# Name Your Colour For the Task: Artificially Discover Colour Naming via Colour Quantisation Transformer Supplementary Material

Shenghan Su[1], Lin Gu[3,2*], Yue Yang[1,4], Zenghui Zhang[1], Tatsuya Harada[2,3]
[1]Shanghai Jiao Tong University, [2]The University of Tokyo, [3]RIKEN AIP, [4] Shanghai AI Laboratory
{su2564468850, yang-yue, zenghui.zhang}@sjtu.edu.cn, lin.gu@riken.jp, harada@mi.t.u-tokyo.ac.jp

## A. Ablation Study

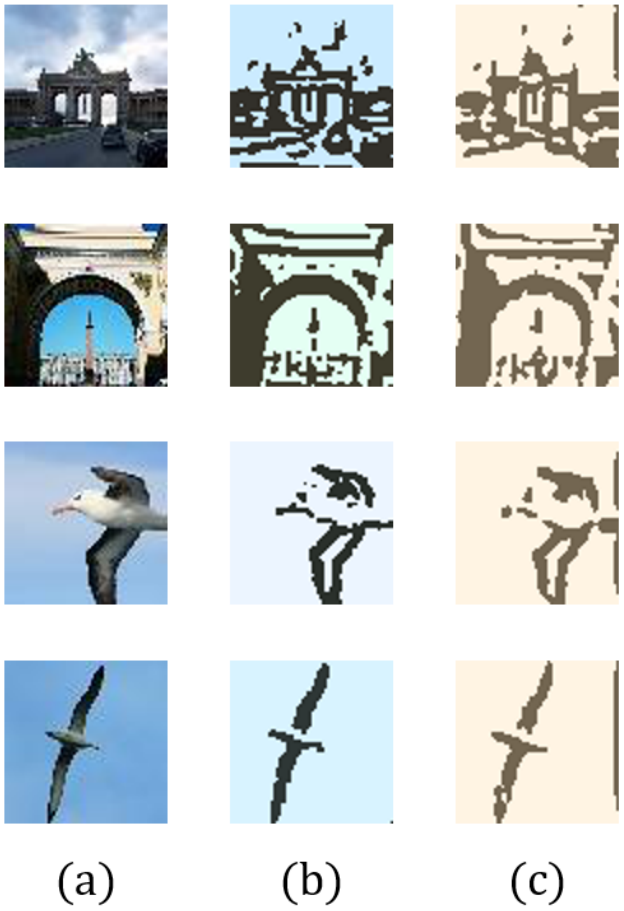



Figure 1. (a) is the original image. (b) is the quantised image with Palette Branch. (c) is the quantised image using a set of centroids instead of Palette Branch.

As shown in Table 1, we ablate the important elements in our CQFormer, using Tiny200 [8] classification dataset under 1-bit colour quantisation. We investigate the effectiveness of the temperature parameter by setting $\tau = 1.0$, perceptual structure loss by setting any terms of $\alpha$, $\beta$ and $\gamma$ as 0 and Palette Branch by replacing it with a set of centroids in colour space. As shown in Table 2, we also investigate the robustness of CQFormer by adding colour jitter and Gaussian blur using CIFAR10 [7] classification dataset under different colour quantisation levels from 1-bit to 4-bit.

**Influence of temperature parameter:** Without the temperature parameter, a severe accuracy drop (-38.8%) has occurred, which shows that the temperature parameter in our CQFormer can further approximate the $\mathrm{One\text{-}hot}(M(\boldsymbol{x}))$ using $m_\tau(\boldsymbol{x})$ during training stage to boost classification accuracy.

**Influence of perceptual structure loss:** With the $R_{\mathrm{Colour}}$, $R_{\mathrm{Diversity}}$ and $\mathcal{L}_{\mathrm{Perceptual}}$, our CQFormer improve the top-1 accuracy by 1.2%, 1.4% and 0.5%, respectively. When we remove all of them, a considerable accuracy drop of -7.8% has occurred. It demonstrates that the perceptual structure loss contributes to better machine accuracy besides maintaining perceptual similarity.

**Influence of Palette Branch:** If we remove Palette Branch and make the CQFormer just learn a set of centroids in RGB colour space, the accuracy is 30.4%, resulting in a severe drop of 14.7%. As shown in Fig. 1, it would make all the images have the same colour palette rather than the same amount of colours, resulting in a loss of perceptual similarity, *e.g.* the blue sky in Col.(c) is represented as light yellow. Therefore, our Palette Branch ensures that reference palette queries are sent to interact with the image and creates the colour palette using both machine preference and image perception features, which maintains the colour specificity of each image.

**Robustness of CQFormer:** We add a colour jitter and Gaussian blur to the colour-quantised image. Results are shown in Table. 2. For example, on the CIFAR10 classification dataset, we achieve 79.3%, 81.6%, 83.4%, and 84.6% top-1 accuracy with a colour jitter from 1-bit to 4-bit

*Corresponding author.

Table 1. Ablation study results under 1-bit colour quantisation, Tiny200 [8] dataset, Renset18 [4] classifier.

| $\tau$ | $R_{\text{Colour}}$ | $R_{\text{Diversity}}$ | $\mathcal{L}_{\text{Perceptual}}$ | Palette Branch | Accuracy |
|---|---|---|---|---|---|
| ✔ | ✔ | ✔ | ✔ | ✔ | **45.1** |
| ✗ | ✔ | ✔ | ✔ | ✔ | 6.9 |
| ✔ | ✗ | ✔ | ✔ | ✔ | 43.9 |
| ✔ | ✔ | ✗ | ✔ | ✔ | 43.7 |
| ✔ | ✔ | ✔ | ✗ | ✔ | 44.6 |
| ✔ | ✗ | ✗ | ✗ | ✔ | 37.3 |
| ✔ | ✔ | ✔ | ✔ | ✗ | 30.4 |

Table 2. Ablation study results of robustness under under different colour quantisation levels from 1-bit to 4-bit, CIFAR10 [7] dataset, Renset18 [4] classifier.

| | 1 bit | 2 bit | 3 bit | 4 bit |
|---|---|---|---|---|
| CQFormer | **80.7** | **83.1** | **83.8** | **85.2** |
| CQFormer (with colour jitter) | 79.3 | 81.6 | 83.4 | 84.6 |
| CQFormer(with Gaussian blur ) | 80.6 | 81.7 | 81.9 | 83.6 |

colour quantisation. The colour jitter causes a little drop of 1.4%, 1.5%, 0.4%, and 0.6%, respectively. Therefore, our CQFormer is robust enough to overcome different noises.

## B. Object Detection with other detector

To evaluate our CQFormer's generalisation on another object detector, we use the popular object detector Faster-RCNN [10] for evaluation. We jointly train our CQFormer with Faster-RCNN for 12 epochs on 4 Tesla V100 GPUs. We adopt the SGD optimiser with an initial learning rate $0.01$, momentum $0.9$ and weight decay $0.0001$, learning rate decay to one-tenth at 8 and 11 epochs. The dataset and other training settings are the same as Sparse-RCNN [12] experiments in Sec. 4.2.

## C. The Reason of Choosing Nafaanra for Colour Evolution Experiments

Colour naming data for Nafaanra [13] were initially collected in 1978 in Banda Ahenkro, as part of the WCS [1], which is a 3-term system, with terms for light ('fiNge'), dark ('wOO'), and warm or red-like ('nyiE'). In 2018, 40 years after the original WCS data collection, Nafaanra colour naming data were collected again in the same town. As technology and lifestyle changed in the community, the Nafaanra colour naming system changed substantially between 1978 and 2018, becoming more semantically fine-grained by adding new colour terms and adjusting the extension of previously existing terms [13]. Therefore, the Nafaanra language is an excellent example for comparing the evolution trajectory of human language using the machine.

## D. The Relationship between the IB Color Naming Model and CQFormer

Fig. 2(a) is the information bottleneck (IB) color naming model proposed by [14]. A straightforward communication scenario, which can be derived from Shannon's communication model [11], serves as the foundation for this theoretical framework.

Fig. 2(b) is our colour quantisation model. The motivation of our colour quantisation model architecture is driven by the IB color naming model [14]. As illustrated in (b), the speaker represents the CQFormer, and the listener represents the classifier. We focus on the case where a colour quantiser (CQFormer) and a classifier communicate about colours. The CQFormer has a "mental" representation, *i.e.* a full-colour image $\boldsymbol{x}$ associated with a prior label $y$ drawn from a prior distribution, and communicates this representation by encoding it into a colour-quantised image $\bar{\boldsymbol{x}}$ according to the perceptual similarity. The classifier receives $\bar{\boldsymbol{x}}$ and attempts to infer from it the full-colour image's label $y$ by predicting the label $\hat{y}$ and constructing another distribution that approximates $y$.

There exists a problem that both the speaker and listener in Fig. 2 (a) have a knowledge of colour naming/recognition at the same time. In contrast, the untrained CQFormer and classifier lack knowledge of colour quantisation and image classification. Therefore, we jointly train both the CQFormer and classifier simultaneously to add prior knowledge of colour quantisation and image classification under a specific bit of colour. Finally, similar to the theoretical limit of semantic efficiency in [14], we obtain optimal image classification accuracy under the specific bit of colour.

Table 3. Object detection results on MS COCO dataset [9] with Faster-RCNN [10] detector, here we report the average precision (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) [5] | 6.5 | 9.7 | 11.4 | 14.0 | **16.8** | 18.1 | - |
| Median Cut (w D) [2] | 7.3 | 8.8 | 11.2 | 13.4 | 14.8 | 15.6 | - |
| OCTree [3] | 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** | - |

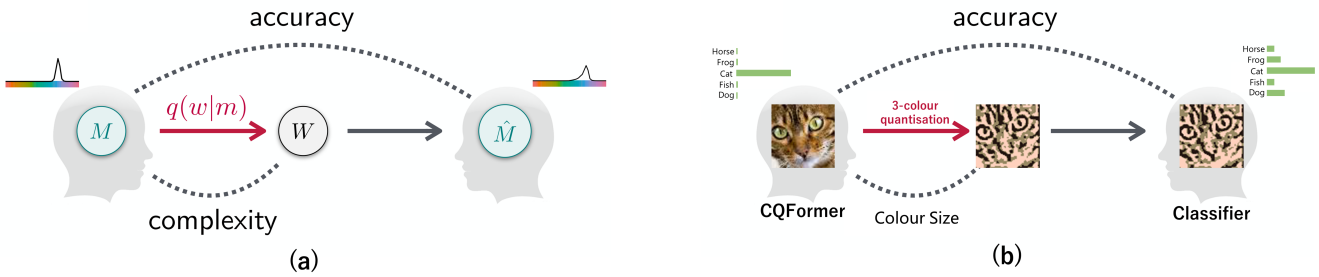



Figure 2. (a) is the IB color naming model proposed by [14].(b) is our colour quantisation model.

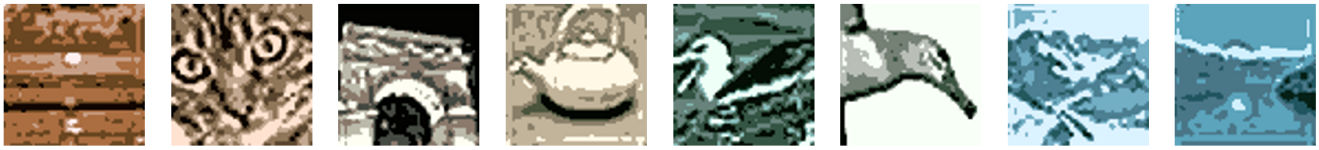


Figure 3. 8-colour quantised images by CQFomer.

## E. Visualisation

As shown in Fig. 4 and Fig. 3, our CQFormer effectively preserves more perceptual structure and similarity. For instance, aeroplane wings, textures of architecture and vehicle windows.

## References


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

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

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

[4] 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*, pages 770–778, 2016.

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

[6] 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*, pages 10116–10125, 2020.

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

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

[9] 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, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

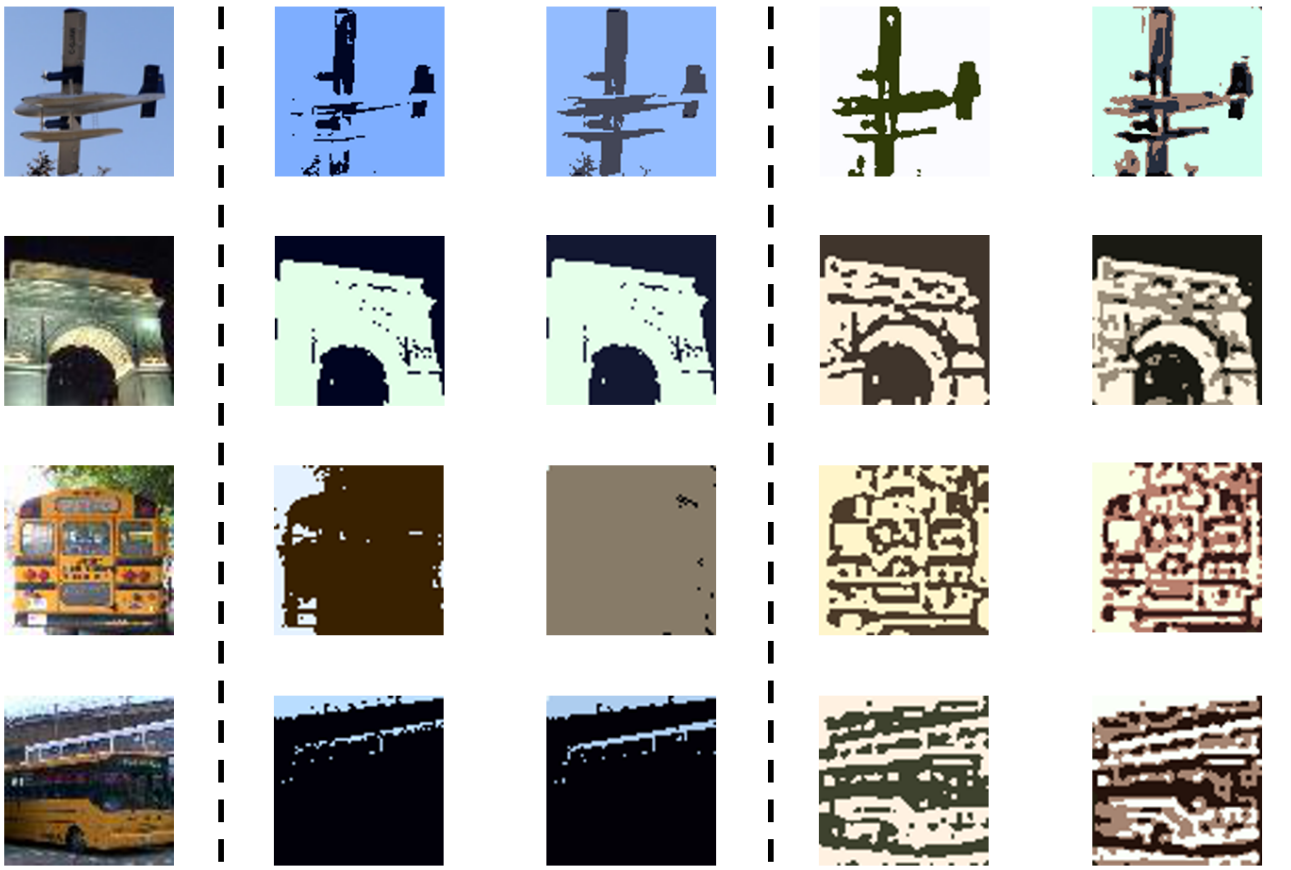
[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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

[12] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[13] 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.

[14] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.



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

Figure 4. Visualisation of 1-bit and 2-bit colour quantisation by ColorCNN [6] and CQformer.