Reprinted from
Journal of Perceptual Imaging® 4(2): 020401-1–020401-10, 2021.

Color Conversion in Deep Autoencoders

Arash Akbarinia[1] and Raquel Gil-Rodríguez[1]

[1]*Department of Experimental Psychology, Justus-Liebig University, D-35394, Giessen, Germany*
E-mail: arash.akbarinia@psychol.uni-giessen.de

Abstract. *While RGB is the status quo in machine vision, other color spaces offer higher utility in distinct visual tasks. Here, the authors have investigated the impact of color spaces on the encoding capacity of a visual system that is subject to information compression, specifically variational autoencoders (VAEs) with a bottleneck constraint. To this end, they propose a framework—color conversion—that allows a fair comparison of color spaces. They systematically investigated several ColourConvNets, i.e. VAEs with different input–output color spaces, e.g. from RGB to CIE L*a*b* (in total five color spaces were examined). Their evaluations demonstrate that, in comparison to the baseline network (whose input and output are RGB), ColourConvNets with a color-opponent output space produce higher quality images. This is also evident quantitatively: (i) in pixel-wise low-level metrics such as color difference (ΔE), peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM); and (ii) in high-level visual tasks such as image classification (on ImageNet dataset) and scene segmentation (on COCO dataset) where the global content of reconstruction matters. These findings offer a promising line of investigation for other applications of VAEs. Furthermore, they provide empirical evidence on the benefits of color-opponent representation in a complex visual system and why it might have emerged in the human brain.* © *2021 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.Percept.Imaging.2021.4.2.020401]

1. INTRODUCTION

Color is an inseparable component of our conscious visual perception with an objective utility spanning over a large set of tasks such as object recognition and scene segmentation [8]. Consequently, color has become a ubiquitous feature in machine vision and image processing. Currently, state of the art and practice in these fields are being dominated by deep learning methods. Thus, progress in these lines requires a better understanding of the networks' underlying mechanism [3] and the color representation learned by them.

The human color vision is a result of three types of cone photoreceptors present in the retina [6]. Thus, models of color perception become defined in a three-dimensional space. In theory, an infinite number of color spaces could be formulated and indeed several of them exist in the literature and industry [55]. RGB color sensors are the standard in off-the-shelf commercial cameras. This makes the RGB color space widely used in computer vision and deep learning applications. We are interested to know whether the choice of color representation influences the capacity of deep networks in visual information processing. This is a generic endeavor not targeted toward a specific application. A common real-world physical restriction to all applications is the bottleneck in information transmission. Hence, autoencoders are a perfect tool to study this question given their objective is simply efficient coding under a similar constraint [50].

To this end, we propose the *color conversion* framework, in which the input–output color spaces are explicitly imposed on deep autoencoders (referred to as *ColourConvNets*). ColourConvNets learn to compress the visual information in their bottleneck while transforming the input to output. Essentially, the output y for input image x is generated on the fly by a transformation $y = T(x)$, where T maps input to output. Color conversion offers a framework to fairly compare the effect of color spaces in a complex visual system that is driven by optimization. Here, we study the choice of color conversion on the quality of reconstructed images, which is an indication of whether the representation of input–output color spaces impacts the network's encoding power.

In this work, we focused on Vector Quantized Variational Autoencoder (VQ-VAE) [52] due to the discrete nature of its latent space. We thoroughly studied five commonly used color spaces by training ColourConvNets for all combinations of input–output spaces. First, we show that ColourConvNets with a decorrelated output color space (e.g. CIE L^*a^*b) convey information more efficiently in their compressing bottleneck, in line with the presence of color opponency in the human visual system [5]. This is evident qualitatively (Figure 7) and quantitatively (evaluated with three low-level and two high-level metrics). We further discuss a potential explanation at the level of embedding vectors linking it to the *histogram equalization* technique [41] and the *efficient coding* theory [4].

2. RELATED WORK

Various color spaces have been explored in classical computer vision to boost the performance of algorithms. Color-opponent spaces (e.g. CIE $L^*a^*b^*$) have been extensively used in applications of image retrieval [42], color constancy [1], color stabilization [19], color transfer [43], color naming [40], texture classification [7], edge detection [2] to name a few. Combinations of intensity, saturation and hue (e.g. HSV) have also been shown effective in applications

Received May 14, 2021; accepted for publication Oct. 1, 2021; published online Oct. 28, 2021. Associate Editor: Pierre Dragicevic.
2575-8144/2021/4(2)/020401/10/$00.00

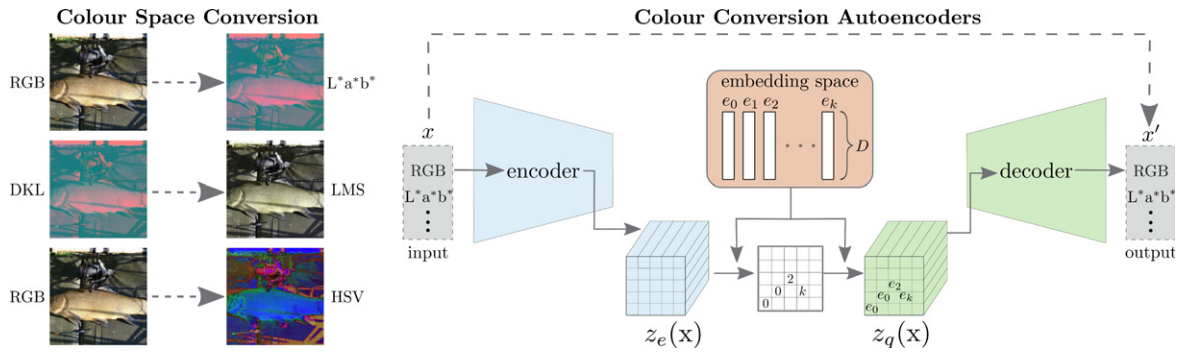

Figure 1. Left: Exemplary conversions across different color spaces. Right: The schematic view of VQ-VAE ColourConvNets.

such as object recognition [18], skin classification [21], and object tracking [39]. In general, the fusion of color spaces is reported to create an optimal feature detector [49].

In comparison to the classical approaches, the utility of color spaces in deep neural networks (DNNs) is understudied. Initial work suggested that non-RGB color spaces do not boost the performance in the ImageNet dataset [36]. Contrary to this, a fusion of three color spaces (RGB, HSV and CIE $L^*a^*b^*$) has improved retinal medical imaging [16]. Similarly, a multi-channel architecture combining three color spaces (RGB, HSV and YCbCr) has been proposed for face identification [29]. Integrating six networks of different color spaces has also been successfully applied to traffic light recognition [23]. In addition to this, color spaces in which luminance and chromatic information have separate channels (e.g. YUV) are in particular helpful in applications such as picture colorization [30] and style transfer [35]. Last but not least, the prediction of luminance from chromatic planes and vice versa has been explored in unsupervised learning [57].

Color spaces have also been a topic of research in the efficient coding literature. The choice of color space influences the degree of image compression and efficient representation [48]. This has made color conversion a standard technique in image compression. In certain on-board systems (e.g. Mars Exploration Rover) the extra computational cost of finding an optimal space for a set of images is justified [56]. Consequently, modern image file formats allow for color-space information to be stored in their metadata [44]. In the case of the commonly used JPEG image compression, it has been specifically shown that RGB is the least and CIE $L^*a^*b^*$ is the most optimal color space [37]. Correspondingly, classical learning-based methods of image compression also use opponency color spaces (i.e. one luminance and two chromatic channels) [9]. To the best of our knowledge, this finding has not been thoroughly examined in modern deep autoencoders [24]. As opposed to classical approaches, current compression studies rely on the encoder capabilities [51], without applying any previous color transformation. In this article, we aim to break this gap by systematically comparing color spaces in the context of deep autoencoders.

3. COLOR CONVERTING AUTOENCODERS

In this article, we propose a novel unsupervised task of color conversion: the network's output color space is independent of its input (see Figure 1). This is inspired by the human visual system, in which the sensory and perceptual systems work in different color spaces. The input to our visual system is triggered by photoreceptors in the back of the retina. Hence, the sensory system is defined in the LMS color space [17]. Before reaching the cortex, this signal is transformed into a cone-opponent space by the opponent cells present in the retina and the lateral geniculate nucleus (LGN) [12]. Behavioral studies suggest that yet another color-opponent space shapes our perceptual system [54]. Last but not least, it has been argued that the current color spaces cannot fully explain the dimension of hue in which colors and objects are associated [26]. This collection of studies in the literature suggests that our visual system functions with different color spaces for distinct goals. A similar observation can be made for machine vision. While the sensory system is in the RGB color space (the input to the system), alternative spaces might be more efficient for other purposes.

A color space is an arbitrary definition of colors' organization in space [27]. Thus, the choice of transformation matrix $\boldsymbol{T}$ in ColourConvNets is perfectly flexible to model any desired space,

$$\mathscr{C}_{in} \xrightarrow{\boldsymbol{T}} \mathscr{C}_{out}, \tag{1}$$

where $\mathscr{C}_{in}$ and $\mathscr{C}_{out}$ are the input and output color spaces. This framework offers a controlled environment to compare color spaces within a complex visual system. Here, we compared them in an information encoding network that is constrained to a bottleneck. This loosely corresponds to the need for signal compression in the human visual system due to present physical constraints. An extension of the proposed framework can encompass other constraints (such as entropy, energy, wiring, etc.) relevant to understanding color representation in complex visual systems. This structure can be further used to compare the autoencoder's latent space across color spaces aiming to decipher the intermediate color representation within these networks [14]. The proposed framework can also be

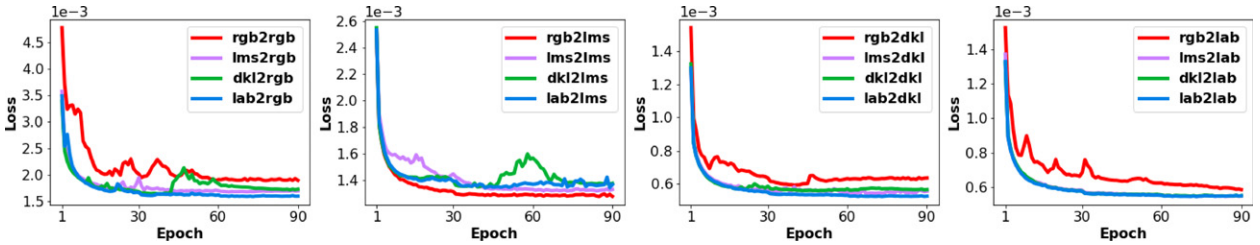

Figure 2. Evolution of losses for VQ-VAEs of K = 8 and D = 128. In each panel, the ColourConvNets have the same output space. Across panels, curves of the same color have the same input space.

employed in applications, e.g., as an add-on optimization capsule to any computer vision application [38], or as a proxy task for visual understanding [30].

3.1 *Networks*

One fundamental property of neural activity in biological brains is "all-or-none" [22]. This, in turn, strengthens the argument of discrete representation [34]. Hence, we studied a particular class of VAEs—Vector Quantized Variational Autoencoder (VQ-VAE) [52]—due to the discrete nature of its latent embedding space, which distinguishes it from other regimes [24].

VQ-VAE consists of three main components (see the right panel in Fig. 1):

1. An `encoder` $f(\mathrm{x})$ that processes the input data x to $z_e(x)$ by non-linear operations;
2. A `embedding space` $\{\mathrm{e}\} \in \mathbb{R}^{K\times D}$, with $\mathbf{K}$ vectors of dimensionality D, mapping the continuous $z_e(x)$ onto a sequence of discrete latent variables $z_q(x)$ by estimating the nearest vector $\mathbf{e}_i$ to $z_e(x)$;
3. A `decoder` $g(\mathrm{e})$ that reconstructs the final output x' with a distribution $p(x|z_q(x))$ over the input data.

The loss function $\mathscr{L}$ is defined as follows,

$$\mathscr{L} = \|\mathrm{y} - g(\mathrm{e})\|_2^2 + \|sg[f(\mathrm{x})] - \mathrm{e}\|_2^2 + \beta\|f(\mathrm{x}) - sg[\mathrm{e}]\|_2^2, \tag{2}$$

where y is the target image (i.e. x in the output color space); sg denotes the stop gradient computation that is defined as the identity during the forward propagation with zero partial derivatives during the backpropagation to refrain its update.

The first term in Eq. (2) corresponds to the quality of the reconstruction image by jointly updating encoder and decoder. The other two terms align the embedding vector with the encoder output. The second term only updates the latent variables (embedding vectors). The third term only updates to the encoder. The hyperparameter $\beta \in \mathbb{R}$ regulates the degree of change for the encoder output. Without a hyperparameter search, we set $\beta = 0.5$ in all conducted experiments.

3.2 *Color Spaces*

We explored five color spaces: RGB, LMS, CIE $L^*a^*b^*$, DKL and HSV. The standard space in digital imaging is RGB that represents colors by three additive primaries in a cubic shape. The LMS color space corresponds to the sensitivity function of cones in the human eye (long, middle, and short wavelengths) [17]. The CIE $L^*a^*b^*$ color space (luminance, red-green and yellow-blue axes) is designed to be perceptually uniform [10]. The DKL color space (Derrington–Krauskopf–Lennie) models the opponent responses of rhesus monkeys in the early visual system [12]. The HSV color space (hue, saturation, and value) is a cylindrical representation of the RGB cube designed by computer graphics.

The input–output to our networks can be in any combination of these color spaces. Effectively, our VQ-VAE models, in addition to learning efficient representation, must learn the transformation function from their input to output color space. It is worth considering that the original images in explored datasets are in the RGB format. Therefore, one might expect a slight positive bias toward this color space given its gamut defines the limits of other color spaces.

4. EXPERIMENTS

4.1 *Training Procedure*

We trained several instances of VQ-VAEs with distinct sizes of embedding space $\{e\} \in \mathbb{R}^{K\times D}$. The training procedure was identical for all networks: trained with Adam optimizer ($lr = 2 \cdot 10^{-4}$) for 90 epochs. To isolate the influence of random variables, all networks were initialized with the same set of weights and an identical random seed was used throughout all experiments. We used the ImageNet dataset [13] for training. This is a visual database of object recognition in real-world images, divided into one thousand categories. The training set contains 1.3M images. At each epoch, a subset of 100K samples was exposed to networks. Input images were of size 224×224 in three color channels. Figure 2 reports the progress of loss function for all ColourConvNets with an embedding space of size $\{e\} \in \mathbb{R}^{8\times 128}$. A similar pattern of convergence can be observed in all trained networks suggesting that the optimization is a fair comparison across different input–output color spaces.

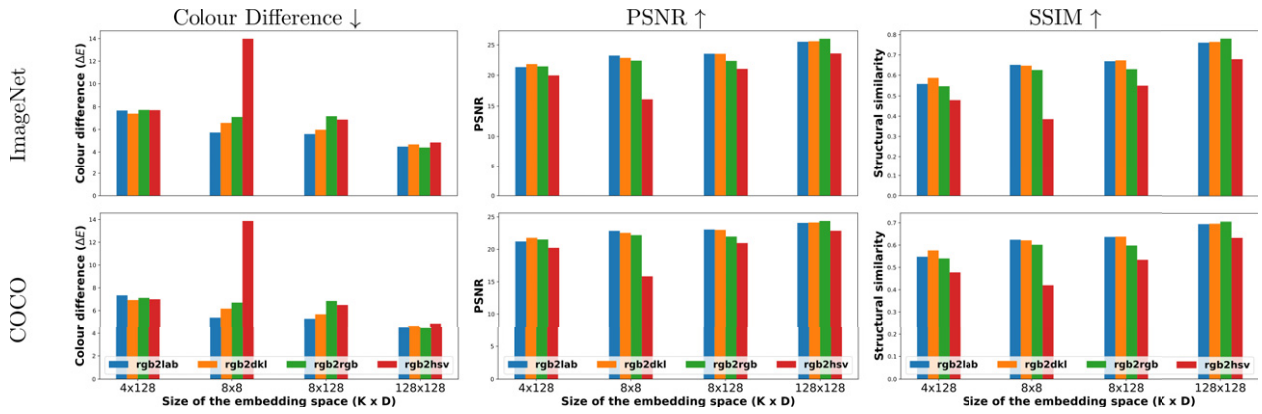

Figure 3. Low-level evaluation for embedding spaces of different sizes. Lower values of color difference and higher values of PSNR and SSIM indicate higher quality of the reconstruction.

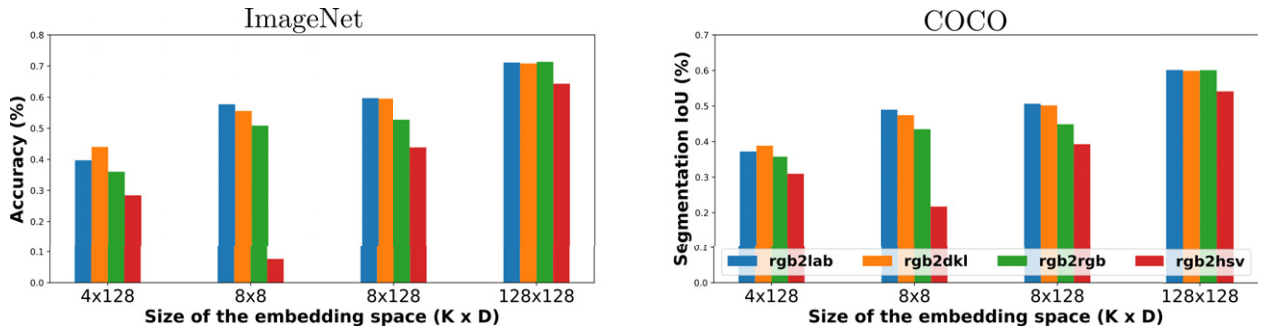

Figure 4. High-level visual task evaluation. ResNet50's classification accuracy on reconstructed images of ImageNet and FPNS's segmentation IoU on reconstructed images of COCO.

4.2 *Evaluation Protocol*

To increase the generalization power of our findings, we evaluated all networks (without any fine-tuning) on the validation set of three benchmark datasets: ImageNet (50K images), COCO (5K images), and CelebA (∼20K images). COCO is a large-scale dataset for object detection and scene segmentation in natural images [32]. CelebA contains facial attributes of celebrities [33]. The types of images in CelebA dataset (close-up faces) rarely appear in the train set of our networks (i.e. ImageNet). We relied on two classes of evaluation: low level, capturing the local statistics of an image; high level, assessing the global content of an image (For reproduction, the source code and experimental data are available: https://github.com/ArashAkbarinia/DecomposeNet).

4.2.1 *Low-level Evaluation*

We computed three commonly used metrics to measure the pixel-wise performance of networks: (i) the color difference CIEDE2000 (ΔE) [47], (ii) peak signal-to-noise ratio (PSNR), and (iii) structural similarity index measure (SSIM) [53]. These metrics are often used in the literature of image quality assessment. Lower values of ΔE and higher values of PSNR and SSIM indicate better performance.

4.2.2 *High-level Evaluation*

Pixel-wise measures are unable to capture the global content of an image and whether semantic information remains perceptually intact. To account for this limitation, we performed a procedure similar to the standard Inception Score [46]: feeding the reconstructed images into two pretrained networks (without fine-tuning) that perform the task of object classification, ResNet50 [20], and scene segmentation, Feature Pyramid Network—FPN [25]. ResNet50 and FPN expect RGB inputs, thus non-RGB reconstructed images were converted to RGB. The evaluation for ResNet50 is the classification accuracy on the ImageNet dataset. The evaluation for FPN is the intersection over union (IoU) on the COCO dataset.

4.3 *Embedding Size*

We first evaluated the influence of embedding size for four regimes of ColourConvNets whose input color space is the original RGB images. The low-level evaluation for the ImageNet and COCO datasets is reported in Figure 3. The most noticeable data point (in all three metrics) is the poor performance of *rgb2hsv* with embedding space $\{e\} \in \mathbb{R}^{8\times8}$. This might be due to the circular nature of the

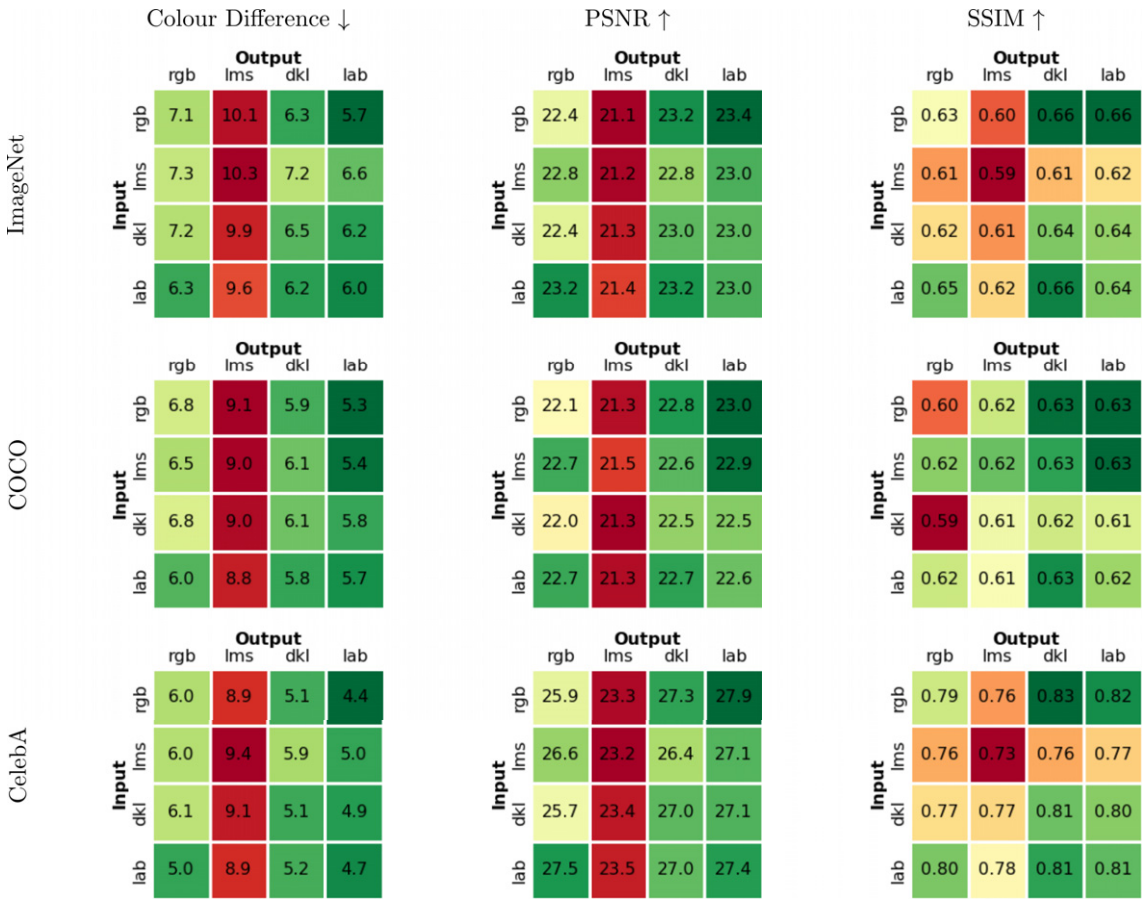

Figure 5. Low-level pairwise comparison in four color spaces. Figures are averaged over two embedding spaces 8×8 and 8×128. Lower values of color difference and higher values of PSNR and SSIM indicate higher quality of the reconstruction. The cells are color-coded accordingly.

hue information that cannot be adequately encoded with low-dimensional vectors (i.e. $D = 8$). For the smallest and the largest embedding space, we observe no significant differences between the four networks. However, for intermediate embedding spaces (i.e. 8×8 and 8×128) an advantage appears for networks whose outputs are opponent color spaces (DKL and CIE L^*a^*b).

The corresponding high-level evaluation is reported in Figure 4. The overall trend is much alike for both tasks. The lowest performance occurs for *rgb2hsv* across all embedding spaces. ColourConvNets with an opponent output color space systematically perform better than *rgb2rgb*, with an exception for the largest embedding space (128×128) where they are on a par with each other (despite the substantial compression, 70% top-1 accuracy on ImageNet and 60% IoU on COCO). The comparison of low- and high-level evaluation for the smallest embedding space (4×128) demonstrates the importance of high-level evaluation. Although in the low-level metrics the four networks perform similarly, in the high-level metrics a large difference appears among them (compare Fig. 4 versus Fig. 3). The classification and segmentation performance is substantially influenced by the choice of color space. Overall, the results of the embedding size experiment suggest that when physical constraints demand heavy compression (i.e. narrow bottleneck) *rgb2lab* and *rgb2dkl* autoencoders better preserve the semantic content of images.

Noise reduction is a primary application of autoencoders. Essentially the imposed bottleneck enforces the system to ignore insignificant information. Correspondingly, we tested all networks after adding different degrees of salt-and-pepper noise to the input images. The ColourConvNets with an opponent output space systematically outperformed the baseline in this experiment as well. While this does not explicitly indicate better noise reduction in these networks, it demonstrates that their efficiency is generalized to out-of-the-distribution conditions.

4.4 *Pairwise Comparison*

For the two embedding spaces 8×8 and 8×128 we conducted an exhaustive pairwise comparison across two regimes of color spaces: sensory (RGB and LMS) versus opponency (DKL and CIE L^*a^*b). The HSV color space was excluded due to the aforementioned reason. Figure 5 presents the low-level evaluation. ColourConvNets with an opponent output space clearly perform better across all measures and datasets. Specifically, in comparison to the baseline (the *rgb2rgb* network) both *rgb2lab* and *rgb2dkl* obtain

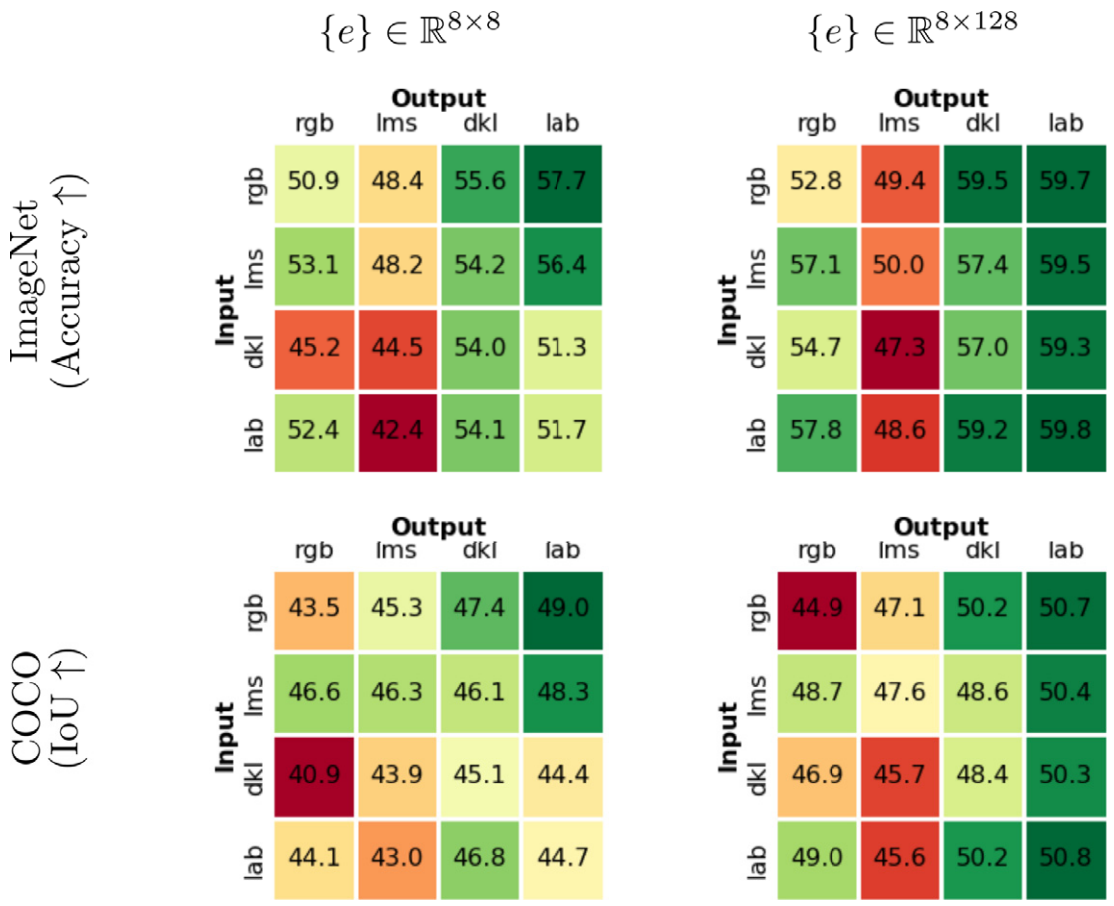

Figure 6. High-level pairwise comparison in four color spaces: sensory (RGB and LMS) and opponency (DKL and CIE L^*a^*b).

substantially lower color differences, and higher PSNRs and SSIMs.

The comparison of rows and columns in Fig. 5 suggests that the quality of compression is more influenced by the output color space in comparison to the input. Specifically, the poor performance of networks whose output space is LMS is noticeable. This can be explained by the high correlation between different channels of the LMS. Essentially, ColourConvNets struggle to accurately decode each of those channels separately. This problem does not occur when LMS is the input color space.

The pairwise high-level evaluation is reported in Figure 6. In agreement to previous findings, the *rgb2lab* network performs best across both datasets and embedding spaces. Overall, ColourConvNets with an opponent output space show a clear advantage: *rgb2lab* and *rgb2dkl* obtain 5–7% higher accuracy and IoU with respect to the baseline (the *rgb2rgb* to network).

4.5 *Qualitative Comparison*

In addition to the quantitative evaluations reported in the previous section, the advantage of utilizing a decorrelated output color space can be appreciated qualitatively. In Figure 7, we have illustrated five representative examples, four images from the dataset of natural scenes and one image from the faces. The Jupyter-Notebook scripts in our GitHub provide more examples and can be executed for user-input images. Overall, the perceptual quality of the image reconstruction in ColourConvNets with an opponent output space (*rgb2dkl* and *rgb2lab*) is visibly higher than the baseline *rgb2rgb*. For instance, in the first row of Fig. 7, the *rgb2rgb* output contains a large number of artifacts on walls and ceilings. In contrast, the output of *rgb2dkl* and *rgb2lab* are sharper. This qualitative difference can also be appreciated on the cabinets of the second row and glasses of the third row (it is best seen in the digital format with full resolution). It is challenging to quantify the types of natural scenes with the greatest advantage for color opponency. This might be better addressed in a more controlled dataset where images are generated from a set of predefined reflectance spectra. Nevertheless, we observed a more prominent effect under two conditions. First, in uniform regions many times the *rgb2rgb* network appears to greatly suffer. For instance, this is evident from the blue sky in the fourth row of Fig. 7. Second, in many instances the *rgb2rgb* fails to faithfully reproduce the color of an object (see the red cloth in the last row).

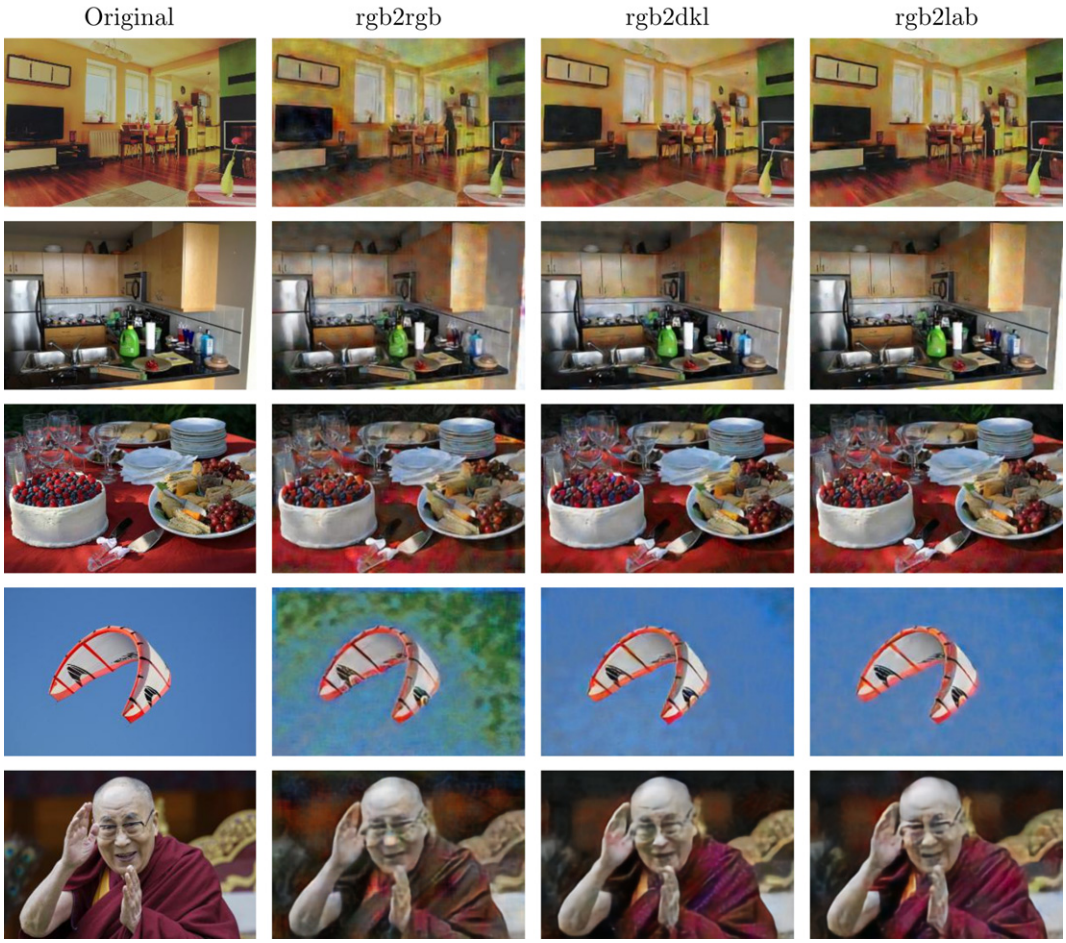

Figure 7. Qualitative comparison of three ColourConvNets (VQ-VAE of K = 8 and D = 128). The first column is the networks' input and the other columns their corresponding outputs. The output images of *rgb2dkl* and *rgb2lab* have been converted to the RGB color space for visualization purposes. The artifacts in *rgb2rgb* are clearly more visible in comparison to the other ColourConvNets.

5. PERFORMANCE ADVANTAGE

The main difference between the two regimes of color spaces (sensory versus opponency) is their intra-axes correlation. In other words, the extent of information independence in each of the three channels. The intra-axes correlation for LMS and RGB is very high, hence referred to as *correlated* color spaces. On the contrary, the intra-axes correlations for CIE $L^*a^*b^*$ and DKL is very low, hence referred to as *decorrelated* color spaces. We computed these correlations r in all images of ImageNet dataset (100 random pixels per image). RGB: $r^{RG} \approx 0.90$, $r^{RB} \approx 0.77$, $r^{GB} \approx 0.89$; LMS: $r^{LM} \approx 1.00$, $r^{LS} \approx 0.93$, $r^{MS} \approx 0.93$; $L^*a^*b^*$: $r^{L*a*} \approx -0.14$, $r^{L*b*} \approx 0.13$, $r^{a*b*} \approx -0.34$; DKL: $r^{DK} \approx 0.01$, $r^{DL} \approx 0.14$, $r^{KL} \approx 0.61$. In biological visual systems, the retinal signal is transformed to opponency before being transmitted to the visual cortex by passing through the physical bottleneck of optical nerve and LGN. This transformation has been argued to optimize the efficiency of color signal transmission in the visual system by reducing redundant information [5].

Interestingly, some works have suggested that deep networks trained to perform high-level visual tasks learn to decorrelate their inputs [45]. Here, our results show a similar phenomenon in deep autoencoders: information compression is more efficient when a network decorrelates the input signal. Contrary to this, the ImageNet classification performance was reported unaltered when input images were explicitly converted from RGB to CIE $L^*a^*b^*$ [36]. This might be explained by the lack of bottleneck constraint in their examined architecture, thus decorrelating color representation leads to no extra advantage. This matches the results we obtained with ColourConvNets of the largest embedding space (128×128), suggesting that decorrelation of color signal become beneficial when the system is constrained in its information flow.

Previous works in the literature [15] have measured the decorrelation characteristics of color-opponent spaces in information-theoretical analysis and demonstrated their effectiveness in encoding natural images. The understanding of how a complex visual system, driven by an error minimization strategy [28], might utilize these properties at the system level is of great interest. We hypothesized that an efficient system distributes its representation across all resources

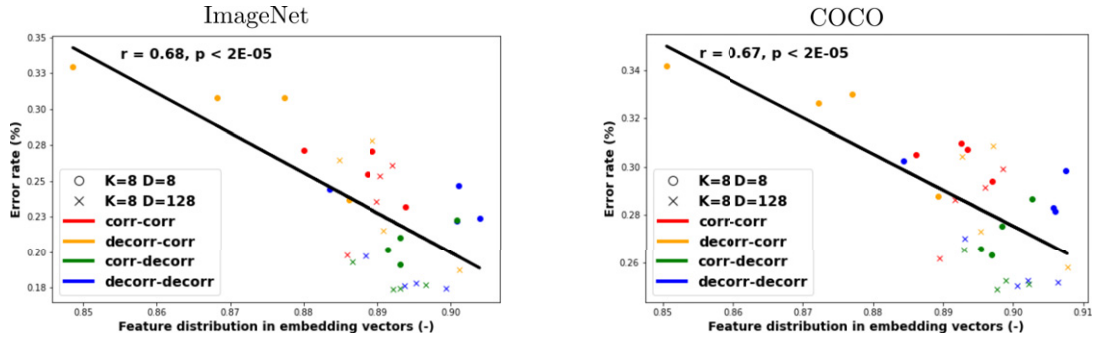

Figure 8. Error rate as a function of the distribution of features in the embedding space. A value of zero in the x-axis indicates all embedding vectors are equally used by the model. Higher values of x indicate that the model relies heavily on certain vectors.

instead of heavily relying on a few components [31]. To measure this, we computed the histogram of embedding vectors across all images of the validation set in the ImageNet (50K) and COCO (5K) datasets. A zero standard deviation in the frequency of selected vectors means embedding vectors are equally used by the network. This can be interpreted as an indication of well-distributed feature representation in the system.

Figure 8 reports the error rate as a function of this measure. A significant correlation emerges in both datasets, suggesting a more uniform contribution of embedding vectors enhances visual encoding in VQ-VAEs. To ensure the obtained correlation is robust, we analyzed the sensitivity of this correlation by means of two methods. (i) To determine highly influential points, we performed the Cook's Distance [11]. No points surpass the standard outlier threshold ($I_t = \frac{4}{n}$). (ii) We performed the Jackknife resampling technique and systematically computed the correlations after leaving out each ColourConvNet. The obtained correlations are in the range of $[0.59, 0.72]$ with an average of 0.67 ± 0.02. Overall, these analyses suggest that there is a correlation between the distribution of features among the embedding vectors and the encoding capacity of the network.

Our findings can be linked to two frameworks of *histogram equalization* and *efficient coding*. The neural model of histogram equalization follows a similar line of reasoning: the materialization of all intensity values [41]. This is achieved by explicitly minimizing a corresponding term in an objective function. This is also consistent with the efficient coding theory for the biological organisms [4], in which the system distributes its encoded representation across all response levels with an equal frequency. Here, we observe a similar phenomenon in VQ-VAEs: ColourConvNets that better materialize all their embedding vectors obtain higher quality in image compression.

6. CONCLUSION

We proposed the unsupervised *color conversion* task to investigate the efficiency of color representation in deep networks. By means of this framework, we studied the impact of color spaces on the encoding capacity of autoencoders, specifically VQ-VAEs whose feature representation is constrained by a discrete bottleneck. The comparison of several ColourConvNets exhibits advantage for a decorrelated output color space. This is evident qualitatively and measured quantitatively with five metrics. Our analysis suggests that this advantage stems from a more uniform distribution of feature representation in networks' embedding space, which is reminiscent of efficient coding and histogram equalization in biological systems.

We propose two lines of investigation for future works. First, integrating the choice of color spaces into the optimization problem, essentially driving the network to explicitly find the most optimum color space for the visual task it is learning. This formulation allows a flexible add-on optimization capsule to any computer vision application. Second, our findings might contribute to the understanding of why the brain's neural network has naturally evolved a particular type of color vision and perception. To better investigate this, we propose to include further biologically motivated constraints (e.g. entropy) on the network. These configurations would perhaps result in the emergence of color categories when a visual scene is being efficiently encoded.

ACKNOWLEDGMENT

This study was funded by Deutsche Forschungsgemeinschaft SFB/TRR 135 (grant number 222641018) TP C2. We would like to thank Matteo Toscani and Alban Flachot for their valuable feedback.

REFERENCES

[1] A. Akbarinia and C. A. Parraga, "Colour constancy beyond the classical receptive field," IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2081–2094 (2017).

[2] A. Akbarinia and C. A. Parraga, "Feedback and surround modulated boundary detection," Int. J. Comput. Vis. **126**, 1367–1380 (2018).

[3] A. Akbarinia and R. Gil Rodríguez, "Deciphering image contrast in object classification deep networks," Vis. Res. **173**, 61–76 (2020).

[4] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," *Sensory Communication* (MIT Press, Cambridge, UK, 1961), pp. 217–234.

[5] G. Buchsbaum and G. Allan, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," Proc. R. Soc. Lond. B **220**, 89–113 (1983).

[6] M. E. Burns and T. D. Lamb, "16. visual transduction by rod and cone photoreceptors," *Visual Neuroscience* (MIT Press, 2003), pp. 215–233, Citeseer.

[7] E. Cernadas, M. Fernández-Delgado, E. González-Rufino, and P. Carrión, "Influence of normalization and color space to color texture classification," Pattern Recognit. **61**, 120–138 (2017).

[8] M. Chirimuuta and F. A. A. Kingdom, "The uses of colour vision: Ornamental, practical, and theoretical," Minds Mach. **25**, 213–229 (2015).

[9] C. Clausen and H. Wechsler, "Color image compression using pca and backpropagation learning," Pattern Recognit. **33**, 1555–1560 (2000).

[10] CIE, "Recommendations on uniform color spaces, color-difference equations, psychometric color terms," Paris: CIE, (1978).

[11] R. D. Cook, "Detection of influential observation in linear regression," Technometrics **19**, 15–18 (1977).

[12] A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque," The J. Physiol. **357**, 241–265 (1984).

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2009), pp. 248–255.

[14] M. Engilberge, E. Collins, and S. Süsstrunk, "Color representation in deep neural networks," *Proc. IEEE Int'l. Conf. on Image Processing* (IEEE, Piscataway, NJ, 2017), pp. 2786–2790.

[15] D. H. Foster, I. Marín-Franch, S. Nascimento, and K. Amano, "Coding efficiency of cie color spaces," *Proc. IS&T/SID CIC16: Sixteenth Color Imaging Conference* (IS&T, Springfield, VA, 2008), pp. 285–288.

[16] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," *Int'l. Conf. on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 48–56.

[17] K. R. Gegenfurtner and L. A. Sharpe, *Color Vision* (Cambridge University Press, Cambridge, UK, 1999).

[18] T. Gevers and A. W. Smeulders, "Color-based object recognition," Pattern Recognit. **32**, 453–464 (1999).

[19] R. Gil Rodríguez, J. Vazquez-Corral, and M. Bertalmío, "Color matching images with unknown non-linear encodings," IEEE Trans. Image Process. **29**, 4435–4444 (2020).

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[21] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," Pattern Recognit. **40**, 1106–1122 (2007).

[22] J. W. Kalat, "Biological psychology," *Nelson Education* (Cengage Learning, MA, 2015).

[23] H.-K. Kim, J. H. Park, and H.-Y. Jung, "An efficient color space for deep-learning based traffic light recognition," J. Adv. Transportation **2018** (2018).

[24] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," Foundations and Trends in Machine Learning **12**, 307–392 (2019).

[25] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 9404–9413.

[26] J. Koenderink, A. v. Doorn, and K. Gegenfurtner, "Colors and things," i-Perception **11**, 1–43 (2020).

[27] J. Koenderink and A. J. van Doorn, *Perspectives on Colour Space* (Oxford University, New York, 2003).

[28] V. Laparra, S. Jiménez, G. Camps-Valls, and J. Malo, "Nonlinearities and adaptation of color vision from sequential principal curves analysis," Neural Computation **24**, 2751–2788 (2012).

[29] K. Larbi, W. Ouarda, H. Drira, B. B. Amor, and C. B. Amar, "Deepcolorfasd: Face anti spoofing solution using a multi channeled color spaces cnn," *Proc. IEEE Int'l. Conf. on Systems, Man, and Cybernetics* (IEEE, Piscataway, NJ, 2018), pp. 4011–4016.

[30] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 6874–6883.

[31] S. Laughlin, "A simple coding procedure enhances a neuron's information capacity," Z. Natforsch. c **36**, 910–912 (1981).

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *Proc. European Conf. on Computer Vision* (Springer, Cham, 2014), pp. 740–755.

[33] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 3730–3738.

[34] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The Bulletin of Mathematical Biophysics **5**, 115–133 (1943).

[35] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Photorealistic style transfer with screened poisson equation," *Proc. The British Machine Vision Conf.* (BMVA Press, UK, 2017), pp. 153.1–153.12.

[36] D. Mishkin, N. Sergievskiy, and J. Matas, "Systematic evaluation of convolution neural network advances on the imagenet," Comput. Vis. Image Underst. **161**, 11–19 (2017).

[37] N. Moroney and M. D. Fairchild, "Color space selection for jpeg image compression," J. Electronic Imaging **4**, 373–382 (1995).

[38] A. Mosleh, A. Sharma, E. Onzon, F. Mannan, N. Robidoux, and F. Heide, "Hardware-in-the-loop end-to-end optimization of camera image processing pipelines," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 7529–7538.

[39] R. Muñoz-Salinas, "A Bayesian plan-view map-based approach for multiple-person detection and tracking," Pattern Recognit. **41**, 3665–3676 (2008).

[40] C. A. Parraga and A. Akbarinia, "Nice: A computational solution to close the gap from colour perception to colour categorization," PloS one **11**, e0149538 (2016).

[41] W. Pratt, *Digital Image Processing*, 4th ed. (John Wiley & Sons, Hoboken, NJ, 2007).

[42] G. Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," Pattern Recognit. **35**, 1675–1686 (2002).

[43] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," IEEE Comput. Graph. Appl. **21**, 34–41 (2001).

[44] M. Rabbani, "Jpeg2000: Image compression fundamentals, standards and practice," J. Electronic Imaging **11**, 286 (2002).

[45] I. Rafegas and M. Vanrell, "Color encoding in biologically-inspired convolutional neural networks," Vis. Res. **151**, 7–17 (2018).

[46] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems* (Curran Associates Inc., NY, 2016), pp. 2234–2242.

[47] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," Color Res. Appl. **30**, 21–30 (2005).

[48] R. Starosolski, "New simple and efficient color space transformations for lossless image compression," J. Vis. Commun. Image Represent. **25**, 1056–1063 (2014).

[49] H. Stokman and T. Gevers, "Selection and fusion of color models for image feature detection," IEEE Trans. on Pattern Anal. Machine Intell. **29**, 371–381 (2007).

[50] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *IEEE Information Theory Workshop* (IEEE, Piscataway, NJ, 2015), pp. 1–5.

[51] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," *Int'l. Conf. on Learning* (ICLR, CA, 2017).

[52] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Advances in Neural Information Processing Systems* (Curran Associates Inc., NY, 2017), pp. 6306–6315.

[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).
[54] S. Wuerger and K. Xiao, *Color Vision, Opponent Theory* (Springer, Berlin, 2016), pp. 413–418.
[55] G. Wyszecki and W. S. Stiles, *Color Science* (Wiley, New York, 1982), Vol. 8.
[56] G. Yu, T. Vladimirova, and M. N. Sweeting, "Image compression systems on board satellites," Acta Astronaut. **64**, 988–1005 (2009).
[57] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 1058–1067.