NAME YOUR COLOUR FOR THE TASK: ARTIFICIALLY DISCOVER COLOUR NAMING VIA COLOUR QUANTISATION TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

The long-standing theory that a colour-naming system evolves under the dual pressure of efficient communication and perceptual mechanism is supported by more and more linguistic studies including the analysis of four decades' diachronic data from the Nafaanra language. This inspires us to explore whether artificial intelligence could evolve and discover a similar colour-naming system via optimising the communication efficiency represented by high-level recognition performance. Here, we propose a novel colour quantisation transformer, CQFormer, that quantises colour space while maintaining the accuracy of machine recognition on the quantised image. Given an RGB image, the annotation branch maps it into an index map before generating the quantised image with a colour palette, meanwhile the palette branch utilises a key-point detection way to find proper colours in palette among whole colour space. By interacting with colour annotation, CQFormer is able to balance both the machine vision accuracy and colour perceptual structure such as distinct and stable colour distribution for discovered colour system. Very interestingly, we even observe the consistent evolution pattern between our artificial colour system and basic colour terms across human languages. Besides, our colour quantisation method also offers an efficient quantisation method that effectively compresses the image storage while maintaining a high performance in high-level recognition tasks such as classification and detection. Extensive experiments demonstrate the superior performance of our method with extremely low bit-rate colours. We will release the source code upon acceptance.

1 INTRODUCTION

Hath not a Jew eyes?
Hath not a Jew hands, organs,dimensions, senses, affections, passions?

William Shakespeare "The Merchant of Venice"

Does artificial intelligence share the same perceptual mechanism for colours as human beings? We aim to explore this intriguing problem through AI simulation in this paper.

Colour involves the visual reception and neural registering of light stimulants when the spectrum of light interacts with cones cells in the eyes. Physical specifications of colour also include the reflective properties of the physical objects, geometry incident illumination *etc*. By defining a colour space (Forsyth & Ponce, 2002), people could identify directly colours according to these quantifiable coordinates.

Compared to the pure physiological nature of hue categorisation, the complex phenomenon of colour naming or colour categorisation spans multiple disciplines from cognitive science to anthropology. Solid diachronic research (Berlin & Kay, 1969) also suggests that human languages are constantly evolving to acquire new colour names, resulting in an increasingly fine-grained colour naming system. This evolutionary process is hypothesised to be under the pressure of both communication efficiency and perceptual structure. Communication efficiency requires shared colour partitioning to be communicated accurately with a lexicon as simple and economical as possible. Colour perceptual structure is relevant to human perceptual in nature. For example, the colour space distance

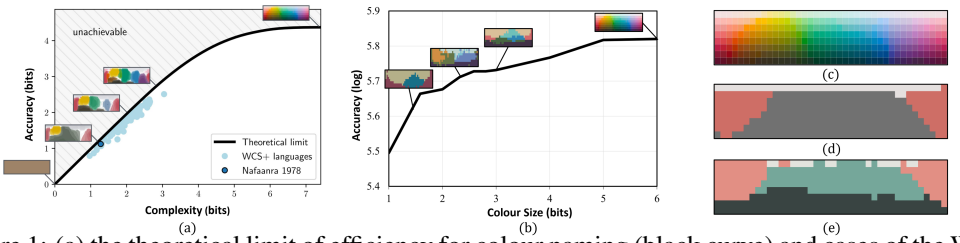


Figure 1: (a) the theoretical limit of efficiency for colour naming (black curve) and cases of the WCS probability map of human colour language copied from (Zaslavsky et al., 2022). (b) the colour size (from 1-bit to 6-bit)-accuracy curve on the tiny-imagenet-200 (Le & Yang, 2015) dataset. The WCS probability maps generated by our CQFormer are also shown along the curve. (c) the colour naming stimulus grid used in the World Colour Survey (WCS) (Kay et al., 2009). (d) the three-term WCS probability map of CQFormer after embedding 1978 Nafaanra three colour system ((light ('fiNge'), dark ('wOO'), and warm or red-like ('nyiE')) into the latent representation. (e) the four-term WCS probability map of CQFormer evolved from (d). The evolved fourth colour, yellow-green, is consistent with the prediction of basic colour term theory (Berlin & Kay, 1969)

between nearby colours should correspond to their perceptual dissimilarity. This structure of perceptual colour space has long be used to explain colour naming patterns across languages. A recent analysis of human colour naming systems, especially in Nafaara, contributes the very direct evidence to support this hypothesis through the employment of Shannon's communication model (Shannon, 1948). Very interestingly, this echos with the research of colour quantisation, which quantises colour space to to reduce the number of distinct colours in an image.

Traditional colour quantisation methods (Heckbert, 1982; Gervautz & Purgathofer, 1988; Floyd & Steinberg, 1976) are *perception-centred* that generates a new image which is as visual perceptually similar as possible to the original image. These methods group similar colours in the colour space and represent each group with a new colour, thus naturally preserves the perceptual structure. Instead of prioritising the perceptual quality, Hou *et al.* (Hou et al., 2020) proposed a *task-centred/machine-centred* colour quantisation method, ColorCNN, focusing on maintaining network classification accuracy in the restricted colour spaces. While achieving an impressive classification accuracy on even few bits image, ColorCNN only identifies and preserves *machine-centred* structure, without directly clustering similar colours in the colour space. Therefore, this pure *machine-centred* strategy sacrifices perceptual structure and often associates nearby colours to different quantised index.

Zaslavsky (Zaslavsky et al., 2022) measures the communication efficiency in colour naming by analysing the informational complexity based on the information bottleneck (IB) principle. Here, we argue that the network recognition accuracy also reflects the communication efficiency when the number of colours is restricted. Since the human colour naming is shaped by both perception structure and communication efficiency (Zaslavsky et al., 2019a), we integrate the need for both *perception* and *machine* to propose a novel end-to-end colour quantisation transformer, CQFormer, to discover the artificial colour naming systems.

As illustrated in Fig. 1.b, the recognition accuracy increases with the number of colours in our discovered colour naming system. Surprisingly, the complexity-accuracy trade offs are similar to the numerical results (Fig. 1.a) independently derived from linguistic research (Zaslavsky et al., 2022). What is more, after embedding 1978 Nafaanra three colour system (Fig. 1.d) into the latent representation of CQFormer, our method automatically evolves the fourth colour closed to yellow-green, matching the basic colour terms theory (Berlin & Kay, 1969) summarised from different languages. Berlin and Kay found universal restrictions on colour naming across cultures and claimed languages acquire new basic colour category in a strict chronological sequence. For example, for a culture has three colour (light ('fiNge'), dark ('wOO'), and warm or red-like ('nyiE') in Nafaanra), the four colour it evolves should be yellow or green, exactly the one (Fig. 1.e) discovered by our CQFormer.

The pipeline of CQFormer, shown in Fig. 2, comprises of two modules: annotation branch and palette branch. The annotation branch annotates each pixel of the input RGB image with the proper quantised colour index before painting the index map with the corresponding colour in colour palette. We localise the colour palette in the whole colour space with a novel palette branch which detects the key-point with attention enquiry of transformer. During the training stage, as illustrated

in the brown dotted line of Fig. 2, the palette branch interacts with input RGB and reference palette query to maintain the perceptual structure by reducing the perceptual structure loss. This *perception centred* design groups similar colours and ensure colour palette sufficiently represents the colour naming system defined by World Color Survey (WCS) colour naming stimulus grids. As shown in Fig. 2, each item in colour palette (noted by asteroid) lies in the middle of corresponding colour distribution in WCS colour naming probability map. Finally, the quantised image is passed to a high level recognition module for *machine accuracy* a tasks such as classification and detection. Through the joint optimisation of CQFormer and consequent high level module, we are able to balance both $perception$ and $machine$. Apart from automatically discovering the colour naming system, our CQFormer also offers an effective solution to extremely compresses the image storage while maintaining a high performance in high-level recognition tasks.

Our contributions could be summarised as following:

- We propose a novel end-to-end colour quantisation transformer, CQFormer, to artificially discover colour naming system by considering the need of both perception and machine. The discovered colour naming system shows similar pattern as human language on colour.
- We propose a novel colour palette generation method, which take the colour quantise as an attention based key-point detection task, input independent attention queries to generate 3D coordinates, then automatically find the selected colour in whole colour space.
- Our colour quantisation achieves superior performance on both classification and object detection with extremely low bit-rate colours. The source code will be release to facilitate the community.

2 Related Works

2.1 Colour Quantisation Methods

Traditional colour quantisation (Orchard & Bouman, 1991; Deng et al., 1999; Liang et al., 2003; Wu, 1992) reduces the colour space while maintaining visual fidelity. These *perception-centred* methods usually cluster colours to create a new image as visual perceptually similar as possible to the original image. For example, MedianCut Heckbert and OCTree, Gervautz & Purgathofer are two representative clustering based algorithms. Dithering (Floyd & Steinberg, 1975) eliminates visual artifacts by including a noise pattern. The colour quantised images can be expressed as indexed colour (Poynton, 2012), and encoded with PNG (Boutell, 1997). Recently, Hou et al. and propose a pure *machine-centred* CNN-based colour quantisation method, ColorCNN, which effectively maintains the informative structures under an extremely constrained colour space. In addition to colour quantisation, Camposeco et al. also design a task-centred image compression method for localisation in 3D map data. Since human colour naming reflects both perceptual structure and communicative need, our CQFormer also considers both *perception* and *machine* to artificially discover the colour naming system.

2.2 World Color Survey (WCS)

The World Color Survey (WCS: Kay et al. (2009)) comprises colour name data from 110 languages of non-industrialised societies (Zaslavsky et al., 2019b), with respect to the stimulus grid shown in Fig. 1(c). There are 320 colour chips in colour naming stimulus grids and each chip is at its maximum saturation for that hue/lightness combination, while columns correspond to equally spaced Munsell hues and rows to equally spaced Munsell values. Participants were asked to name the colour of each chip to record the colour naming system.

3 Methodology

3.1 Problem Formulation

For an image-label pair $(\boldsymbol{x}, y)$ in dataset $\mathcal{D}$, the recognition network $f_\theta(\cdot)$ encodes input image $\boldsymbol{x}$ to predict its label $\hat{y}$ (*i.e.* class probability in classification task). $f_\theta(\cdot)$ can be optimised by minimising

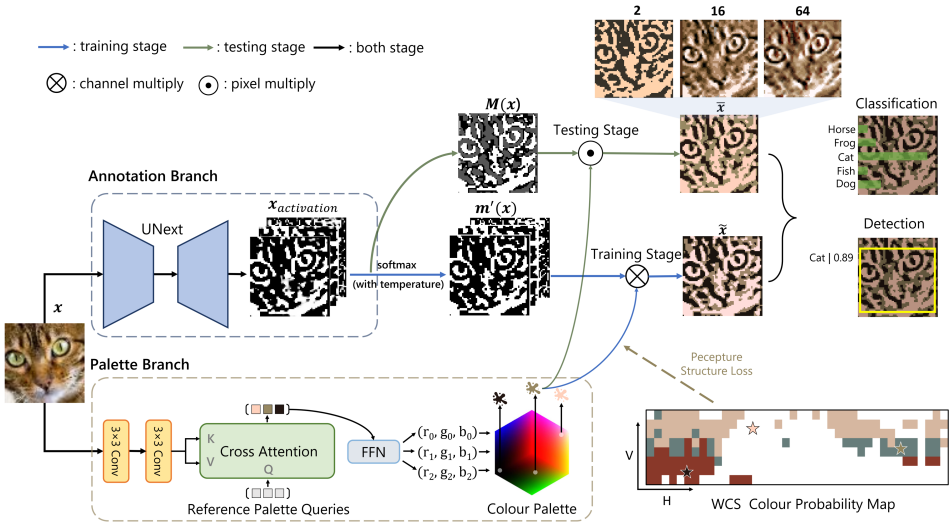


Figure 2: Detailed structure of CQFormer. Blue lines are only for training phase, green lines are only for testing phase, black lines are both used in training and testing phase.

loss $\mathcal{L}$ between prediction $\hat{y} = f_\theta(x)$ and ground truth y, to find the appropriate parameter $\theta^\star$:

$$\theta^\star = \arg\min_{\theta} \sum_{(\boldsymbol{x},y)\in D} \mathcal{L}\left(y, f_\theta(\boldsymbol{x})\right). \tag{1}$$

We aim to discover the artificial colour naming system by meeting the need for both machine accuracy and human perception. Therefore, CQFormer not only strives for recognition accuracy on the colour quantised images, but also maintains the perceptual structure. The function in Eq. 1 can be rewritten as follow:

$$\psi^\star, \theta^\star = \arg\min_{\psi,\theta} \sum_{(\boldsymbol{x},y)\in D} \mathcal{L}\left(y, f_\theta\left(g_\psi(\boldsymbol{x})\right)\right) + \gamma R, \tag{2}$$

where ψ, θ respectively denotes the parameter of CQFormer g and the following high level recognition network f, which we jointly optimise g and f to find optimal parameters $\psi^\star$ and $\theta^\star$. R is the regularisation terms and γ is a non-negative parameter to balance loss weights.

3.2 CQFormer Architecture

As illustrated in Fig. 2, our CQFormer mainly consists of two branches: an annotation branch and a palette branch. Annotation branch annotates each pixel of the input RGB image with the proper quantised colour index. Given the input RGB image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$, the annotation branch generates a softmax (with a temperature parameter) probability map $m^{'}(\boldsymbol{x}) \in \mathbb{R}^{H\times W\times C}$ during training stage or a colour index map $M(\boldsymbol{x}) \in \mathbb{R}^{H\times W}$ during testing time. H and W denotes height and width of input image $\boldsymbol{x}$ while C represents the number of quantised colours. Softmax probability map represents the possibility of individual pixel in input image that belongs to certain quantised colour c.

Here, we also define WCS colour probability map on the colour naming stimulus grids in the world colour survey (WCS). Each grid is associated with a colour index and the corresponding probability. The map for human language is collected by WCS and the machine one is counted across the training or testing dataset.

We also propose a palette branch to acquire quantised colour palette. Here we define a explicit reference palette queries $Q \in \mathbb{R}^{C\times d}$ consisting of C learnable vectors of dimension d. Each vector represents an automatically mined colour. The reference palette queries Q interacts with input RGB to generate the final colour palette $P(\boldsymbol{x}) \in \mathbb{R}^{C\times 3}$ that comprises of C triples of (R, G, B). Each triple represents C colours discovered by machine.

CQFormer then outputs quantised image by painting the colour index map $M(\boldsymbol{x})$ with colour palette $P(\boldsymbol{x})$. Finally, we feed the quantised image into the high-level recognition module for recognition tasks, *i.e.* classification, object detection.

Annotation Branch: As illustrated in Fig. 2, the first component is an UNeXt (Valanarasu & Patel, 2022) encoder which outputs per-pixel category. Given the input RGB image $\boldsymbol{x}$, UNeXt encoder would generates a class activation map $\boldsymbol{x}_{activation} \in \mathbb{R}^{H\times W\times C}$. Pass by a softmax function, the probability map $m^{'}(\boldsymbol{x}) \in \mathbb{R}^{H\times W\times C}$ would generated from $\boldsymbol{x}_{activation}$. During testing procedure, we could get the quantised colour index map $M(\boldsymbol{x}) \in \mathbb{R}^{H\times W}$ from the indices of the maximum probability over the C-channel probability map $m^{'}(\boldsymbol{x})$, we then utilise colour index map $M(\boldsymbol{x})$ and the colour palette $P(\boldsymbol{x})$ to generate the test-time colour quantised image $\bar{\boldsymbol{x}}$ via:

$$\bar{\boldsymbol{x}} = \sum_{c} [P(\boldsymbol{x})]_c \cdot \mathbb{I}(M(\boldsymbol{x}) = c), \tag{3}$$

where c is the colour index , $[P(\boldsymbol{x})]_c \in \mathbb{R}^3$ represents the c-th RGB value of the colour palette, $\mathbb{I}(\cdot)$ is an indicator function taking the value 1 if its argument is true and 0 else, $\cdot$ denotes the point-wise multiplication.

During training stage, since $\arg\max$ function is not differentiable, here we utilise the C-channel softmax probability map $m^{'}(\boldsymbol{x})$ instead of the colour index map $M(\boldsymbol{x})$ generated by $\arg\max$ function. To avoid over-fitting, we add a temperature parameter τ (He et al., 2018) to the original $\mathrm{softmax}$ function, forcing the softmax probability map distribution closer to an one-hot vector, the train-time softmax probability map $m^{'}(\boldsymbol{x})$ is computed as:

$$m^{'}(\boldsymbol{x}) \quad = \quad \mathrm{softmax}(\frac{\boldsymbol{x}_{activation}}{\tau}), \tag{4}$$

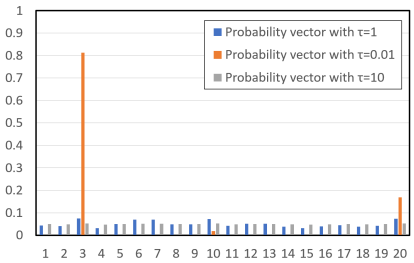


Figure 3: The influence of the temperature parameter τ on Softmax function.

where $\tau > 0$ denotes the temperature parameter, which plays a significant role in the final probabilities distribution. As shown in Fig. 3, when in the low temperature ($\tau < 1$), the output is similar to a one-hot vector with large diversity. With high temperature($\tau > 1$), the output distribution is similar to a uniform distribution with small diversity (He et al., 2018). Therefore, we fix the $0 < \tau < 1$ to approximate the $\arg\max$ function during the training procedure. Finally, we change the table look-up operation from the original forward pass into a weighted sum of the colour palette $P(\boldsymbol{x})$. For quantised colour with index c, we use $[m^{'}(\boldsymbol{x})]_c$ (c colour of $m^{'}(\boldsymbol{x})$ selected by palette $P(\boldsymbol{x})$) as the intensity of expression over the whole image. The train-time colour quantised image $\widetilde{\boldsymbol{x}}$ is generated as:

$$\widetilde{\boldsymbol{x}} \quad = \quad \sum_{c} [P(\boldsymbol{x})]_c \cdot [m^{'}(\boldsymbol{x})]_c. \tag{5}$$

Palette Branch: The palette branch is proposed to generate the suitable colour palette that benefits the high-level recognition tasks and maintains perceptual structure. Here, we locates the representative colour by attention based key-point detection strategy, which is originally designed to utilise transformer queries to automatically find key-point location (*i.e.* bounding-box (Carion et al., 2020), human pose (Li et al., 2021)) by attention mechanism. Specifically, we reformulate the problem of colour quantisation as a 3D spatial key-point localisation task among whole colour space.

Given the input image $\boldsymbol{x}$, two stacked convolutions would encode $\boldsymbol{x}$ into a high-dimensional feature with lower resolution $\boldsymbol{x}_{feat} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times d}$, where d is the dimension of the encoded feature $\boldsymbol{x}_{feat}$. Since the attention block expects sequence as input, we collapse the spatial dimensions of $\boldsymbol{x}_{feat}$ into one dimension, resulting in a $\boldsymbol{x}^{'}_{feat} \in (\frac{HW}{16}) \times d$ feature map. After that, the keys $K \in \mathbb{R}^{N\times d}$ and values $V \in \mathbb{R}^{N\times d}$ would generated from the encoded features $\boldsymbol{x}^{'}_{feat}$, here we use depth-wise convolution (Han et al., 2021) to complete position encoding, which is suitable for different input resolutions.

We define an explicit colour palette queries called reference palette queries Q to represent each automatically mined colour. As shown in Fig. 1, the reference palette queries Q is a series of learnable embeddings initialised with Gaussian distribution, which would sent into attention block to make attention multiply with K and V, different from DETR (Carion et al., 2020), we utilise cross attention between the Q,K and V instead of multi-head self-attention. Finally the attention block's outputs would pass by a FFN (feed forward network, which consist of two linear layers) to get the

final 3D coordinates in whole colour space like (r_1, g_1, b_1). Then the generated 3D coordinates would mapping back to corresponding colours in colour space, thus generate colour palette $P(\boldsymbol{x})$.

High-Level Recognition Module: High-level recognition module takes generated $\bar{\boldsymbol{x}}$ or $\widetilde{\boldsymbol{x}}$ as inputs, we adopt the popular Resnet (He et al., 2016) as the classifier and the Faster-RCNN (Ren et al., 2015) as the detector in our experiments.

3.3 Perceptual Structure Loss

To balance the machine accuracy and perceptual structure, we design the perceptual structure loss, including perception regularisation, purity regularisation and perceptually similarity regularisation. CQFormer only minimises this loss in the training, forcing the method to discover the proper colour naming system.

Perception Regularisation: The perception regularisation term $R_{\text{perception}}$ is a mean squared error (MSE) between the quantised image $\widetilde{\boldsymbol{x}}$ and the origin image $\boldsymbol{x}$. We adopt $R_{perception}$ to ensure the colour palette $P(\boldsymbol{x})$ lies in the centre of corresponding colour distribution in the WCS colour naming probability map. Hence, the colour palette generated by palette branch has a solid connection with the softmax probability map generated by annotation branch, which ensures the colour quantised image generated by CQFormer has high perceptual quality and visual similarity to the original image. The $R_{\text{perception}}$ is formulated as:

$$R_{\text{perception}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}[i] - \widetilde{\boldsymbol{x}}[i])^2 \tag{6}$$

where $N = H \times W \times 3$ represents the number of all values on the image.

Purity Regularisation: In order to further approximate the $\arg\max$ function during training, we utilise the purity regularisation term R_{purity}. Purity is a metric that forces each pixel to belong to only one colour. During training, the purity regularisation term R_{purity} enables each pixel's probability distribution of the softmax probability map $m(\boldsymbol{x})$ to be more like a one-hot vector (Hou et al., 2020). The R_{purity} is calculated as:

$$R_{\text{purity}} = \log_2 C \times \left(1 - \frac{1}{C} \times \sum_{c} \max_{(u,v)}[m(\boldsymbol{x})]_{u,v}\right) \tag{7}$$

Perceptually Similarity Regularisation: After colour quantisation, the pixels in the origin image belongs to C colours. Now the pixel colours of each new colour could form a cluster and covers parts of WCS colour probability map. We propose the perceptually similarity regularisation term $R_{\text{similarity}}$ to minimise the intra-cluster Munsell HSV space distance of each colour cluster. In other words, we make the colours covered by each new colour term to be as perceptually similar as possible. We at first calculate the centroid colour value $centroid_c$ of each colour cluster in the Munsell HSV colour space. Then, the average squared colour distance of all pixels in colour cluster c to the centroid d_c is computed in the conical representation of the HSV colour space (see in Eq. 8). Finally, we calculate the mean value of squared intra-cluster distance for each colour cluster and obtain the perceptually similarity regularisation term $R_{\text{similarity}}$.

$$\begin{aligned}&\text{HSV-Distance}(\boldsymbol{x}_{HSV,1}, \boldsymbol{x}_{HSV,2}) = \text{HSV-Distance}\left([h_1, s_1, v_1], [h_2, s_2, v_2]\right)\\ &= \sqrt{(v_2 - v_1)^2 + s_1^2 v_1^2 + s_2^2 v_2^2 - 2 s_1 s_2 v_1 v_2 \cos\left(h_2 - h_1\right)}\end{aligned} \tag{8}$$

Therefore, we re-define the loss function $\mathcal{L}$ as :

$$\mathcal{L} = \mathcal{L}_{High} + \alpha R_{\text{Similarity}} + \beta R_{\text{purity}} + \gamma R_{\text{perception}} \tag{9}$$

where $\mathcal{L}_H$ is the high-level recognition loss (*e.g.* cross entropy loss for classification task), $R_{\text{similarity}}$, R_{purity} and $R_{\text{perception}}$ represent the regularisation term of perceptually similarity, purity and the perception, respectively. The α , β and γ denote their weight.

3.4 Embedding Colour Naming of Human Language

To exploit the human colour naming, we aim to embed the colour naming system of human language into the latent representation of our CQFormer. Here, we propose two kinds of embedding solutions: colour probability map embedding and central colour embedding.

Colour Probability Map Embedding: This embedding forces our CQFormer to have the identical WCS colour probability map with certain language. At first, we train CQFormer to have the same colour number C of designed human colour system. As illustrate in Fig. 4, for each pixel in the RGB input image (Fig. 4(a)), we could find its corresponding coordinate at the WCS colour probability map of certain human language (Fig. 4(c)) according to its H and V value. After knowing the human colour term c for this pixel, we assign this probability to the same image position of c channel softmax probability map. After performing above operations on all pixels of input image, we could have C channels human language softmax probability maps. Then we force the C channels softmax of CQFormer to match the human language softmax probability maps with a cross-entropy loss.

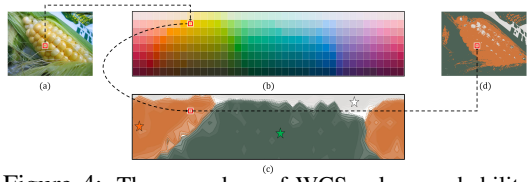


Figure 4: The procedure of WCS colour probability map embedding. (a) the origin RGB image. (b) the WCS colour naming stimulus grids. (c) the WCS colour probability map of 1978 Nafaanra three colour system. (d) the colour index map for Nafaanra-1978.

Colour Evolution: Our CQFormer could evolve to acquire fine-grained new colour. For example, we at first train a CQFormer with three colour setting as teacher model. When training a CQFormer with four colour setting, we match its first three channels of softmax colour probability map with the softmax probability map of teacher model. In this knowledge distillation way, we are able to make CQFormer to inherit the colour naming system of less colour number model. Similarly, using above colour probability map embedding, we could even evolve from existing human colour system as detailed in Sec.4.5. Specifically, the four colour naming system shown inf Fig.1.e is evolved from the three colour in Nafaanra language. Very interestingly, the evolved fourth colour is exactly green-yellow predicted by basic colour term theory (Berlin & Kay, 1969).

Central Colour Embedding: The centroid embedding only embeds the central colour (noted by asteroid) of WCS colour probability map into the latent space of CQFormer. At first, we calculate the H and V value of centroid coordinates of each colour in the human colour system. Then we replace the centroid centre $centroid_c$ of Perceptually Similarity Regularisation with aforementioned central human colours. Specifically, we obtain the artificial colour probability map shown in Fig. 1(b) by embedding the central colour shown in Fig. 1(b).

4 Experiments

We evaluate our CQFormer model on mainstream benchmark datasets in both image classification (Sec. 4.2) task and object detection task (Sec. 4.3). Additionally, we specifically design a colour evolution experiment (Sec. 4.5) to demonstrate how the CQFormer automatically evolves increasing fine-grained colours. For the visualisation of the quantised image and the ablation study, please also refer to Appendix 6.

4.1 Experiment Setup

Datasets: We use various datasets to test our CQFormer. For classification task, we utilise CIFAR10, CIFAR100 (Krizhevsky et al., 2009), STL10 (Coates et al., 2011) and tiny-imagenet-200 (Tiny200) (Le & Yang, 2015) dataset. Both CIFAR10 and CIFAR100 contain 60000 images, covering 10 and 100 classes of general objects respectively. The STL10 consists of 13000 images (5000 for training and 8000 for testing) with 10 classes. The Tiny200 is a subset of ImageNet (Deng et al., 2009) dataset which contains 200 categories with 110000 images. All train images are random cropped, random horizontal flipped and resized into their respective original resolutions while all test images are resized into their respective original resolutions. For object detection task, we utilise MS COCO (Lin et al., 2014) dataset for detection evaluation. MS COCO contains $\sim$118k images with bounding box annotation in 80 categories. Here, we use COCO `train2017` set as train set and use COCO `val2017` set as test set.

Evaluation Metrics: For classification, we report top-1 classification accuracy and top-1 classification accuracy as the evaluation metric. For object detection, we report average precision (AP) value in COCO evaluation setting (see Table. 1).

Implement Details: *(i) Upper bound:* We utilise the performance of classifier/detector without additional colour quantisation methods in full colour space (24 bit) as the upper bound. For clas-

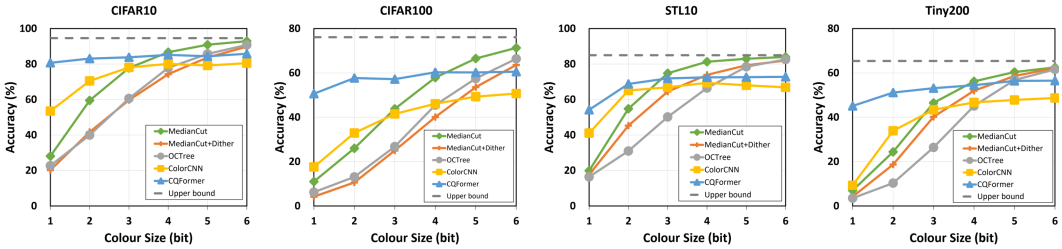


Figure 5: Top-1 classification accuracy of colour quantised images on four datasets with Resnet-18 (He et al., 2016) networks. The dark blue curve represents our CQFormer's performance under low-bit colour space. The grey dotted line denotes the upper bound on each dataset.

sification upper bound, we adopt Resnet-18 (He et al., 2016) network. For detection upper bound, we adopt Faster-RCNN (Ren et al., 2015) network with ResNet-50 (He et al., 2016) backbone and FPN (Lin et al., 2017) neck.

(ii)Training Settings: For classification, we cascade CQFormer with classifier and jointly train from scratch for 60 epochs on single GeForce RTX A6000 GPU, we set the temperature parameter $\tau = 0.01$ and the regularisation term $\alpha = 1$, $\beta = 0.3$, $\gamma = 1$, batch size is set to 128 (STL10 is set to 32), we adopt SGD optimiser with initial learning rate 0.1, momentum is 0.5 and weight decay is 0.001. Additionally, we utilise the Cosine-Warm-Restart (Loshchilov & Hutter, 2016) as the learning rate scheduler with a peak learning rate at 0.05, a minimal learning rate of 0. For detection, we train the model for 12 epochs on 4 Tesla V100 GPUs, with regularisation term $\alpha = 0$, $\beta = 1$, $\gamma = 0.1$, batch size is set to 8, we also adopt SGD optimiser with initial learning rate 0.01, momentum 0.9 and weight decay is 0.0001, learning rate decay to one-tenth at 8 and 11 epoch.

Comparison Methods: As illustrated in Fig. 5, we compare with three traditional perception-centred methods: MedianCut (green solid line) (Heckbert, 1982) , MedianCut+Dither (orange solid line) (Floyd & Steinberg, 1976) and OCTree (gray solid line) (Gervautz & Purgathofer, 1988), and a task-centred CNN-based method ColorCNN (yellow solid line) (Hou et al., 2020). The gray dotted line is the upper bound and the dark blue solid line represents the performance of our CQFormer.

4.2 Classification Task Evaluation

Fig. 5 presents comparisons to the state-of-the-art methods on four datasets. Our proposed CQFormer (blue curve) has a consistent and obvious improvement over all of the other methods in extremely low-bit colour space (less than 3-bit). Moreover, our CQFormer archives a superior performance than the task-centred method ColorCNN (Hou et al., 2020) under all colour quantisation level from 1-bit to 6-bit. Taking the 1-bit quantisation level as an example, our CQFormer reaches 80.7%, 50.6%, 60.2% and 45.1% accuracy on the four datasets, respectively. Although the CIFAR100 and Tiny200 dataset have more categories and higher classification difficulty, our CQFormer still maintain the advantageous performance. As the number of the quantised colours increases, the accuracy of the CQFormer continues to improve, which is similar to the complexity-accuracy trade offs (Fig. 1(a)) independently derived from linguistic research (Zaslavsky et al., 2022).

4.3 Detection Task Evaluation

Table. 1 shows the object detection results on COCO dataset, we report the average precision value (AP) of CQFormer and the comparison methods, we could see that our CQFormer also gain superior performance in object detection task, especially gain best performance in most of bit number.

4.4 Performance on Larger Colour Space

Similar to task centred colorCNN (Hou et al., 2020), we are also inferior to the traditional method under a large space. As extensively discussed in (Hou et al., 2020), this is very normal since traditional methods clusters all pixels to enforce intra-cluster similarity to make a collective decision. Actually, this also explains why our method consistently outperforms pure machine-centred ColorCNN as our method considers the perceptual structure.

In addition, we are quite satisfied with the current superior performance on limited colour numbers as most of human languages only use fewer than 3-bit colour terms. This implies that discovering more colours, not only compromises the principle of efficiency, but also goes contrary to the expectation of better perceptual effects. In other words, the performance of the CQFormer on limited coluor categories may hint at the optimal outcome restricted by the unique neurological structure of human vision and cognition, which are in turn reflected in a wide array of languages.

Table 1: Object detection results on MS COCO dataset (Lin et al., 2014), we report AP value.

Method	1-bit	2-bit	3-bit	4-bit	5-bit	6-bit	Full Colour (24-bit)
Upper bound	-	-	-	-	-	-	37.2
Median Cut (w/o D) (Heckbert, 1982)	6.5	9.7	11.4	14.0	**16.8**	18.1	-
Median Cut (w D) (Floyd & Steinberg, 1976)	7.3	8.8	11.2	13.4	14.8	15.6	-
OCTree (Gervautz & Purgathofer, 1988)	8.7	9.0	9.8	10.9	13.4	14.5	-
CQFormer	**8.9**	**11.2**	**12.8**	**14.5**	16.6	**19.4**	-

4.5 Colour Evolution

Setup: The evolution of human language on CQFormer is divided into two stages: 1. embedding colour probability map stage that embeds 1978 Nafaanra three colour system into the latent representation of CQFormer and 2. colour evolution stage that evolves the fourth colour from the 1978 Nafaanra three colour system. The experiment is based on the classification task on CIFAR10 (Krizhevsky et al., 2009) dataset using CQFormer and Resnet-18 (He et al., 2016). We set $C = 4$, $\tau = 0.1$ $\alpha = 0$, $\beta = 1$, $\gamma = 0$.

In the first embedding stage, we force the last 2 colour indexes to represent the same colour without using the loss function of Eq. 9. Hence, for the initial 40 epochs, the number of the quantised colour is forced to be three and we only minimise two cross-entropy loss, including *(i)* cross-entropy between predicted classification label $f_\theta(g_\psi(\boldsymbol{x}))$ and ground truth classification label y, *(ii)* cross-entropy between colour index map $M(\boldsymbol{x})$ and the colour index map of Nafaanra-1978 $M_{\text{Nafaanra-1978}}$.

In the evolution stage, we remove the restriction on the last 2 colour indexes and minimise the loss function in Eq. 9. Hence, for the subsequent 20 epochs, the number of the quantised colour is four. Based on the three colours generated in the prior embedding stage (Fig. 1(d)), the CQFormer automatically evolves the fourth colour in the evolution stage. As illustrated in Fig. 1(e), the fourth colour automatically evolved by our method CQFormer is closed to yellow-green, matching the basic colour terms theory (Kay & McDaniel, 1978).

5 Limitation and Discussion

While the complexity-accuracy tradeoff of machine-discovered colour terms, as shown in Fig. 1.b, is quite similar to the numerical limit of categorical counterparts for human languages, the current work is still preliminary. As shown in Fig. 1, the newly discovered WCS colour probability map is still quite different from the human one. A more accurate language evolution replication needs to consider more complex variables such as environmental idiosyncrasies, cultural peculiarities, functional necessities, technological maturity, learning experience and intercultural communication.

Another promising direction would be associating the discovered colours to the human colour terms. This would involve much research on Nature Language Processing and we hope to discuss with experts from different disciplines on future works. Last but not the least, the AI simulation outcome contributes to the long-standing universalist-relativist debate of the linguistic community on colour cognition. Though not entirely excluding the cultural specificities of the colour schemes, the machine finding strongly supports the universalist view that an innate, physiological principle constrains if not determines the evolutionary sequence and distributional possibilities of basic colour terms in communities of different cultural traditions. The complexity-efficiency principle is confirmed by the finding that the numerical limitation of colour categories could lead to a superior performance on colour-specific tasks, contrary to the intuitive expectation that complexity breeds perfection. The independent AI discovery of the "green-yellow" category on basis of the fundamental tripartite "dark-light-red" colour scheme points to the congruence of neural algorithms and human cognition, and opens a new frontier to test contested hypothesis in social sciences through machine simulation. We would be more than delighted that if this tentative attempt would prove to be a bridge to link scholars of different disciplines for more collaboration and generate more fruitful results.

References

B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969. ISBN 9780520076358.

Thomas Boutell. Png (portable network graphics) specification version 1.0. Technical report, 1997.

Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7653–7662, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Yining Deng, Charles Kenney, Michael S Moore, and BS Manjunath. Peer group filtering and perceptual color image quantization. In *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 4, pp. 21–24. IEEE, 1999.

R. W. Floyd and L. Steinberg. An adaptive algorithm for spatial gray scale. *Proceedings of the Society for Information Display*, 17, 1975.

R. W. Floyd and L. Steinberg. An adaptive algorithm for spatial grayscale. *Proceedings of the Society for Information Display*, 17, 1976.

David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

Michael Gervautz and Werner Purgathofer. A simple method for color quantization: Octree quantization. In *New trends in computer graphics*, pp. 219–231. Springer, 1988.

Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yu-Lin He, Xiao-Liang Zhang, Wei Ao, and Joshua Zhexue Huang. Determining the optimal temperature parameter for softmax function in reinforcement learning. *Applied Soft Computing*, 70: 80–85, 2018. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2018.05.012. URL `https://www.sciencedirect.com/science/article/pii/S1568494618302758`.

Paul Heckbert. Color image quantization for frame buffer display. *ACM Siggraph Computer Graphics*, 16(3):297–307, 1982.

Yunzhong Hou, Liang Zheng, and Stephen Gould. Learning to structure an image with few colors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10116–10125, 2020.

Paul Kay and Chad K McDaniel. The linguistic significance of the meanings of basic color terms. *Language*, 54(3):610–646, 1978.

Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. *The world color survey*. CSLI Publications Stanford, CA, 2009.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1944–1953, 2021.

DianLong Liang, Lizhi Cheng, and Zenghui Zhang. General construction of wavelet filters via a lifting scheme and its application in image coding. *Optical Engineering*, 42(7):1949–1955, 2003.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

M.T. Orchard and C.A. Bouman. Color quantization of images. *IEEE Transactions on Signal Processing*, 39(12):2677–2690, 1991. doi: 10.1109/78.107417.

Charles Poynton. *Digital video and HD: Algorithms and Interfaces*. Elsevier, 2012.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. *arXiv preprint arXiv:2203.04967*, 2022.

Xiaolin Wu. Color quantization by dynamic programming and principal analysis. *ACM Transactions on Graphics (TOG)*, 11(4):348–372, 1992.

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11(1):207–219, 2019a.

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Communicative need in color naming. *Cognitive Neuropsychology*, 2019b. doi: 10.1080/02643294.2019.1604502.

Noga Zaslavsky, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, 2022. ISSN 2058-458X. doi: 10.1093/jole/lzac001.

6 APPENDIX

6.1 ABLATION STUDY

τ	$R_{\text{similarity}}$	R_{purity}	$R_{\text{perception}}$	Accuracy
✔	✔	✔	✔	**45.1**
✗	✔	✔	✔	6.3
✔	✗	✔	✔	43.9
✔	✔	✗	✔	43.7
✔	✔	✔	✗	44.6

Table 2: Ablation Study results under 1-bit colour quantisation, Tiny200 (Le & Yang, 2015) dataset, Renset18 (He et al., 2016) classifier.

We remove the temperature parameter ($\tau = 1.0$) or regularisation term, to see the CQFormer performance without either of those. Results are shown in Table 6.1. First, we find that removing the temperature parameter causes a severe 38.8% accuracy drop. In fact, the softmax function without temperature parameter can introduce more colours in the train-time quantised image $\widetilde{\boldsymbol{x}}$, which increases the domain gap between train-time quantised images $\widetilde{\boldsymbol{x}}$ and test-time quantised images $\bar{\boldsymbol{x}}$ and result in overfitting. Second, removing the either of the three regularisation term leads to a 1.2% , 1.4% and 0.5% accuracy drop, respectively.

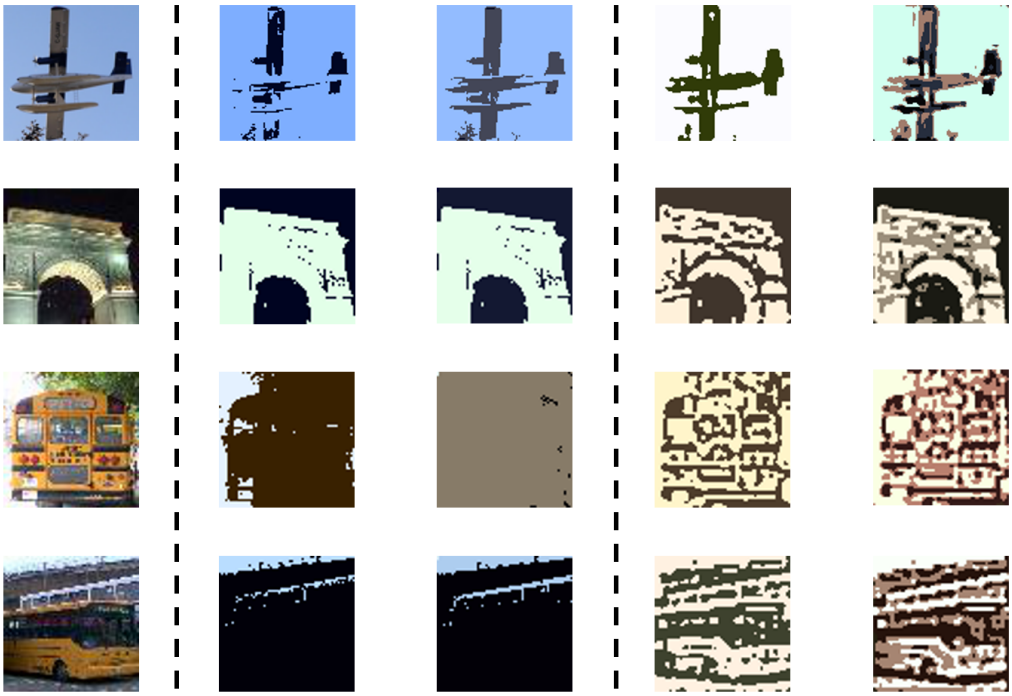
6.2 DETAILED CLASSIFICATION RESULT

Evaluating Datasets	Methods	Colour size						
		1-bit	2-bit	3-bit	4-bit	5-bit	6-bit	Full Colour (24-bit)
CIFAR10 (Krizhevsky et al., 2009)	baseline	-						94.6
	MedianCut (Heckbert, 1982)	28.1	59.5	77.8	86.7	90.9	92.9	-
	MedianCut+Dither (Floyd & Steinberg, 1976)	20.2	41.7	59.7	74.3	83.7	89.8	
	OCTree (Gervautz & Purgathofer, 1988)	22.7	40.0	60.6	77.9	85.7	90.8	
	ColorCNN (Hou et al., 2020)	53.5	70.5	78.1	80.1	79.2	80.5	
	CQFormer	80.7	83.1	83.8	85.2	84.4	85.8	
CIFAR100 (Krizhevsky et al., 2009)	baseline	-						76.3
	MedianCut (Heckbert, 1982)	11.0	26.0	43.8	57.9	66.5	71.3	-
	MedianCut+Dither (Floyd & Steinberg, 1976)	4.2	10.6	25.0	40.1	53.6	63.6	
	OCTree (Gervautz & Purgathofer, 1988)	6.2	13.1	26.8	45.5	57.5	66.4	
	ColorCNN (Hou et al., 2020)	17.6	32.9	41.5	46.1	49.3	50.7	
	CQFormer	50.6	57.7	57.2	60.3	60.2	60.6	
STL10 (Coates et al., 2011)	baseline	-						84.3
	MedianCut (Heckbert, 1982)	19.9	54.8	74.8	81.3	83.0	84.0	-
	MedianCut+Dither (Floyd & Steinberg, 1976)	17.7	45.3	64.5	73.9	79.3	82.0	
	OCTree (Gervautz & Purgathofer, 1988)	16.4	31.0	50.2	66.5	78.5	82.6	
	ColorCNN (Hou et al., 2020)	41.1	65.1	66.9	69.45	68.0	66.9	
	CQFormer	54.2	68.8	71.9	72.5	72.6	72.8	
Tiny200 (Le & Yang, 2015)	baseline	-						69.1
	MedianCut (Heckbert, 1982)	7.0	24.4	46.5	56.2	60.4	62.5	-
	MedianCut+Dither (Floyd & Steinberg, 1976)	4.2	18.7	40.2	52.0	58.6	62.2	
	OCTree (Gervautz & Purgathofer, 1988)	3.5	10.3	26.4	45.1	56.8	61.6	
	ColorCNN (Hou et al., 2020)	9.3	33.9	43.4	46.7	47.8	48.7	
	CQFormer	45.1	51.2	53.1	54.5	56.4	55.5	

Table 3: Top-1 classification accuracy of colour quantised images on four datasets with three networks. We observe that our CQFormer is significantly superior to MedianCut (Heckbert, 1982), OCTree (Gervautz & Purgathofer, 1988), MedianCut+Dither (Floyd & Steinberg, 1976) and ColorCNN (Hou et al., 2020) under low-bit (less than 3-bit) quantisation levels.

6.3 VISUALISATION

As shown in Fig. 6, our CQFormer effectively preserves more perceptual structure and similarity. For instance, airplane wings, Textures of architecture and vehicle windows.



(a) Original (b) ColorCNN (1-bit) (c) ColorCNN (2-bit) (d) CQFormer (1-bit) (e) CQFormer (2-bit)

Figure 6: Visualisation of 1-bit and 2-bit colour quantisation.