



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/220375/>

Version: Accepted Version

Proceedings Paper:

Gilooly, Thomas, Thomas, Jean-Baptiste, Hardeberg, Jon Yngve et al. (2025) Image Adaptation for Colour Vision Deficient Viewers Using Vision Transformers. In: IEEE/CVF Winter Conference on Applications of Computer Vision 2025. IEEE/CVF Winter Conference on Applications of Computer Vision 2025, 28 Feb 2025 - 04 Mar 2026 IEEE Winter Conference on Applications of Computer Vision. IEEE, USA, pp. 5646-5655.

<https://doi.org/10.1109/WACV61041.2025.00551>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Image Adaptation for Colour Vision Deficient Viewers Using Vision Transformers

Tom Gillooly¹ Jean-Baptiste Thomas^{1,4} Jon Y. Hardeberg¹ Giuseppe Claudio Guarnera^{2,3}

¹NTNU, Norway ²University of York, UK ³Lumirithmic, UK ⁴Université de Bourgogne, France

thomas.b.gillooly@ntnu.no, Jean-Baptiste.Thomas@u-bourgogne.fr,

jon.hardeberg@ntnu.no, claudio.guarnera@york.ac.uk

Abstract

Colour Vision Deficiency (CVD) occurs when anomalous retinal cone spectral responses impact the ability to distinguish between certain colours. To enhance image quality and viewing experience, recolouring algorithms seek to modify pixel values so that this does not lead to a loss of detail or image quality. Recent approaches to recolouring for CVD viewers employ neural models which exploit higher order features to direct colour adaptation. In this work, we build upon the idea that visual neural models exhibit emergent behaviour which mimics the human visual system. We make use of these learned behaviours to guide the colour adaptation process by considering regions of the image that are the most semantically meaningful for a non-CVD viewer and compensate for them appropriately if they are absent or distorted in a CVD-simulated version of the image. We find that a minimal algorithm built atop a pre-trained model produces results that substantially boost contrast and salience for viewers affected by CVD. We also investigate a few cases where modifications are absent, indicating that a neurally guided salience-based model may also provide a means of determining when recolouring is not necessary. Additionally, we introduce a novel metric that quantifies the contrast increase or decrease under changes in image colour.

1. Introduction

The three types of cone photoreceptors in the human eye cover different spectral bands; long (L), medium (M), or short (S) wavelength. Their spectral responses determine our ability to distinguish colour therefore anomalies in these responses can impact our ability to discern certain colours, termed Colour Vision Deficiency (CVD) [24]. Protanopia, deuteranopia, and tritanopia refer to the different CVD types associated with anomalous L-, M-, or S-cones, respectively. Protanopic and deuteranopic CVD both correspond to impacted ability to distinguish colours which

differ along the red-green axis, and tritanopia to colours along the blue-yellow axis. When two regions of impacted distinguishability neighbour one another in an image, this can affect perceived image clarity and quality [21]. Recolouring algorithms for viewers with Colour Vision Deficiency (CVD) aim to improve image quality and viewing experience when the viewer has impacted sensitivity to particular colour channels. Methods for recolouring images involve modifying pixel values to compensate for these deficient cone responses on either a per-pixel basis, or by considering relative intensities in a local neighbourhood [19]. However, such methods may overlook more abstract image content. Evidence suggests that context can play a greater role in image classification and recognition than colour alone [6], indicating that colour adaptation algorithms should incorporate this broader context. Further, a key challenge in recolouring images for CVD viewers is enhancing contrast while preserving naturalness. Maximising contrast by indiscriminately modifying pixel values can lead to unnatural-looking results. Therefore, the process must be constrained to ensure that modifications are performed selectively. Modern deep learning architectures not only extract image features but also establish their mutual relevance using attention mechanisms [3]. Since these features correspond to high-level image content, they indicate how well image content is preserved after recolouring. Attention mechanisms further guide this process by determining which features should be emphasised and which can be safely disregarded (see Fig. 1).

1.1. Attentional methods for image recolouring

This work aims to develop a recolouring method that enhances contrast based on the importance of image regions. Attention methods naturally lend themselves to this process, as they assess the similarity of feature pairs to determine their contribution to the output [27]. The Vision Transformer (ViT) architecture [3] extends this method to the image domain, where the computed attention maps can be interpreted as an indication of which image regions are most

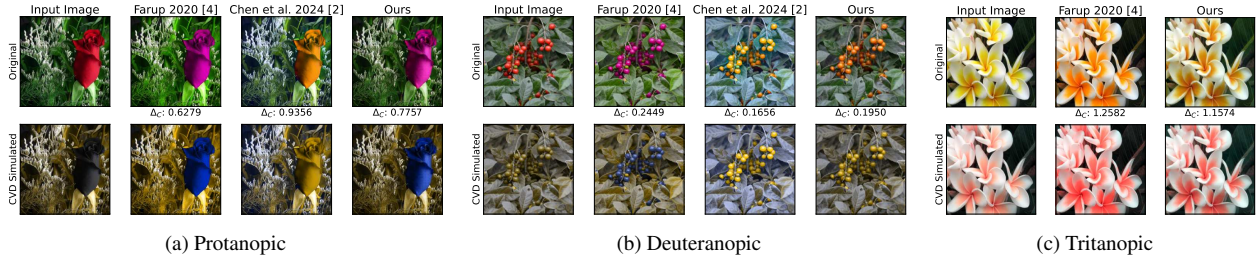


Figure 1. Examples of our method used to recolor images to improve contrast under protanopic, deuteranopic, and tritanopic colour vision deficiency, from left to right. The top row shows the unaltered images and their recoloured versions using three different methods, while the bottom row shows their CVD simulated versions i.e. as they would appear to a viewer with colour vision deficiency. In each case, the recolouring process has modified image regions so that they stand out more when colour vision deficiency is simulated by removing pixel intensity in colour channels corresponding to the affected cone response.

critical for the target application. Since no ground truth exists for recolouring for CVD viewers, one option is to repurpose a ViT model trained for a tasks like image classification. Since it is unclear how much correspondence there is between a specific task like image classification and image recolouring, another option is to use a model which has been trained to produce general-purpose features. The ViT trained with DiNO [1] is trained unsupervised and produces emergent features that can be adapted to specific downstream tasks via transfer learning. However, in this work we do not perform such specialisation. Instead, we evaluate the effectiveness of these general, emergent features for recolouring, without a specific objective or recolouring ground truth. We further note that emergent properties of neural networks have been shown to mimic the human visual system [5, 25, 29], suggesting that features deemed relevant by a model can be assumed to be sufficiently similar to those which would be relevant to a human observer, with a ViT’s attention weights quantifying this relevance.

1.2. Contributions

The main contributions of our work are:

- Leveraging emergent properties of a pre-trained network to highlight salient features of image, allowing for more context-aware recolouring;
- Using saliency to guide an adjoint method for recolouring for CVD, which considers overall semantic content, thus improving the visual experience for viewers.
- Proposing a new metric that combines local contrast analysis with a structure-aware approach to better reflect perceived contrast changes and naturalness in comparison to existing metrics, providing a quantitative measure for assessing the effectiveness of recoloring methods.

We have made our code available on [GitHub](#)¹.

¹The GitHub link for our code has been removed for peer review.

2. Related work

Recolouring for colour vision deficient viewers Common recolouring methods typically identify colours in the image located along what are known as confusion lines (i.e. colours on the gamut which differ only by power in missing cones [11]), and adjust them to enhance distinguishability for subjects with deficient cone responses.

However, modifying pixel values without considering spatial context can lead to unnatural-looking images. To mitigate this issue, methods such as [4, 9, 13, 31] incorporate energy constraints to preserve the naturalness and contrast during optimisation. While these approaches consider neighboring pixels, they primarily focus on raw pixel values. Consequently, although factors such as edge intensity and image smoothness are considered, they may not fully capture more abstract image content.

Deep learning for recolouring Prior work using deep learning techniques for recolouring include [18], which combines traditional image processing techniques with verification by a CNN, and DeepCorrect [17], which trains a GAN-style network for colour correction, using a loss derived from a network trained for image classification. While these techniques incorporate pre-trained networks for verification, they essentially act as black boxes, limiting interpretability. The method in [14] uses diffusion to build a saliency map, boosting contrast in a CVD image, then converts it back to a colour image using CycleGAN. While the concept of building a saliency map is similar to the approach taken in this work, we replace the handcrafted saliency map with a neural network output. Unlike these works, which require networks specifically trained for recolouring, our approach sidesteps this necessity, avoiding the need for GANs, which can be difficult and unstable to train. GANs are also used in [10], where cyclical consistency loss preserves image content while altering style. This approach regularises the latent space such that it can be traversed in specific di-

mensions to yield images with varying degrees of recolouring, corresponding to CVD severity. However, as the architecture is decoder-only, it cannot recolour specific input images. In [2], a Transformer architecture generates recoloured images from unaltered inputs, using an unsupervised approach that aims to preserve image contrast and naturalness by balancing objectives pertaining to image contrast and structural similarity between the input and modified images. The encoder-decoder model proposed in [8] converts input images to recoloured versions that minimise absolute colour difference after CVD simulation. As in [2] the loss function is per-pixel, with no emphasis on visually salient regions.

Evaluation metrics Metrics for quantitative evaluation in prior work include chromaticity difference, global contrast, local contrast [2], and histogram distance [10]. However, these metrics often discard relevant information, such as lightness or image structure, and may not adequately account for perceptual distance, making them insufficient. Further discussion is provided in the Supplementary Material. In [10] the authors use the VGG model [23] to express a perceptual loss for the purposes of quantitative evaluation, a method commonly used to evaluate generative models as in *e.g.* [7], where features taken from the final stages represent image content and therefore taking their difference quantifies content difference. At training time, the VGG model uses an RGB shift augmentation, which suggests that the model could extract features which are invariant to colour shifts within the image. However, [12] (from which the RGB shift augmentation in VGG is taken) states that this augmentation simulates variations in illumination intensity and colour, to which the network should be invariant. The augmentation described is applied uniformly across the whole image. To boost contrast post-CVD simulation, colours of certain regions must change independently, *i.e.*, non-uniformly. Therefore, the VGG network cannot be guaranteed to be invariant to the necessary colour changes and could report a perceptual loss when in actuality there is none. Further, the training dataset for the VGG network is ImageNet-1k, consisting of natural images. In many real-world applications the images to be processed may not be natural images. Rather, they could be artworks or have unique stylistic properties. In these out-of-domain cases it is unclear how the VGG model will perform.

While using VGG to evaluating image content consistency is reasonable, further study is needed to ensure its reliability under the proposed transformations. Without confirming its invariance, it is unclear how much of the result is due to variations in the VGG model performance versus changes in image content we wish to measure.

3. Method

In this work, we leverage the emergent properties observed in self-supervised neural networks [5, 25] along with the intuitive correspondence to saliency employed by attention-based neural network architectures. We use a pre-trained Vision Transformer model trained with DINO and use the output attention maps to guide an adjoint recolouring method, ensuring saliency is preserved for both normal and deficient colour vision.

3.1. Workflow

To determine how a CVD viewer would perceive an image, we use the simulation matrices M as defined in [15]. The image $I_{CVD} = MI$ then refers to the input image I as perceived by CVD viewer. Similarly, $I_d = MI_r$ refers to the CVD simulation of the recoloured image I_r . The overall workflow for our proposed method is shown in Figure 2. The aim is to converge to a recoloured RGB image I_r that best preserves the saliency maps of the original RGB image I once CVD simulation has been applied.

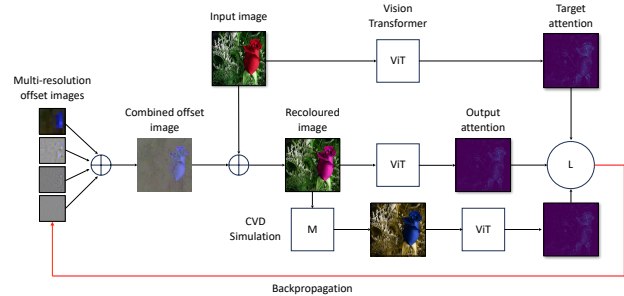


Figure 2. Full workflow for the daltonisation process. A breakdown of each step is in Sec. 3.1. The goal is to modify the image offsets so that the attention maps output by the Vision Transformer agree. Backpropagation updates the pixels of the offset images rather than the parameters of the Transformer network.

We achieve this by using an adjoint method to find an offset image ΔI that minimises the objective function \mathcal{L} :

$$\mathcal{L} = \sum_{\ell=1}^L L_a(I_r, I, \ell) + L_a(I_d, I, \ell) \quad (1)$$

The first loss term expresses the difference between the logarithm of the attention maps (*i.e.* the ViT outputs) for the recoloured image I_r and the input image I , while the second expresses the same for the recoloured image’s CVD simulation I_d and the input image I . The subscript ℓ indicates the layer ℓ from which the attention map is taken. The summation is across the first L attention layers of the transformer model. In our work we use the first four layers. The attention loss function L_a is defined as:

$$L_a(I_0, I_1, \ell) = \alpha_1^\ell L_2(\log \alpha_0^\ell, \log \alpha_1^\ell) \quad (2)$$

Where L_2 is the mean squared error and α_i^ℓ is the output of attention block ℓ of DINO for input I_i . We use the small version of DINO with 6 attention heads and patch size 8×8 .

We compute the recoloured image $I_r = \sigma(I + \Delta I)$, where σ is the hard sigmoid function, clamping the range of I_r to $[0, 1]$. This prevents the optimisation routine from boosting the image unrealistically by adding image intensity outside the final pixel value range.

The offset image ΔI is constructed recursively from a sequence of offset maps $\{I_m \mid M \geq m > 0\}$ of different resolutions:

$$\Delta I = \Delta I_M + g(\Delta I_{M-1}, H_M, W_M) \quad (3)$$

The function g is an upscaling function that interpolates an offset image $\Delta I_{m-1} \in \mathbb{R}^{H_{m-1} \times W_{m-1}}$ to the size of the offset image of the next highest resolution, i.e., $H_m \times W_m$. In our method we use four resolutions of ΔI_m ; 8×8 , 16×16 , 32×32 and 64×64 , each with three colour channels. Bilinear interpolation is used for the function g .

After each optimisation step, we subtract the average values of each ΔI_m so that the mean of the offset image is zero. Empirically, this helps prevent changes to the overall colour cast of the image. The impact of this step is further discussed in our ablation study (Sec. 4.2).

3.2. Dataset

Following prior work [2, 10], we use subsets of the *Oxford Flowers* (~8100 images) [16], *Places365* (~41000 images) [30], *Abstract Art* (~8100 images) [26], and *WikiArt Abstract* (~4400 images), *Still Life* (~2500 images), and *Landscape* (~12000 images) [20] datasets. As the images show different degrees of contrast loss under CVD simulation, we resample the datasets to ensure balanced contrast loss, following an approximately Gaussian distribution. Further details are in the Supplementary Material.

3.3. Structured local contrast metric

To evaluate our method, we quantify the contrast improvement in the modified CVD image over the unmodified one, i.e., retention of image structure while undergoing a shift in overall colour varying with CVD severity. To overcome limitations of existing metrics (Sec. 2), we adopt a sliding window local contrast approach and treat the outputs as high-dimensional features to derive a structurally-aware distance between image regions. Details can be found in the Supplementary Material, but at a high level we decompose the local contrast difference into a similarity and dissimilarity term, where the similarity counts positively towards the image score and the dissimilarity counts negatively. This decomposition means that if contrast is augmented over

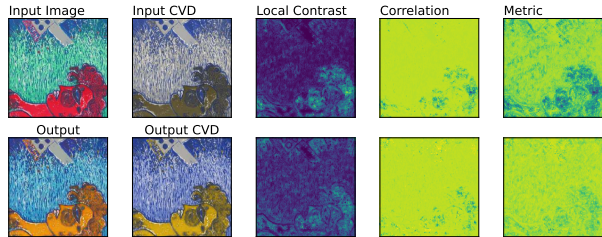


Figure 3. Visualisation of our proposed metric. The top row shows the unaltered image and its CVD simulation, while the bottom row shows the recoloured image and its CVD simulation. The local contrast image shows the absolute difference between the local contrast of the unaltered input and each of the CVD simulations. The metric images shows the same for our proposed metric. The correlation image represents the correlation between the absolute difference and our proposed metric. In the altered image, the local contrast highlights large differences the centre, while the metric image shows high similarity in this region, thus the difference is due to the contrast being enhanced. Note that augmenting contrast is not always desirable as it can give an unnatural result, which is why we incorporate naturalness into the metric.

the input image it is not penalised, while the dissimilarity term means that deviation from the input image structure is treated as noise and penalised. Figure 3 shows this effect. Contrast is measured as the Euclidean distance in CIELAB space, accounting for its perceptual non-uniformity. The total difference metric is the per-pixel difference of the two local contrast images.

3.3.1 Evaluating naturalness

Defining naturalness is challenging. We assume that input images are natural, and small perturbations minimally affect this quality. Therefore, an image with fewer perturbations is considered more natural than one with greater alterations. We quantify the perturbation size as the mean energy, i.e. the mean Frobenius norm of the offset image ΔI .

Note that we have a *relative* measure of naturalness; we cannot report the naturalness of a single image, but we can state that one image is more natural than another. Hence when comparing different methods in Sec. 4.1 and Sec. 4.2, we scale the contrast improvement score by the ratio of the energy of each image so that the final scaled metric value incorporates both contrast improvement and naturalness.

3.4. CVD simulation and impact on saliency

Figure 4 shows a sample image under CVD simulation with varying degrees of severity using the transformation matrices from [15], while Figure 5 shows the attention maps for each severity level.

The 8, 12, and 20nm labels represent the wavelength deviation of the cone spectral response curve, correspond-

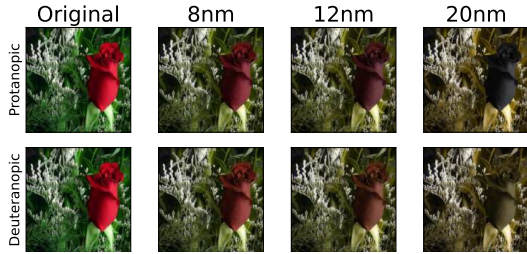


Figure 4. Sample image with CVD simulation applied with varying degrees of severity. The 8, 12, and 20nm shifts refer to the degree of cone response anomaly and correspond to mild, moderate, and high severity, respectively.

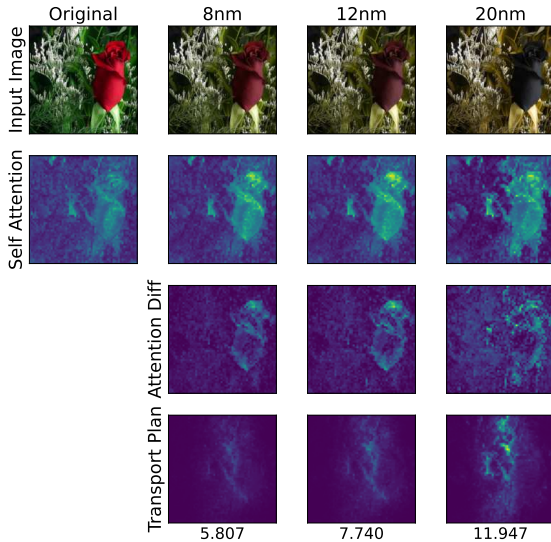


Figure 5. Impact on saliency for a sample image for a viewer with protanopia. The saliency of the rose becomes more dispersed as the severity of CVD increases.

ing to mild, moderate, and high severity, respectively. We selected these values in line with the severity levels used in [10]. The colour difference between the rose and its background decreases with increasing severity, and a corresponding dispersion is observed in the attention maps as the colour features which contribute to saliency deteriorate. The bottom row shows the Wasserstein-2 distance [28] between attention maps, representing the transport plan to move attention weights in the CVD attention map to match those of the unmodified image.

4. Results

This section presents the results of our algorithm Section 3.1, comparing it with the deep learning approach in [2] and the non-deep learning method in [4]. We selected the former on the grounds of its recency and architectural sim-

ilarity i.e. both our method and that of [2] use a Vision Transformer. We selected the latter as it is an improvement on [22] which at time of publication was state of the art.

4.1. Recolouring results

Table 1 presents the results for each method on each tested dataset across the three CVD types and severities. Detailed results for each CVD type and severity can be found in the Supplementary Material.

The columns in Tab. 1 show the raw contrast increase (Δ_L), the mean offset image energy (Δ_N) and the combined score (Δ_C). Δ_C is calculated by multiplying Δ_L by the ratio of our method’s Δ_N to the compared method’s Δ_N , thus incorporating a naturalness measure into the contrast score, as discussed in Sec. 3.3. According to the combined score Δ_C , our method performs best across all datasets except *Flowers*. This is likely because our method refrains from altering images where no significant saliency-driven modification is necessary. We observed that many images in the *Flowers* dataset required no change, as the unmodified input image retained sufficient contrast and general saliency was unaffected. Some of these examples are shown in Figure 6. This lack of modification impacts our method’s metric, as it results in zero contrast boost for these images, lowering the average score. While the approach from [2] performs best on the *Places365* dataset, and our method ranks second, it is important to note that [2] was specifically trained on this domain, tuning the model for optimal performance on this dataset. In contrast, our model is not tailored to any dataset but still approaches the performance of [2]. In terms of raw contrast score Δ_L , the other methods outperform ours. However, our method consistently applies the smallest change to the original image to achieve the recolouring result. This is expected, as our approach is guided by saliency, meaning not every pixel is considered salient, and thus, not every pixel will be modified, leading to less overall change. In contrast, the methods in [2] and [4] apply modifications indiscriminately to every pixel, resulting in greater overall changes, even in areas where the impact on image saliency is minimal.

Figure 7 shows some results under protanopic colour vision deficiency, i.e. diminished sensitivity in the red wavelengths. In all cases, the method of [4] results in perhaps the largest contrast boost but at the expense of large changes in the original image.

Figure 8 shows sample results for deuteranopic colour vision deficiency. As in the protanopic case, our method increases contrast with minimal changes. Notably, in the second image, both our method and [4] enhance the blue channel in the flower petals, but ours confines modification to the flower, the image’s focal point, leaving the background unaffected. In contrast, [4] alters the background intensity.

Figure 9 shows sample results for tritanopic colour vi-

Dataset	Ours			Swin ViT [2]			Halo-Free [4]		
	Δ_L	Δ_N	Δ_C	Δ_L	Δ_N	Δ_C	Δ_L	Δ_N	Δ_C
<i>Abstract Art</i> [26]	0.3215	0.0519	0.3215	0.7018	0.1588	0.2986	0.4319	0.1128	0.2341
<i>Flowers</i> [16]	0.1716	0.0614	0.1716	0.5564	0.1938	0.2443	0.6295	0.1234	0.3777
<i>Places365</i> [30]	0.4792	0.0616	0.4792	0.8998	0.1476	0.5095	0.3307	0.0832	0.2821
<i>WikiArt Abstract</i> [20]	0.2541	0.0456	0.2541	0.5440	0.1411	0.2243	0.1325	0.0976	0.0748
<i>WikiArt Landscape</i> [20]	0.4787	0.0348	0.4787	0.5807	0.1005	0.2407	-0.0180	0.0670	0.0008
<i>WikiArt Still Life</i> [20]	0.4307	0.0375	0.4307	0.5722	0.1198	0.2249	-0.0448	0.0700	-0.0434

Table 1. Results for each dataset, averaged across all CVD types and severities. Δ_L represents the raw score from our proposed metric (Sec. 3.3, reflecting contrast alone and not accounting for naturalness, thus not a complete performance indicator). Δ_N is the mean energy of the offset image, quantifying the perturbation required to produce the recoloured image. Lower energy indicates lower perturbation and a more natural recoloured image. Δ_C is Δ_L scaled by the ratio of Δ_N , reflecting both contrast enhancement and naturalness.

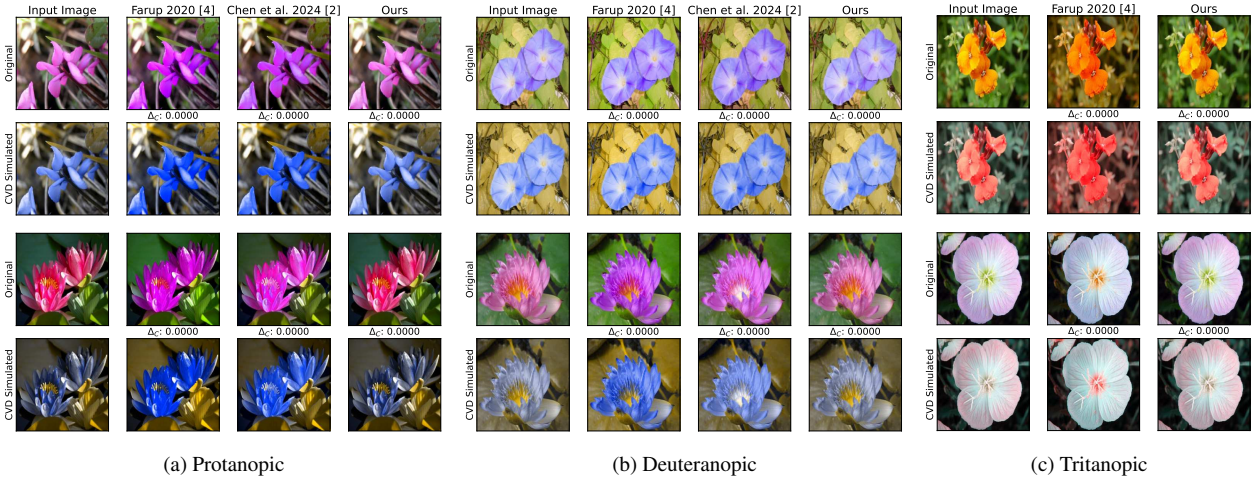


Figure 6. Example images under protanopic (6a), deuteranopic (6b) and (6c) that remain unchanged by our recoloring workflow. As our method only highlights regions it deems salient and there is still sufficient contrast post-CVD simulation, it makes no alterations to the input image.

sion deficiency. As no dataset exists for tritanopic CVD, there are no results for the Swin Transformer method of [2]. Hence, we compare only with [4]. Similar to the other cases, [4] increases contrast but significantly alters the image colour, affecting the naturalness of the result.

More examples are in the Supplementary Material.

4.2. Ablation study

In this section, we examine the impact of removing three key elements from the pipeline: multiresolution images, bias removal from the total offset image, and combining the attention map losses for multiple attention layers within the transformer model. We briefly describe each ablated element and summarise the resulting impact on contrast improvement. Further detail and example images for each ablated element are provided in the Supplementary Material.

Bias removal Since we do not apply regularisation to the offset images, we found that they often exhibit bias drift,

appearing as a colour shift in the output image. While including the bias sometimes improves metric score despite the colour shift, we found that removing bias generally enhances contrast without introducing a colour shift. Example images are available in the Supplementary Material.

Multi-resolution offsets As discussed above, the ViT operates on image crops of size 8×8 . When optimising for an offset map using only the maximum resolution level, the resulting image showed block artifacts comparable in size to the 8×8 patches that the model was trained on. Using lower resolution offsets avoids high frequency noise, but using a single image of the highest resolution possible while still avoiding block artifacts does not produce results of the same quality as combining multiple resolutions, as the model cannot produce fine detail in the final offset image. Producing the best quality image is then a matter of balancing high frequency noise against high frequency detail. Examples showing the visual impact of different resolution levels can

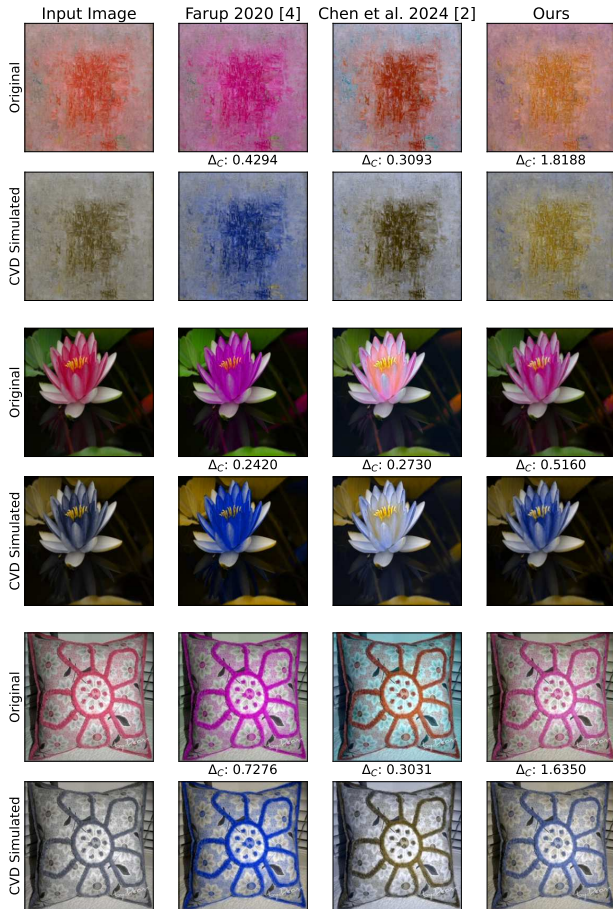


Figure 7. Sample results for protanopic CVD at high severity. Our method does not boost the contrast as much as [4], but there is less overall image change which corresponds to a higher naturalness.

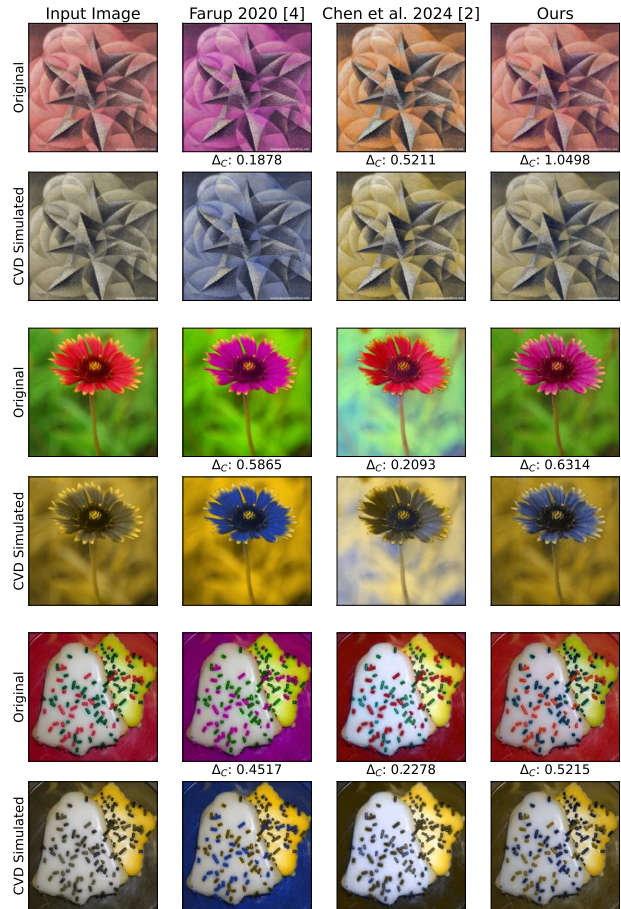


Figure 8. Sample results for severe deuteranopic CVD. Our method balances contrast boost against input image perturbation. Notably, in the flower image, the flower is highlighted as the most salient object while the background remains relatively unchanged.

be found in the Supplementary Material.

Limiting attention layers The Vision Transformer model consists of a patch feature extractor followed by 12 attention layers. In our final ablation, we limit the number of attention layers over which we apply the loss function (Eq. (1)).

While reducing the number of attention layers does not necessarily produce disruptive artifacts, it leads to less contrast improvement. Example images with varying numbers of attention layers are provided in the Supplementary Material.

Attn Layers		✓		✓		✓		✓
Multires			✓	✓			✓	✓
Bias removal					✓	✓	✓	✓
<i>Abstract Art</i> [26]	-0.8926	0.3401	-0.1212	0.3527	-7.5866	-1.4343	-1.8622	0.4628
<i>Flowers</i> [16]	-2.6534	0.0921	-0.3161	0.1740	-10.8010	-0.1242	-1.2945	0.2084
<i>Places365</i> [30]	-0.6078	0.9186	0.3461	0.7379	-4.2247	0.8928	-0.3236	0.9164
All	-1.3681	0.4566	-0.0257	0.4257	-7.4955	-0.2185	-1.1553	0.5350

Table 2. Ablation study results for three datasets of different image categories under protanopic CVD with high severity. Pipeline choices affect datasets differently, but on average, the method with all options active produces the best average result across all datasets

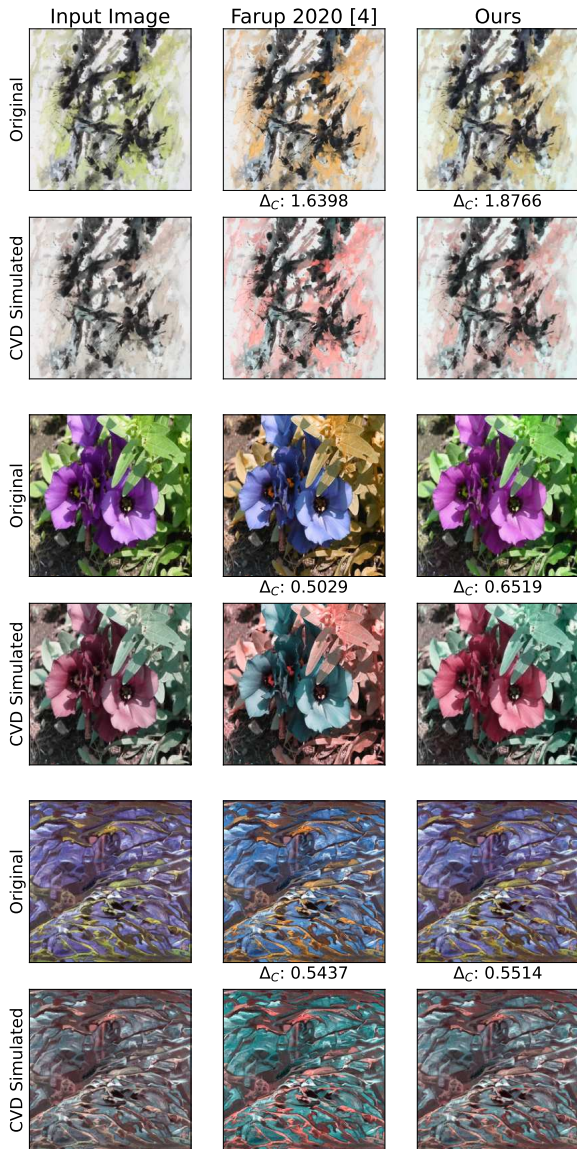


Figure 9. Sample results for tritanopic CVD at high severity. The flower image shows how our method boosts contrast without drastically changing the input image content.

4.2.1 Ablation results

The results of the ablation study are shown in Tab. 2. The top portion of the table shows which parts of the pipeline are active. The bottom row shows the average across all the datasets. The leftmost column shows the results for a single attention layer in the loss function, retaining bias, and using a single resolution offset image. We select three of our chosen datasets for evaluation to test across different types of images. The results show that some architectural choices benefit certain datasets while others do not. For example,

the *Places365* dataset yields better results using a single resolution offset image and preserving bias (though a disruptive colour cast from bias retention is not always reflected in the metric score). The non-ablated pipeline achieves the best results across all datasets, suggesting it is the best general-purpose choice. However, for specific datasets, disabling certain options may yield better outcomes.

5. Conclusion

We have demonstrated a method for recolouring images by saliency preservation using a pre-trained Vision Transformer model. Our results shown that attention-based models that are trained on broad datasets learn regularities that can be used to solve the recolouring problem, even in the absence of a formal ground truth. The generality of the foundation model means that our method shows good baseline performance across a range of datasets. Additionally, since the raw output features from the original model can be used to for video segmentation without fine-tuning, our method should also be applicable to recolouring video data, though this remains to be evaluated in future work.

The model generally enhances energy in colour channels not attenuated for those affected by CVD. Modifications are computed using an adjoint method with gradient descent on a per-image basis, but this approach could also generate data to train a model for direct single-step modifications. This work focuses on leveraging self-supervised models to exploit emergent properties that mimic those of the human visual system. By not relying on expensive annotated data, our method offers greater flexibility and could be extended to images of different modalities, such as the multispectral domain, where there may not be enough annotated data to train a supervised model.

We see this work as a first step towards solving the recolouring problem with foundation models, establishing a baseline for future methods that use the same methodology. By evaluating on a broad range of datasets, we hope that this work will serve as a benchmark for future research.

In addition, we addressed the shortcomings of metrics used to assess recolouring in prior work and proposed a novel metric that quantifies contrast increase or decrease through a colour change transformation. This metric will be valuable for better assessing recolouring methods in future studies. However, further subjective evaluation is needed to determine how well this metric aligns with perceived contrast for a larger population of human observers.

Acknowledgments This research has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 814158 and by the BBSRC grant BB/X01312X/1.

References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2
- [2] Ligeng Chen, Zhenyang Zhu, Wangkang Huang, Kentaro Go, Xiaodiao Chen, and Xiaoyang Mao. Image recoloring for color vision deficiency compensation using swin transformer. *Neural Computing and Applications*, pages 1–16, 2024. 3, 4, 5, 6
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1
- [4] Ivar Farup. Individualised halo-free gradient-domain colour image daltonisation. *Journal of Imaging*, 6(11):116, 2020. 2, 5, 6, 7
- [5] Roland W Fleming and Katherine R Storrs. Learning to see stuff. *Current opinion in behavioral sciences*, 30:100–108, 2019. 2, 3
- [6] Brian Funt and Ligeng Zhu. Does colour really matter? evaluation via object classification. In *Color and Imaging Conference*, volume 26, pages 268–271. Society for Imaging Science and Technology, 2018. 1
- [7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3
- [8] Satyam Goyal, Kavya Sasikumar, Rohan Sheth, Akash Seelam, Taeyeong Choi, and Xin Liu. Encolor: Improving visual accessibility with a deep encoder-decoder image corrector for color vision deficient individuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23335–23342, Mar. 2024. 3
- [9] Mohd Fikree Hassan and Raveendran Paramesran. Naturalness preserving image recoloring method for people with red–green deficiency. *Signal Processing: Image Communication*, 57:126–133, 2017. 2
- [10] Shuyi Jiang, Daochang Liu, Dingquan Li, and Chang Xu. Personalized image generation for color vision deficiency population. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22571–22580, October 2023. 2, 3, 4, 5
- [11] Neda Milić Keresteš, Stefan urević, Dragoljub Novaković, Miroslav Zarić, Nemanja Kašiković, Sandra Dedijer, and Gojko Vladić. Customized daltonization: adaptation of different image types for observers with different severities of color vision deficiencies. *Universal Access in the Information Society*, pages 1–17, 2021. 2
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 3
- [13] Giovane R. Kuhn, Manuel M. Oliveira, and Leandro A. F. Fernandes. An efficient naturalness-preserving image-recoloring method for dichromats. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1747–1754, 2008. 2
- [14] Jinjiang Li, Xiaomei Feng, and Hui Fan. Saliency consistency-based image re-colorization for color blindness. *IEEE Access*, 8:88558–88574, 2020. 2
- [15] Gustavo M. Machado, Manuel M. Oliveira, and Leandro A. F. Fernandes. A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1291–1298, November/December 2009. 3, 4
- [16] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 4, 6, 7
- [17] Gajo Petrovic and Hamido Fujita. Deep correct: Deep learning color correction for color blindness. In *SoMeT*, pages 824–834, 2017. 2
- [18] Dhruv Rathee and Suman Mann. Daltonizer: A cnn-based framework for monochromatic and dichromatic colorblindness. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, pages 1–5, 2022. 2
- [19] Madalena Ribeiro and Abel J. P. Gomes. Recoloring algorithms for colorblind people: A survey. *ACM Comput. Surv.*, 52(4), aug 2019. 1
- [20] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016. 4, 6
- [21] Lindsay T Sharpe, Andrew Stockman, Herbert Jägle, and Jeremy Nathans. Opsin genes, cone photopigments, color vision, and color blindness. *Color vision: From genes to perception*, 351:3–52, 1999. 1
- [22] Joschua Thomas Simon-Liedtke and Ivar Farup. Multiscale daltonization in the gradient domain. *Journal of Perceptual Imaging*, 1(1):10503–1, 2018. 5
- [23] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14. Computational and Biological Learning Society, 2015. 3
- [24] Andrew Stockman and Lindsay T. Sharpe. *Human Cone Spectral Sensitivities and Color Vision Deficiencies*, pages 307–327. Humana Press, Totowa, NJ, 2008. 1
- [25] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021. 2, 3
- [26] Greg Surma. Abstract art dataset, 2019. version 1, accessed on 16 February 2024. 4, 6, 7
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

- [28] Cédric Villani. *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. [5](#)
- [29] Wei Yu, Kuiyuan Yang, Yalong Bai, Tianjun Xiao, Hongxun Yao, and Yong Rui. Visualizing and comparing alexnet and vgg using deconvolutional layers. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016. [2](#)
- [30] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. [4](#), [6](#), [7](#)
- [31] Zhenyang Zhu, Masahiro Toyoura, Kentaro Go, Issei Fujishiro, Kenji Kashiwagi, and Xiaoyang Mao. Processing images for red–green dichromats compensation via naturalness and information-preservation considered recoloring. *The Visual Computer*, 35:1053–1066, 2019. [2](#)