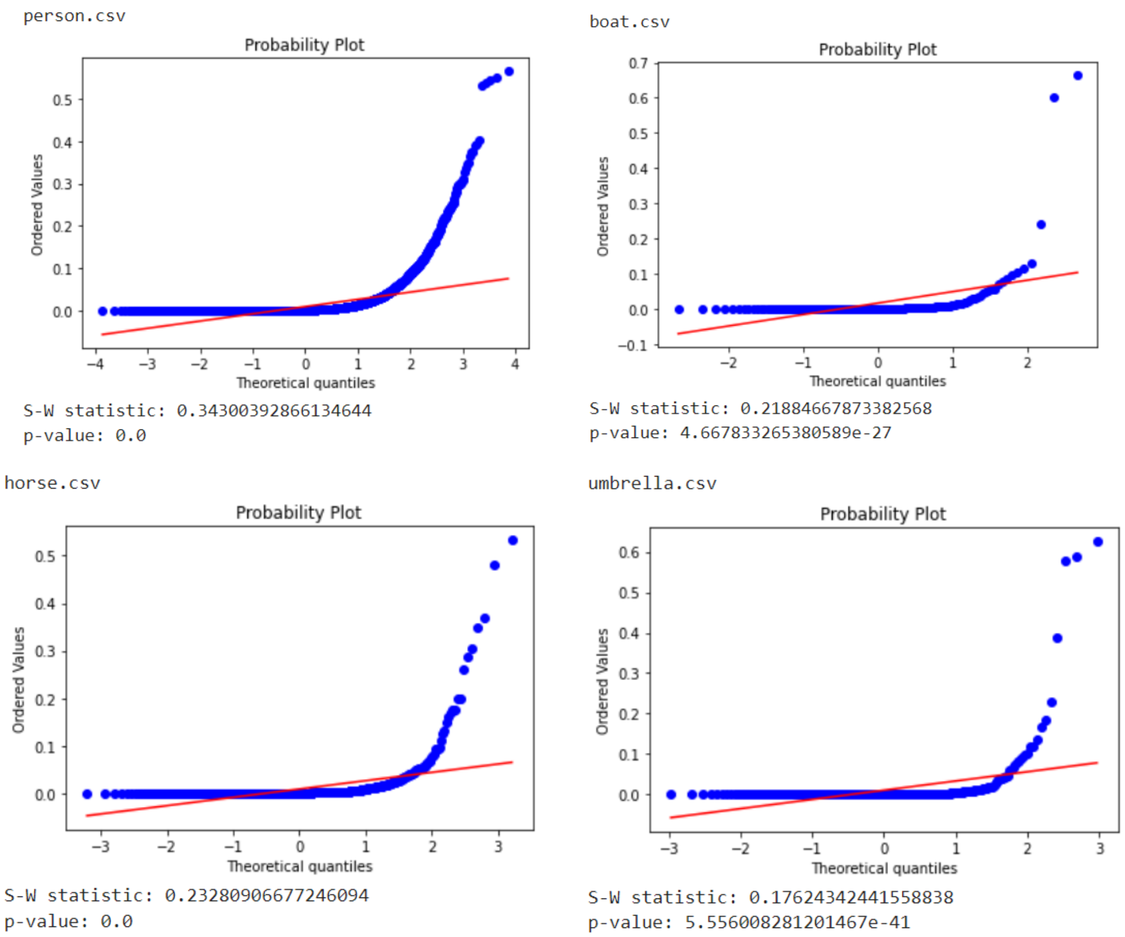
Finally, the final score for the test image has calculated by combining the weighted average score of the objects in the image with the image's un-segmented background score in a 1:5 ratio.

4.5. Alternative consideration & rejected

Many approaches were considered during the algorithm's development but ultimately discarded because they generated inconsistencies or did not adhere to specified statistical principles.

4.5.1. Normalisation

To normalise the pixel occurrence in the generated train data before using the ranking technique. The min-max normalisation was used to obtain the plausibility score. However, the normalised values of pixel occurrence seemed inexact due to the higher variation between the values of pixel occurrence. So, a normality test is performed for each class pixel's occurrence data to confirm the distribution of the data by plotting the graph and using the Shapiro normality test.



```
person.csv
S-W statistic: 0.34300392866134644
p-value: 0.0
```

```
boat.csv
S-W statistic: 0.21884667873382568
p-value: 4.667833265380589e-27
```

```
horse.csv
S-W statistic: 0.23280906677246094
p-value: 0.0
```

```
umbrella.csv
S-W statistic: 0.17624342441558838
p-value: 5.556008281201467e-41
```

Fig. 13 Q-Q plot & Shapiro-Wilk Test results of pixels columns

The pixels data were not evenly distributed, as seen by the Q-Q plot below, and the Shapiro-Wilk test's P-value was less than.05, indicating that the null hypothesis of the Shapiro-Wilk test is rejected. This indicates that we have statistically significant evidence to conclude that the sample data does not follow a normal distribution, has extreme values, and has a substantial number of values that are near zero. Hence the normalisation approach cannot be directly performed since the data is not normally distributed.

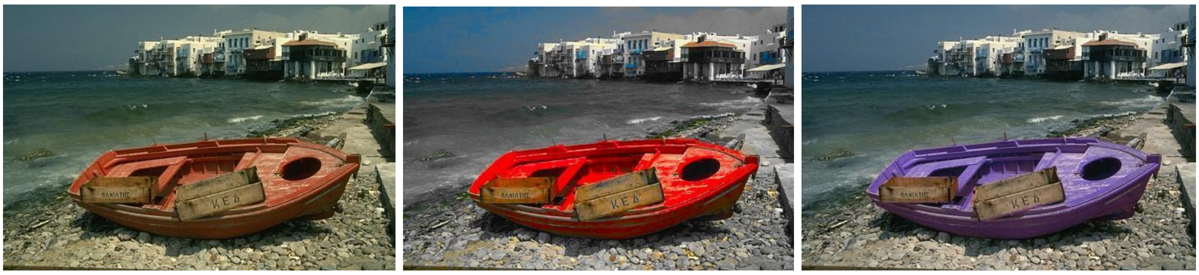
4.5.2. Euclidean distance

The distance between two points was calculated using the Euclidean distance. Previously before adopting the sphere Euclidean distance, To assign the plausibility score, the usual Euclidean distance was employed to locate the closest colour for the colours contained in the test picture using the produced train data. However, the Euclidean distance was well-defined on a flat surface and can be used only to find the distance between two points on a plain surface, Whereas the HSV colour model Hue is a sphere in structure represented by degrees of 0 to 360. So the nearest value for 354 should be 0 rather than 340, which would be unattainable using traditional Euclidean distance.

4.6. Approach

4.6.1. Version 1

Initially, the same pipeline was used to generate training data from the training dataset in the first version. The method for retrieving the test image's colour information remains the same, from segmenting it with detectron2 to transforming its colour data into a dataframe called the Final_dataset. The plausibility scoring technique utilises all three properties H, S, and V of HSV colour models to find the exact or nearest colour from generated train data using Sphere Euclidean distance to provide the plausibility score for the colour. The final score is calculated by adding up all the colour values, and plausibility scores of each object and multiplying with their subject weights to get the score.



(A) (B) (C)

Fig. 14 The above images are evaluated using version 1

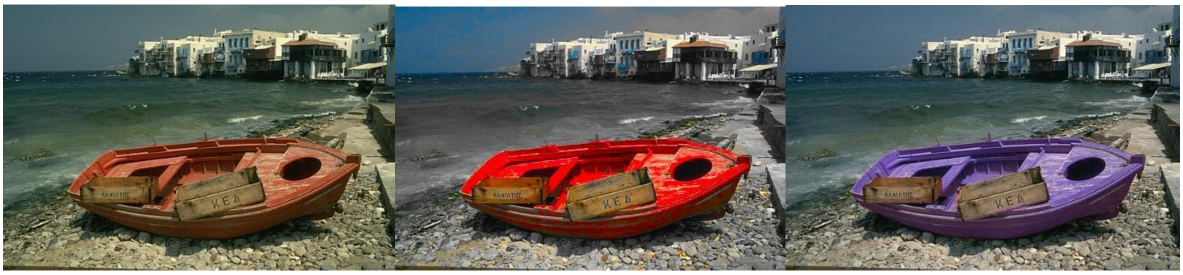
Table 1 The score produced by evaluation algorithm for the above-mentioned images

Label	Image Name	Final Score
A	118020_gt.jpg	32.74
B	118020AC_4.jpg	55.99
C	118020AF_12.jpg	17.26

4.6.2. Version 1.1

The primary goal of the research is to evaluate greyscale images which have been colourized via the colourization algorithm. As a result, the goal of v1.0 is to rebuild the algorithm relevant to the functionality of the colourization approaches. So, most colourization algorithms will convert the image to CIEL*A*B from RGB to format and only consider the A and B channels for prediction. The L channel will be the Luminance is the total quantity of light that a surface reflects in the directions of an observer's eyes, also known as brightness, also referred to as Value in the HSV colour model, Which will remain unmodified in the vast majority of colourization algorithms.

So, by considering all the above. Usually, sphere Euclidean distances have calculated between point 1 test image colour HSV colour code and point 2 from the generated train data. The distance calculation between point 1 and point 2 has dropped from three vector points to two vector points, assuming just Hue and Saturation to identify the nearest colour for point 1. This improved the algorithm in simplifying the computation of Euclidean distance, resulting in a shorter runtime, Since the method can calculate and locate the nearest colour for point 1 by considering only two vectors among 1000s of rows.



(A) (B) (C)

Fig. 15 The above images are evaluated using version 1.1

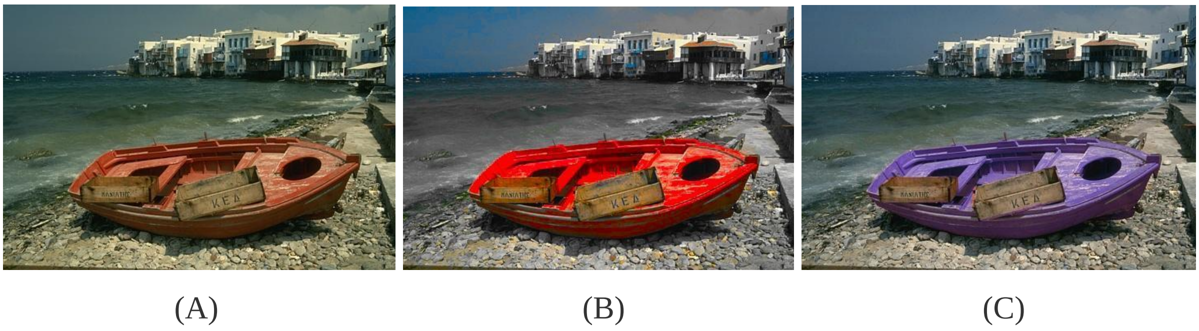
Table 2 The above image score is generated by only considering H & S

Label	Image Name	Final Score
A	118020_gt.jpg	83.83
B	118020AC_4.jpg	92.70
C	118020AF_12.jpg	41.44

4.6.3. Version 2

The goal of version 2 is to fine-tune the algorithm so that it does more than only assess and evaluate the fidelity of a given image. Since the majority of image evaluation metrics were designed to assess the image fidelity of a given image to its reference image rather than the image quality. Also, just because an image varies from the original does not always indicate that the image is of worse quality (Hasler and Suesstrunk 2003). Image colours can be damaged in two different ways: by colour casts or by a loss of colourfulness. So, the model has been tuned to compute the image's colourfulness and colour cast for the unsegmented background, which is included in the final score. The standard pipeline was used to score the given test image, from segmentation to determine the weighted average.

However, in between the background of the image is obtained by eliminating all detected objects from the given image. This is used to compute the colour cast and colourfulness factored into the image's final score The model is set up in such a way that if the background value exceeds the colour cast threshold of 1.5 and the colourfulness threshold of 100, the background's colourfulness value is divided by 2. Because if the image is influenced by a colour cast, then a majority of the pixels will be coloured, resulting in a high colourfulness score. Furthermore, none of the natural photographs will have a colourfulness score greater than 100. If this is the case, the picture will most likely be more vibrant than usual, which makes it appear unnatural. The background score is then added to the weighted average computation to get the image's final score.



(A) (B) (C)

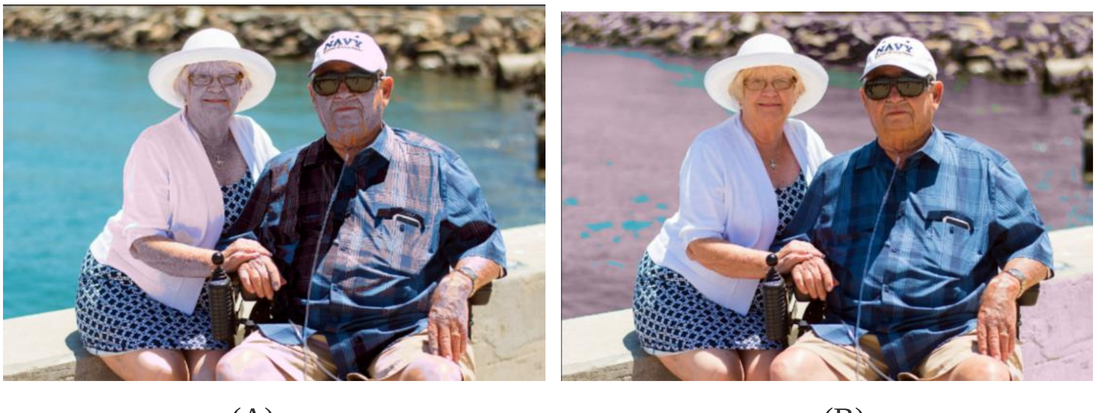
Fig. 16 The above images are evaluated using version 2

Table 3The above image score is generated by only considering H & S

Label	**Image Name**	**Object score**	**Background score**	**Final Score**
A	118020_gt.jpg	83.83	22.69	35.78
B	118020AC_4.jpg	92.70	27.39	41.36
C	118020AF_12.jpg	41.44	20.93	25.38

4.6.4. Version 2.1

V2.1 is a slightly tweaked version of version 2 with a few improvements in combining detected object scores and background values in the final score computations. The evaluation algorithm has accomplished the milestone of assessing and scoring every test image pixel after integrating the colourfulness and colour cast approach. however, there is a minor abnormal fluctuation in the final result. After analysing the whole flow of the evaluation process, it was observed that there was a flaw in allocating the weights of the detected objects and background in the final result. However, the background score was considered in the final result based on pixel weights. The problem is that the background contributes to around 3/4 or 2/4 of the image's weights in most images. Meanwhile, the detected objects contribute less than half of the image's weights, Which is unacceptable to use a pixel-based weighted score in the final result since foreground objects have a more significant impact on the image than background objects.



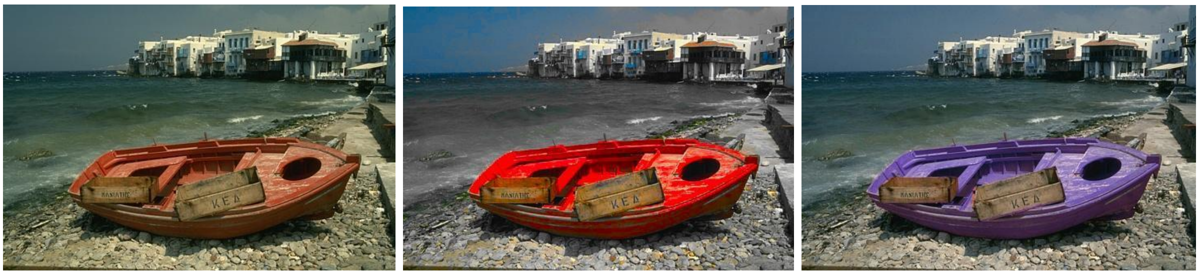
(A) (B)

Fig. 17 Images that has be gone through colourization

By looking at the below image (a), it is visible that the image is quite acceptable even though the ¾ of the image falls under the background, which is not fully coloured; in an image (b), the image feels odd even though the background that holds higher weights is perfectly colourized whereas the object in the image has significantly fewer weights which are uncoloured or with skewed colours.

To support this discussion, an analysis had performed using the obtained detected objects score and background score together with user rating data from the HBCD-Colorization dataset. The correlation between the user rating data, the detected object score, and the same image's background score are computed attached in appendix. The user rating strongly correlates with the image's foreground (detected objects) and has a substantial negative correlation with the background score. It suggests that viewers commonly give the foreground object more priority, even though it has fewer contributions to the image than the background.

Perhaps instead of allocating weights to the image's foreground (detected objects) and background based on pixels occupied in the final score. The weights had shared in an 80:20 ratio, with the foreground (detected items) contributing 80% of the final score and the image's background contributing 20% of the final score.



(A) (B) (C)

Fig. 18 The above images are evaluated using version 2

Table 4 The above image score is generated by only considering H & S

Label	**Image Name**	**Object score**	**Background score**	**Final Score**
A	118020_gt.jpg	83.83	22.69	71.61
B	118020AC_4.jpg	92.70	27.39	79.64
C	118020AF_12.jpg	41.44	20.93	37.34

4.6.5. Version 3

V3 have specially designed to handle the segmented class known as people. The human presence in the image have segmented under the person class, and the difficulty here is that the person is generally segmented from head to toe, including clothes and accessories. Globally, clothing and accessories are available in all possible colours and shades, and it lacks any plausible colour. So, while producing the training dataset, if colour values from the person are scraped without implementing any skin detection, all of the colour data of a segmented person, including the colour of the clothing, is scraped and stored in the person class data. This causes generated train person data to be created, with all the colours and shades considered the person's plausible colour.

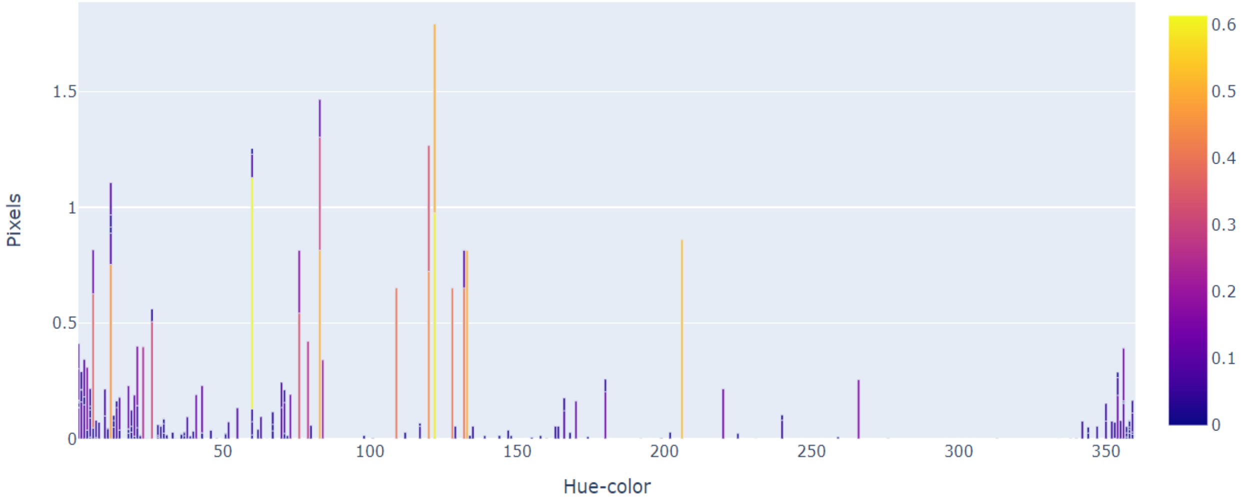


Fig. 19 Histogram of person-generated train data before implementing skin extraction

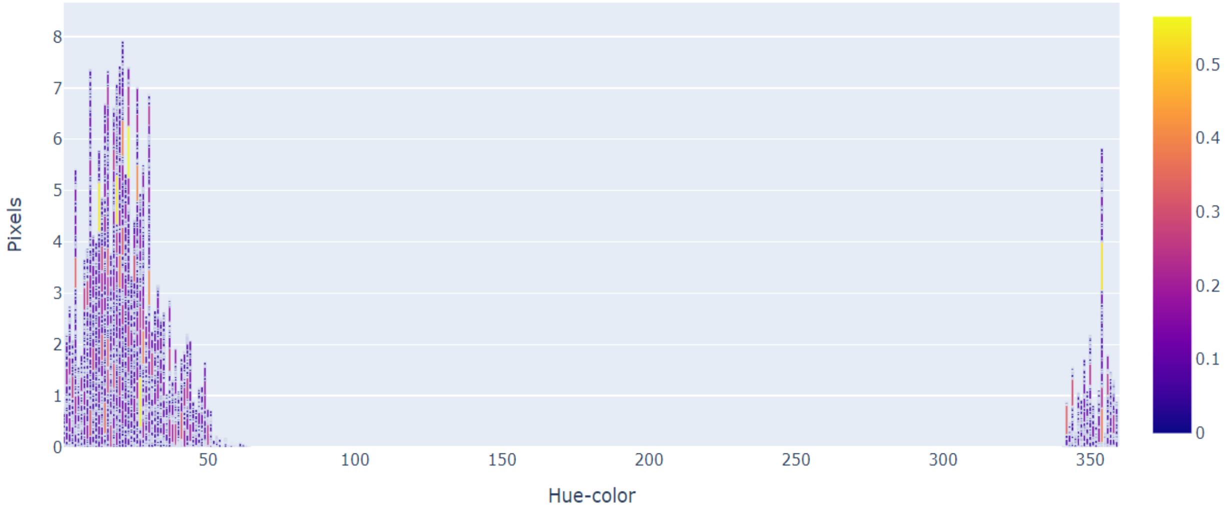


Fig. 20 Histogram of person-generated train data after implementing skin extraction.

The above histogram has hue colour of 0 to 360 in x axis and pixels weights in the y axis where the higher the pixel weight for the HSV value lighter the colour in histogram. For example, if the algorithm is tested with an image of a person with blue skin to evaluate the score with that generated data, the evaluation metrics provide a higher score since the blue colour is a realistic colour based on the generated train human data. Probably, The algorithm is unable to assign scores by distinguishing between the colour of the clothing and the colour of the skin. The image's major flaw, skin colour, was missed by evaluation measures. So, in the segmented human, the skin alone is extracted, excluding the clothes and accessories, which is unnecessary.

5. Results & Evaluation

After addressing many algorithm errors, the algorithm for evaluating the colourised image was successfully developed. It will go through the testing procedure by analysing the algorithm's outputs for various images from the HBCD dataset.

Check 1 - ***Does the proposed evaluation metrics algorithm generates a score such as human perception?***

The algorithm's ability in HVS-based evaluation could be evaluated by evaluating the correlation between the user score and the algorithm-generated score for the HECD dataset sample images.

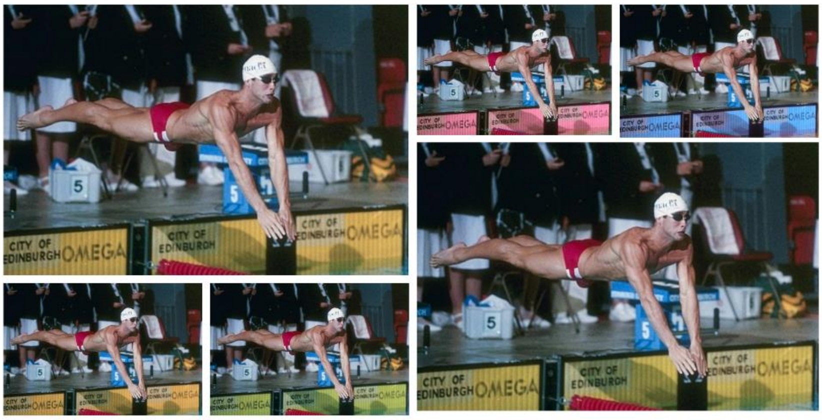


Fig. 21 Sample images of 153093

Table 5 The set of sample images scored by both user and algorithm

filename	*user_ratings_mean*	*new_colour_value*
153093AB_1.jpg	2.315789474	34.48
153093_gt.jpg	4.416732439	85.43
153093AG_1.jpg	4.05	78.66
153093AC_4.jpg	3.368421053	66.66
153093AF_9.jpg	4.473684211	73.05
153093D_4.jpg	2.45	63.34
153093F_7.jpg	3.894736842	78.21

The table contains data from photographs that have been scored by the algorithm on a scale of 0 to 100, and the mean value of the user score is used to determine the user's rating for the image. The algorithm's score for the images above is relatively realistic. In contrast, the only way to evaluate the algorithm's effectiveness is to determine whether its outcomes are identical to or comparable to the average user score.

The correlation between user and algorithm ratings for the sample images is computed to achieve this. The user score comes under the Likert scale because it has a scale of 0 to 1, which is continuous data that follows a monotonic relationship or ordinal data. In addition, the algorithm's score is on a scale of 0 to 100, which is a continuous variable. Therefore, one of the first and most significant approaches is verifying the correlation coefficient's relevance through hypothesis testing using Spearman's correlation.

Hypothesis testing

Null Hypothesis: H0:ρ=0 *Alternate Hypothesis*: Ha:ρ≠0

α = 0.05

Null Hypothesis H0 : The two samples of the score given by the user and algorithm are independent and have no correlation.

Alternate Hypothesis H1 : There is a dependency between the samples and highly correlated.

The p-value is 0.036

The p-value, 0.026, is less than the significance level of α=0.05.

Decision: the null hypothesis is rejected.

Conclusion: There is statistically significant evidence to reject the null hypothesis and conclude there is a significant linear relationship between the score given by the user and the algorithm for an image. because there is a substantial correlation of 0.786, the correlation coefficient is considerably different from zero.

Check 2 - *Does the evaluation algorithm outperform the popular objective-based full reference evaluation metrics*

The initial goal is to create an evaluation algorithm that coincides with the HVS. The second most objective is to produce a better score than the objective based on full reference IQA evaluation metrics. PSNR, the most popular metric for estimating quality, and The SSIM was related to the HVS's quality and perception, is chosen for comparison. So the spearman correlation is calculated with the user score, and all among algorithms ,score, PSNR and SSIM to check have a better correlation with the user score.

Check 2.1 - ***Visual representation***

The correlation coefficient of the user score with an algorithm-generated score, PSNR, and SSIM are represented visually.

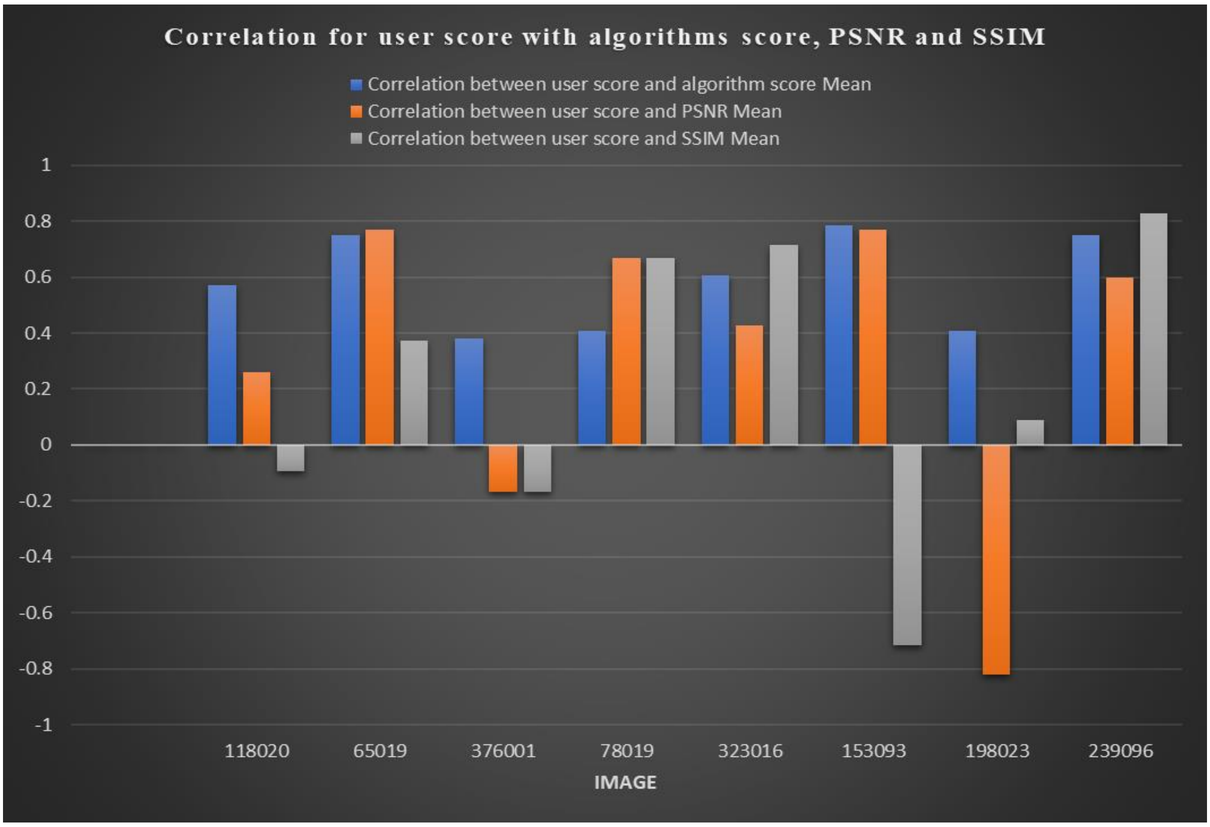


Fig. 22 Correlation co efficient of user score with algorithm score, PSNR, and SSIM

The above chart makes it quite convincing that the proposed evaluation algorithm has performed well in most of the sample images compared to the PSNR and SSIM. The PSNR and SSIM have a negative correlation for a few images.

Check 2.2 Statistical Representation

Although the suggested evaluation technique outperformed both PSNR and SSIM in most sample images, a statistical comparison was performed to get a more detailed observation.

Table 6 Correlation coefficient of user score mean with algorithms score, PSNR and SSIM

Image	***Correlation between user score & algorithm score***		***Correlation between user score and PSNR***		***Correlation between user score and SSIM***	
	Mean		***Mean***		***Mean***	
	Pearson correlation	***Fishers z***	***Pearson correlation***	***Fisher z***	***Pearson correlation***	***Fisher z***
118020	*0.789*	*1.0688*	*0.356*	*0.3723*	*-0.208*	*-0.2111*
65019	*0.806*	*1.1155*	*0.926*	*1.6296*	*0.433*	*0.4636*
376001	*0.378*	*0.3977*	*0.623*	*0.7299*	*0.007*	*0.0070*
78019	*0.368*	*0.3861*	*0.658*	*0.7893*	*0.575*	*0.6550*
323016	*0.741*	*0.9527*	*0.355*	*0.3712*	*0.221*	*0.2247*
153093	*0.841*	*1.2246*	*0.683*	*0.8347*	*-0.554*	*-0.6241*
198023	*0.444*	*0.4772*	*-0.751*	*-0.9752*	*0.599*	*0.6916*
239096	*0.837*	*1.2111*	*0.426*	*0.455*	*0.818*	*1.1507*
Average z		*0.8542*		*0.5258*		*0.2947*
Average Correlation		***0.693***		***0.482***		***0.286***

The above table contains the correlation coefficient data of user score mean with algorithm score, PSNR and SSIM on the same sample images. The correlation coefficient for each set of sample images is calculated for all the evaluation metrics, then the correlation coefficient of each sample. To eliminate bias, according to (Alexander 1990), the correlation coefficient of each sample is transformed to fisher z. Each measure's average fisher z is calculated and back transferred to the correlation coefficient. Looking at the average correlation coefficient of all three metrics with the user score, the proposed colourization evaluation metrics have a stronger correlation to the user score.

Evaluation

The algorithm produces the score as much as humans, and now let us dive into deep analysis. Even though the user score is ordinal and the algorithms score is a continuous variable, it allows us to test the spearman's correlation (Johnson and Creech 1983). “This rationale centres on the fact that Likert or ordinal variables with five or more categories can often be used as continuous without harming the analysis”. So the Pearson correlation is also taken into consideration in the upcoming analysis.

The mode is generally the most prevalent score among a collection of numbers that may be utilised as numerical data. Also, it is often used to find a central tendency for categorical data. So the mode of the user score is computed to calculate the correlation with the algorithms score.

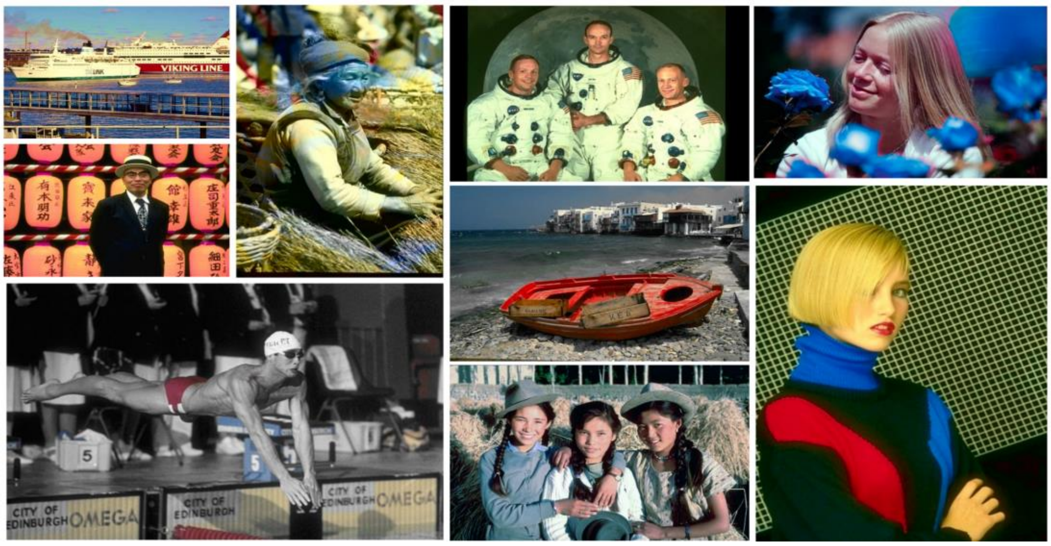


Fig. 23 Sample images from HECD dataset

Table 7 The Correlation between user score and score produced by the algorithm

Image	***Mean***		***Mode***	
	Pearson correlation	***Spearman correlation***	***Pearson correlation***	***Spearman correlation***
118020	*0.789*	*0.571*	*0.667*	*0.600*
065019	*0.806*	*0.750*	*0.802*	*0.777*
376001	*0.378*	*0.381*	*0.669*	*0.77*
078019	*0.368*	*0.406*	*0.727*	*0.648*
323016	*0.741*	*0.607*	*0.845*	*0.797*
153093	*0.841*	*0.786*	*0.718*	*0.852*
198023	*0.444*	*0.408*	*0.608*	*0.546*
239096	*0.837*	*0.750*	*0.909*	*0.926*

The above table represents the correlation between the rating given by the users and the score produced by the algorithm to a set of sample images in each series. Each sample image mean and the mode was considered for computing both Pearson and spearman correlation.

Discussion about Pearson & Spearman correlation

Let's looks into the variations between the two correlation methods

There is a considerable variation between the Pearson correlation and the Spearman correlation.

Table 8 153093 sample images data

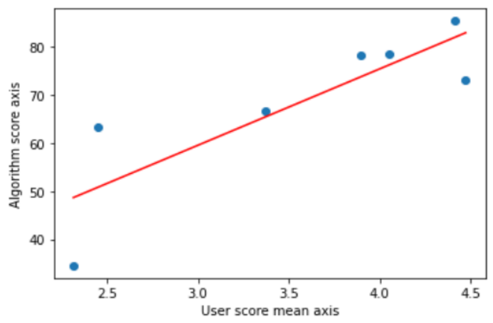
filename	***User_Ratings_Mean***	***User_Score_mode***	***Algorithm score***
153093AB_1.jpg	2.315789474	2	34.48
153093_gt.jpg	4.416732439	5	85.43
153093AG_1.jpg	4.05	5	78.66
153093AC_4.jpg	3.368421053	3	66.66

153093AF_9.jpg	4.473684211	5	73.05
153093D_4.jpg	2.45	1	63.34
153093F_7.jpg	3.894736842	4	78.21

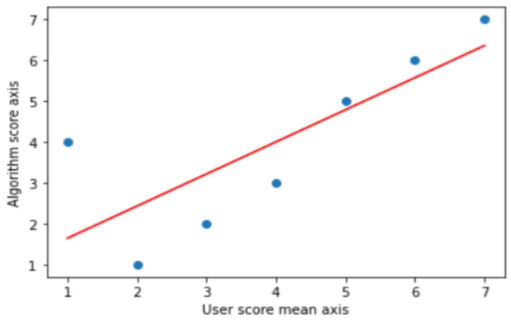
The above table has chosen to conduct a comparison between the Pearson and spearman correlation with both mode as well as mean.

Pearson correlation

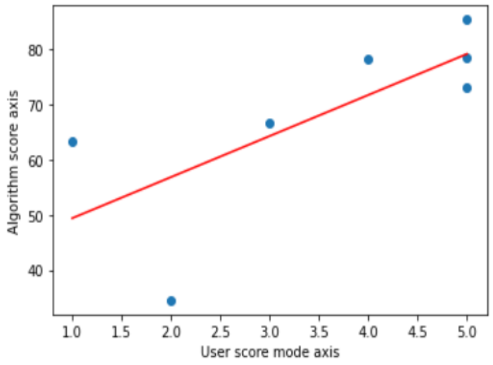
spearman correlation



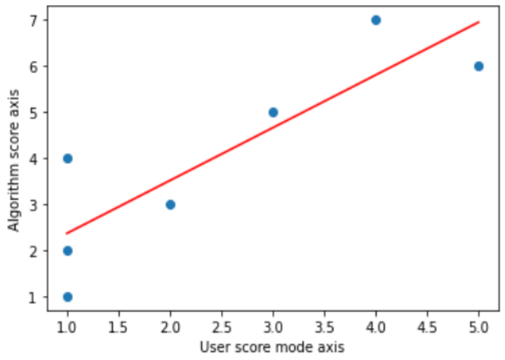
Correlation with mean = *0.841*



Correlation with mode=*0.786*



Correlation with mean =*0.718*



Correlation with mode =*0.852*

The correlation coefficients for user score with algorithm score is positively correlated in both mean and mode. Since Spearman correlation coefficients only quantify monotonic correlations, the variables generally move in the same or opposite direction but may not always do so steadily. On the other hand, with a linear connection, the rate remains constant. In contrast, Pearson correlation coefficients assess the linear relationship among two variables, their values are nearly identical.

6. Conclusion

The fundamental purpose of the research is to build a colourization evaluation metric that collects the colour distribution of all objects through segmentation and uses the generated colour distribution of each object to evaluate the colourized image using mathematical formulas, This was accomplished by using pre-trained instance segmentation detectron2 model to separate the objects from the picture and the colours from the objects to generate the colour distribution. That can be used to evaluate the colourized image. As a significant milestone, the un-segmented pixels of the image are evaluated based on HVS using colourfulness and colour cast. The final score has been produced by considering both the segmented objects assessed based on generated colour distribution using train data and un-segmented pixels that fall in the background section assessed using the colourfulness and colour cast. The results concluded that the proposed colourization evaluation algorithm performed better in the sample images in the HECD dataset than the PSNR and SSIM.

In comparison, the PSNR and SSIM are the FR-IQA that require the reference image to review only the fidelity towards its original image. In contrast, the proposed evaluation metrics are complete NR-IQA metrics that evaluate the image based on the HVS. However, the proposed evaluation metrics are not much correlated to the human score for the sample data because The NR-IQA is much more challenging than the FR-IQA and RR-IQA. (Kamble and Bhurchandi 2015) Because there is no reference picture, The statistics of the reference picture must be reproduced, which fulfils the nature of the HSV and the influence of distortions on picture statistics in an unsupervised manner.

6.1 Error Analysis

Also, Table 7 shows a considerable difference between the mean and mode in both Pearson and spearman correlation. That may be due to the larger standard deviation between the scores in the user data for each image, which may affect the mean. As a result, the statistical test is performed to see if the algorithm score is affected by the standard deviation of each sample picture.

Table 9 Variance between mean & mode for both correlation.

Image	***Variance between mean & mode***		***Overall variance***
	For Pearson correlation	***For Spearman correlation***	
118020	*18.3*	*4.8*	*23.1*
065019	*0.5*	*3.5*	*4.0*
376001	*43.5*	*50.5*	*94.0*
078019	*49.4*	*37.3*	*86.7*
323016	*12.3*	*23.8*	*36.1*
153093	*17.1*	*7.7*	*24.9*
198023	*27.0*	*25.3*	*52.2*
239096	*7.9*	*19.0*	*26.9*

The above table represents the variance percentage between the Mean and mode of the user rating and the algorithms score. Here ***376001*** image series has a much variance percentage among all the image series. So ***376001*** sample images have been chosen to analyse the mean and mode.

Table 10 Data of sample images from 376001 image series.

Filename	**user_ratings_mean**	**user_ratings_mode**	**Algorithm score**
376001A.jpg	4	4	51.4
376001AF_13.jpg	4.058823529	5	68.15
376001B_2.jpg	3.75	3	39.57
376001F_2.jpg	4.611111111	5	88.59
376001AF_6.jpg	4.2	4	44.34
376001AG_1.jpg	4.3	4	51.25
376001C_2.jpg	3.85	4	87.3

376001AG_2.jpg	3.9	4	50.63

The above tables holds the data for 36001 sample images mean & mode score assigned by the user and then algorithms score.

Table 11 standard deviation on user score and percentage error between user score and algorithms score

Image	**Standard deviation**	**Percentage error**
376001A.jpg	*0.725*	*35.8*
376001AF_13.jpg	*0.966*	*16.0*
376001B_2.jpg	*1.020*	*47.2*
376001F_2.jpg	*0.502*	*3.9*
376001AF_6.jpg	*0.768*	*47.1*
376001AG_1.jpg	*0.571*	*40.5*
376001C_2.jpg	*0.933*	*13.2*
376001AG_2.jpg	*1.165*	*35.1*

For each image, the percentage error is calculated by taking the mean of the human score as the actual value and the algorithm score as the experimental value. The standard deviation for sample data for 376001 is derived by considering the human score for each image. The data must be examined to see if the error between the actual human score and the algorithm score is independent or has a correlation. That may be accomplished by running the hypothesis test.

HYPOTHESIS TESTING

Null Hypothesis: $H0:\rho=0$ *Alternate Hypothesis*: $Ha:\rho\neq0$

$$\boldsymbol{\alpha = 0.05}$$

Null Hypothesis H0 : The two samples of the score given by the user and algorithm are independent and have no correlation.

Alternate Hypothesis H1 : There is a dependency between the samples and highly correlated.

The p-value is 0.621

The p-value, 0.621, is greater than the significance level of α=0.05.

Decision: DO NOT REJECT the null hypothesis.

Conclusion: There is no statistically significant evidence to conclude that there is a significant linear relationship between percentage error and standard deviation. because the correlation coefficient is NOT significantly different from zero and has a correlation of **0.208**

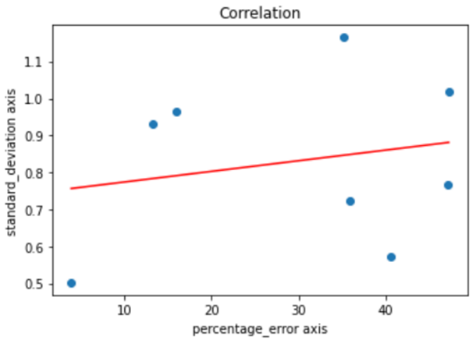


Fig. 24 Scatter plot for percentage error and standard deviation.

Data bias

The collection contains 7.5 times more photos of light-skinned individuals than dark-skinned people, twice as many men as females, and even fewer representations of dark-skinned females. According to (Zhao et al. 2021), The top two Types appear to be Fitzpatrick Skin 1 (15.4%) and Fitzpatrick Skin 2 (31.5%), whereas Types 6 and 5 account for just 1.7% and 1.9% of the cases, correspondingly. The lack of representation of people with darker complexion is an example of representational damage in and of itself.

6.2 Methodological limitations

The findings of this study, like the majority of research, have been examined in light of two limitations imposed by computational limits. The first is identifying objects that are not included in the coco dataset. As a result, colour distributions for objects other than those in the coco dataset cannot be created. The second method is to extract the skin based on colour space, which cannot discriminate between skin and clothing in skin tone.

6.3 Future works

The limitations can be overcome by using panoptic segmentation, assigning a class label to each pixel in the image, which allows collection colour distribution for every possible object (Panoptic Segmentation. [no date]). Also, use a deep learning approach to detect the skin.

References

Erdoğan, K. and Yılmaz, N. 2014. *Shifting Colors to Overcome not Realizing Objects Problem due to Color Vision Deficiency*. doi: 10.15224/978-1-63248-034-7-27.

H, K.V. and K, B.S. 2018. A Comprehensive Investigation of Color Models used in Image Processing. *International Journal of Computer Applications* 180(22), pp. 19–24.

Hasler, D. and Suesstrunk, S. 2003. Measuring Colourfulness in Natural Images. *Proceedings of SPIE - The International Society for Optical Engineering* 5007, pp. 87–95. doi: 10.1117/12.477378.

Iizuka, S., Simo-Serra, E. and Ishikawa, H. 2016. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics* 35(4), p. 110:1-110:11. doi: 10.1145/2897824.2925974.

Levin, A., Lischinski, D. and Weiss, Y. 2004. Colorization using optimization. *ACM Transactions on Graphics* 23(3), pp. 689–694. doi: 10.1145/1015706.1015780.

Panetta, K., Gao, C. and Agaian, S. 2016. Human-Visual-System-Inspired Underwater Image Quality Measures. *IEEE Journal of Oceanic Engineering* 41(3), pp. 541–551. doi: 10.1109/JOE.2015.2469915.

Su, J.-W., Chu, H.-K. and Huang, J.-B. 2020. Instance-Aware Image Colorization., pp. 7968–7977. Available at: https://openaccess.thecvf.com/content_CVPR_2020/html/Su_Instance-Aware_Image_Colorization_CVPR_2020_paper.html [Accessed: 23 October 2022].

Vitoria, P., Raad, L. and Ballester, C. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*., pp. 2434–2443. doi: 10.1109/WACV45572.2020.9093389.

Yang, M. and Sowmya, A. 2015. An Underwater Color Image Quality Evaluation Metric. *IEEE Transactions on Image Processing* 24(12), pp. 6062–6071. doi: 10.1109/TIP.2015.2491020.

Žeger, I., Grgic, S., Vuković, J. and Šišul, G. 2021. Grayscale Image Colorization Methods: Overview and Evaluation. *IEEE Access* 9, pp. 113326–113346. doi: 10.1109/ACCESS.2021.3104515.

Zhang, R., Isola, P. and Efros, A.A. 2016. Colorful Image Colorization. In: Leibe, B., Matas, J., Sebe, N., and Welling, M. eds. *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 649–666. doi: 10.1007/978-3-319-46487-9_40.

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A.S., Yu, T. and Efros, A.A. 2017. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics* 36(4), p. 119:1-119:11. doi: 10.1145/3072959.3073703.

Alexander, R.A. 1990. A note on averaging correlations. *Bulletin of the Psychonomic Society* 28(4), pp. 335–336. doi: 10.3758/BF03334037.

Johnson, D.R. and Creech, J.C. 1983. Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error. *American Sociological Review* 48(3), pp. 398–407. doi: 10.2307/2095231.

Dai, J., He, K. and Sun, J. 2016. *Instance-Aware Semantic Segmentation via Multi-task Network Cascades*., p. 3158. doi: 10.1109/CVPR.2016.343.

Girshick, R. 2015. Fast R-CNN. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, pp. 1440–1448. Available at: http://ieeexplore.ieee.org/document/7410526/ [Accessed: 23 October 2022].

Girshick, R., Donahue, J., Darrell, T. and Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*., pp. 580–587. doi: 10.1109/CVPR.2014.81.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. 2017. Mask R-CNN. In: *2017 IEEE International Conference on Computer Vision (ICCV)*., pp. 2980–2988. doi: 10.1109/ICCV.2017.322.

Huang, Z., Huang, L., Gong, Y., Huang, C. and Wang, X. 2019. Mask Scoring R-CNN. IEEE Computer Society, pp. 6402–6411. Available at: https://www.computer.org/csdl/proceedings-article/cvpr/2019/329300g402/1gyrVSGQAXC [Accessed: 23 October 2022].

Lee, Y. and Park, J. 2020. *CenterMask: Real-Time Anchor-Free Instance Segmentation*., p. 13912. doi: 10.1109/CVPR42600.2020.01392.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. 2017. *Feature Pyramid Networks for Object Detection*., p. 944. doi: 10.1109/CVPR.2017.106.

Qiao, S., Sun, Y. and Zhang, H. 2020. Deep Learning Based Electric Pylon Detection in Remote Sensing Images. *Remote Sensing* 12, p. 1857. doi: 10.3390/rs12111857.

Ren, S., He, K., Girshick, R. and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6), pp. 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

Yagüe, F., Diez-Pastor, J., Latorre-Carmona, P. and García-Osorio, C. 2022. *Defect detection and segmentation in X-Ray images of magnesium alloy castings using the Detectron2 framework*.

Berns, R. and Reiman, D. 2002. Color managing the 3rd edition of Billmeyer and Saltzman's Principles of Color Technology. *Color Research & Application* 27, pp. 360–373. doi: 10.1002/col.10083.

Defining and Communicating Color The CIELAB System 2013. - References - Scientific Research Publishing. [no date]. Available at: https://www.scirp.org/%28S%28lz5mqp453edsnp55rrgjct55%29%29/reference/referencespapers.aspx?referenceid=2443075 [Accessed: 23 October 2022].

Image color cast detection algorithm, fast speed, good effect, shared with everyone. _Invincible triangle cat's blog - CSDN blog _ color cast detection. [no date]. Available at: https://blog.csdn.net/fightingforcv/article/details/52724848 [Accessed: 24 October 2022].

Kamble, V. and Bhurchandi, K. 2015. No-reference image quality assessment algorithms: A survey. *Optik - International Journal for Light and Electron Optics* 126, pp. 1090–1097. doi: 10.1016/j.ijleo.2015.02.093.

Li, P., Yu, H., Li, S. and Xu, P. 2021. Comparative Study of Human Skin Detection Using Object Detection Based on Transfer Learning. *Applied Artificial Intelligence* 35(15), pp. 2370–2388. doi: 10.1080/08839514.2021.1997215.

Ly, B., Dyer, E., Feig, J., Chien, A. and Bino, S. 2020. Research Techniques Made Simple: Cutaneous Colorimetry: A Reliable Technique for Objective Skin Color Measurement. *The Journal of investigative dermatology* 140, pp. 3-12.e1. doi: 10.1016/j.jid.2019.11.003.

Mandal, M. 2021. CNN for Deep Learning | Convolutional Neural Networks. Available at: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/ [Accessed: 24 October 2022].

Niveditta, T. and Swapna, D. 2011. A new Method for Color Image Quality Assessment. *International Journal of Computer Applications* 15. doi: 10.5120/1921-2565.

Panoptic Segmentation. [no date]. Available at: https://openreview.net/forum?id=RRruz0LqpMT [Accessed: 24 October 2022].

Preiss, J. 2015. *Color-Image Quality Assessment: From Metric to Application*. Ph.D. Thesis, Darmstadt: Technische Universität. Available at: https://tuprints.ulb.tu-darmstadt.de/4389/ [Accessed: 23 October 2022].

Reinhard, E., Khan, E., Akyüz, A. and Johnson, G. 2008. Color Imaging Fundamentals and Applications. Available at: https://avesis.metu.edu.tr/yayin/85890f29-dab5-4916-88b9-3c661ed1de60/color-imaging-fundamentals-and-applications [Accessed: 23 October 2022].

Teng, X., Li, Z., Liu, Q., Pointer, M.R., Huang, Z. and Sun, H. 2021. Subjective evaluation of colourized images with different colorization models. *Color Research & Application* 46(2), pp. 319–331. doi: 10.1002/col.22593.

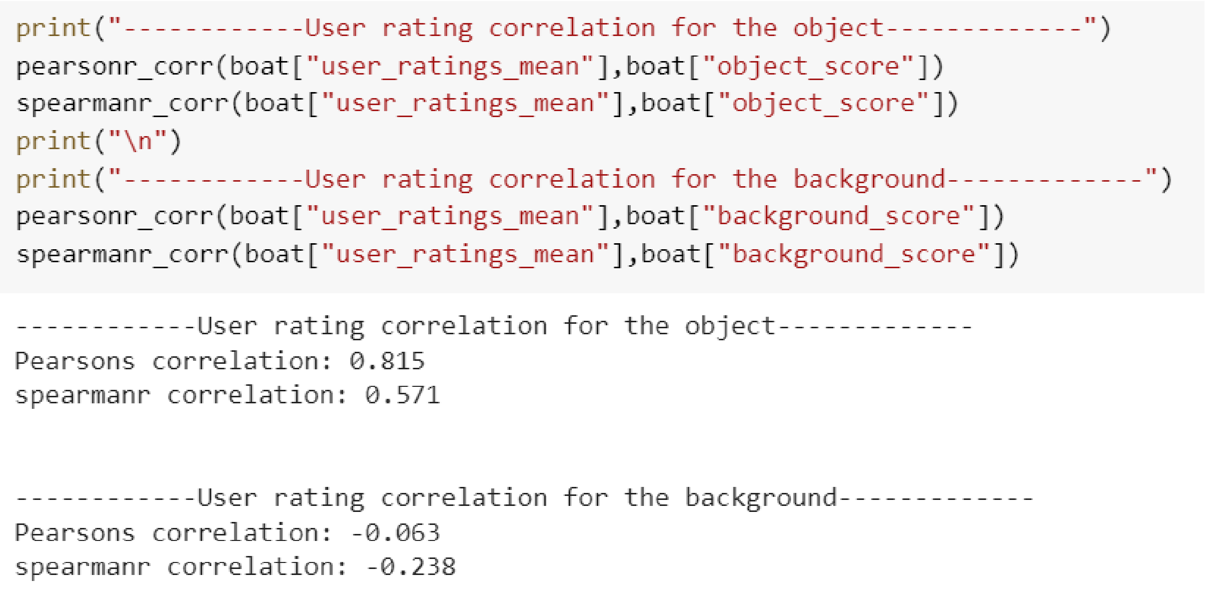
Thung, K. and Raveendran, P. 2010. A survey of image quality measures., pp. 1–4. doi: 10.1109/TECHPOS.2009.5412098.

Yamashita, R., Nishio, M., Do, R.K.G. and Togashi, K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging* 9(4), pp. 611–629. doi: 10.1007/s13244-018-0639-9.

Zhao, D., Wang, A. and Russakovsky, O. 2021. *Understanding and Evaluating Racial Biases in Image Captioning*., p. 14820. doi: 10.1109/ICCV48922.2021.01456.

7. Appendix 1

The correlation for the background score and the user score is calculated with the same correlation coefficient method to check whether the user consider much about the background.



```
print("------------User rating correlation for the object-------------")
pearsonr_corr(boat["user_ratings_mean"],boat["object_score"])
spearmanr_corr(boat["user_ratings_mean"],boat["object_score"])
print("\n")
print("------------User rating correlation for the background-------------")
pearsonr_corr(boat["user_ratings_mean"],boat["background_score"])
spearmanr_corr(boat["user_ratings_mean"],boat["background_score"])
```

```
------------User rating correlation for the object-------------
Pearsons correlation: 0.815
spearmanr correlation: 0.571


------------User rating correlation for the background-------------
Pearsons correlation: -0.063
spearmanr correlation: -0.238
```