Supplementary Material - Image Adaptation for Colour Vision Deficient Viewers Using Vision Transformers

Tom Gillooly[1] Jean-Baptiste Thomas[1,4] Jon Y. Hardeberg[1] Giuseppe Claudio Guarnera[2,3]

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

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

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

This supplementary document contains additional detail relating to the paper Image Adaptation for Colour Vision Deficient Viewers Using Vision Transformers, specifically, a description of the dataset balancing strategy, an analysis of evaluation metrics used in prior works, a description of our local contrast evaluation metric, and examples of the visual impact of the ablations presented in the main work. The final sections show further recolouring results for each type of colour vision deficiency.

1. Dataset resampling

Some of the selected datasets (*e.g*. *Places365*) show greater average contrast loss than others under CVD simulation, so a naïve sampling method (*e.g*. taking the first n images when ordered by contrast loss) could result in an artificially inflated score for these datasets. Therefore, we resample the data to compare the recolouring procedure's performance across different datasets without conflating improved performance with large unmodified contrast loss. We apply our metric (discussed in Section 3) to determine the loss of contrast between each original image I and its CVD simulation I_{CVD}. We sample an approximate Gaussian distribution centred at a contrast loss of -2.5 with a standard deviation of 0.5, yielding minimum and maximum contrast losses of -4 and -1, respectively.

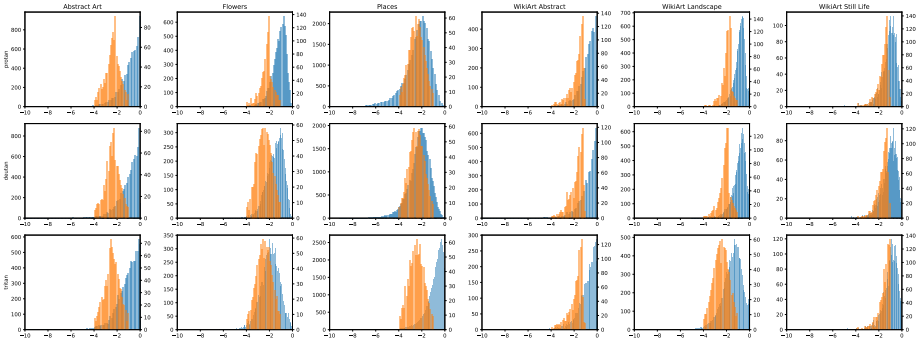


Figure 1. Contrast loss histograms for each dataset across different CVD types. The histogram for the full dataset is shown in blue, while the resampled histogram is shown in orange. Note that the two plots have different y axes. We sampled as closely to a Gaussian as possible, with 1000 samples per dataset.

Figure 1 shows the results of resampling. We bin the datasets by contrast loss and draw samples from each bin according to the Gaussian distribution described above. In cases where there are insufficient samples in a particular bin to draw the required amount to fit to a Gaussian, we push the leftover samples to the next highest bin and attempt to sample this combined amount.

2. Analysis of existing evaluation metrics

This section discusses evaluation metrics used in prior work and analyses their suitability for assessing contrast improvement across a recolouring transformation.

Chromaticity difference Coordinates in CIELAB space consist of three components: lightness (L^*), red-green colour value (a^*), and blue-yellow colour value (b^*). The chromaticity difference, as used in [1], is defined as the Euclidean norm in CIELAB space, excluding the lightness component. Ignoring the lightness component overlooks an aspect that significantly influences perceived colour difference. For instance, this means that light and dark blue would be considered identical according to the metric. Therefore, the chromaticity difference does not correspond to actual perceptual differences, making this metric an unreliable indicator of the viewing experience of a human observer.

Total colour contrast The total colour contrast of an image is determined by the combined mean of its global and local contrast [1]. Global contrast is computed by randomly sampling pixel pairs within an image and calculating their absolute difference. Local contrast represents colour variation statistics for pixels within a sliding window, averaging the absolute differences between each pixel and the centre pixel. The value calculated for each region is averaged to find the local contrast value across the entire image. As colour difference is affected by surround and the proximity of the two samples being considered [4], we find it questionable how representative the global contrast metric is of perceived contrast when sampled pixels might be distant from one another. While local contrast provides a clearer representation of local structure, averaging the result removes spatial information from the region under consideration, complicating image comparisons. For example, a region with an edge of fixed magnitude that is vertical in a test image and horizontal in a reference image would yield the same local contrast value, despite representing different structures.

Histogram distance The histogram distance metric used in [3] measures the Hellinger distance of an image's colour histogram before and after CVD simulation. This is a reasonable quantity to measure when considering the input and recoloured images, as it quantifies how much recolouring has changed the overall colour of the image, and can therefore serve as an indicator of image naturalness. However, it is less clear what value this affords when comparing the recoloured image and its CVD simulation. In this case colour changes are expected, *e.g.* applying protanopic simulation to an image with a red colour cast must necessarily change as energy in the red channel is attenuated. However, it does not follow that the quality of the result has deteriorated or that the result can be considered unnatural.

3. Structured local contrast metric

Our proposed evaluation metric computes local contrast and compares the outputs in terms of structural similarity. Contrast is taken as pixel difference in CIELAB space, following the local contrast metric used in [1]. Although Euclidean distance in CIELAB space does not perfectly conform to perceptual difference, there is better agreement than when using RGB pixel values. As our metric is based on local contrast, i.e., relative pixel difference between centre and neighbouring pixels, it is logical to work in the colour space where this distance better corresponds to perceptual difference according to a standard human observer.

First, we convert input images from RGB to CIELAB space and compute the local contrast tensor C:

$$C(I)_{i,j,k,l} = \theta_{k,l}\sqrt{(I_{m,n} - I_{m+k,m+l})^2} \tag{1}$$

where $I_{m,n}$ denotes the pixel at $(m,n) = (i+K, j+K)$ in the Lab-space image I, $\theta_{k,l}$ is element (k,l) of a 2D Gaussian kernel centred at pixel (m,n) with $l,k \in [-K,K]$ (i.e., the kernel size is $2K+1$).

The resulting output is an image $C \in \mathbb{R}^{+^{h\times w\times D}}$, representing the Euclidean distance of each pixel value to that of its neighbours in the support of the kernel θ, with $w = W - 2K$, $h = H - 2K$, and $D = (2K+1)^2 - 1$. The CIELAB colour space is not perfectly perceptually uniform: a pair of colours separated by a distance of $\Delta E = 1$ in one region of colour space may not have the same perceptual difference as another pair separated by the same distance elsewhere. A distance of 2.3 is typically used as an average Just Noticeable Difference (JND) [5]. This perceptual non-uniformity also means that quantitative comparisons grow inaccurate with increasing distance. Therefore, we clamp C to the interval $[2,5]$ and shift it down by 2, resulting in $\hat{C} \in [0,3]^{h\times w\times D}$

Treating $\hat{C}$ as single point in feature space and using the ℓ_2 norm of its distance from another point to determine the difference between two contrast images will not yield the desired behaviour, as equally distant points from a reference point

may not represent the same contrast difference. Each point represents a contrast pattern, so we must consider the meaning of the difference between two points. In this space, the origin $\mathbf{0}$ represents a completely uniform region. A test point t which is a scalar multiple of a reference point r represents the same structure with scaled contrast: a factor greater than one should correspond to a positive score (amplified differences), and less than one a negative score (diminished differences). Test points along the reference line intersecting $\mathbf{0}$ and r represent similar structures with varying contrast, and distance from this line indicates deviation from the reference structure.

We change the test point coordinates to define them in terms of components co-linear and orthogonal to the reference line. The co-linear component is scored positively if it is farther from the origin than the reference point, negatively if it is closer, or zero if it is the same. The orthogonal component should negatively affect the metric score. Given a reference point r and a test point t, we take the co-linear component u and the orthogonal component v:

$$\begin{aligned} u &= \frac{< r, t > r}{||r||^2} \\ v &= r - u \end{aligned} \tag{2}$$

The energy of the co-linear component represents the similarity of r and t. We subtract the energy of the reference point so that the similarity term is zero when $t = r$, and subtract the square of the orthogonal component energy to decrease the metric score as the size of the orthogonal component increases. The full structured local contrast metric $d(t, r)$ is:

$$d(t, r) = ||u|| - ||r|| - \frac{||v||^2}{(1 + ||r||)^2} \tag{3}$$

The orthogonal component is scaled by the inverse of the target energy, so that when the contrast of the reference is greater, deviations need to be larger to impact the score. The added 1 in the denominator handles the case where the reference energy is zero, i.e., the reference is a uniform region. The metric then represents a quadratic roll-off with distance from the reference line. We found heuristically that squaring the orthogonal energy term agrees better with the sample images tested, but subjective evaluation is required to verify general applicability.

We calculate the contrast difference of two images ΔC as:

$$\Delta C(I_0, I_1) = \frac{1}{hw} \sum_{i=0,j=0}^{h,w} d(\hat{C}(I_0)_{i,j}, \hat{C}(I_1)_{i,j}) \tag{4}$$

Using this measure, our goal is to find a recoloured image that better resembles the contrast structure of the input image I than the CVD simulation I_{CVD} of the latter. We compute the contrast boost of each image as $\Delta C(I_d, I) - \Delta C(I_{CVD}, I)$, where $I_{CVD} = MI$.

3.1. Visualisation of ablated results

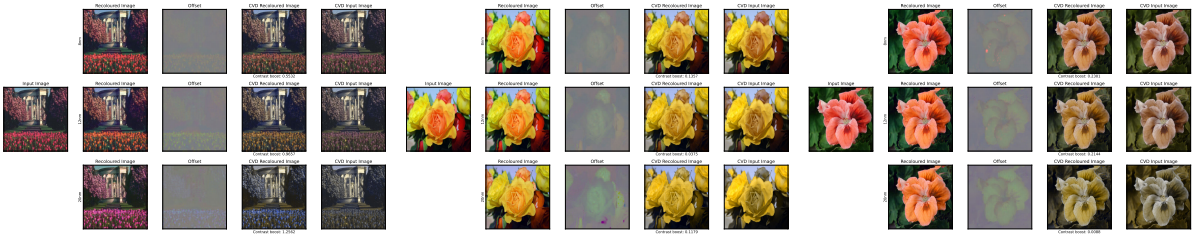
This section shows example output images for each of the ablated elements discussed in the ablated study; bias removal, multi-resolution offsets, and limited attention layers.

3.1.1 Bias removal

Figure 2 shows examples of the impact of bias removal on the output. Examples 1 and 2 have taken on a colour shift due to offset bias. In the first case, the image appears washed out, and in the second the image has taken on a green-blue hue, despite the contrast metric increasing from the full model which has the bias term removed. Example 3 shows an example where including the bias has resulted in a more vivid image and an increased metric score. However, even with bias removal, the metric score increases over the unmodified input, so we elect to remove bias for all images. Adapting the pipeline to include bias where helpful is left as an avenue for future research.

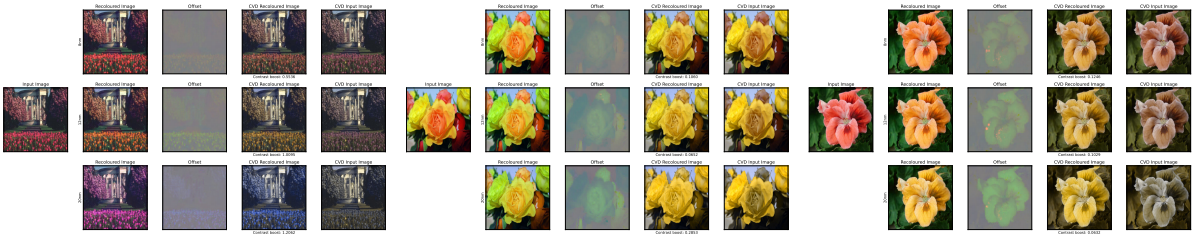
3.1.2 Multi-resolution offsets

Figure 3 shows an example output image for model using only a single offset image at full resolution. The resulting image shows block artifacts comparable in size to the 8×8 patches that the model was trained on. We found that using the 16×16 variant of the ViT model results in block artifacts of size 16×16, so it appears that these artifacts are intrinsic to the ViT architecture, and a full resolution offset image will not smooth these out without some imposing some constraint.



(a) Example 1 (b) Example 2 (c) Example 3

Bias removed



(d) Example 1 (e) Example 2 (f) Example 3

Bias retained

Figure 2. Results with bias removed (top row) and retained (bottom row)

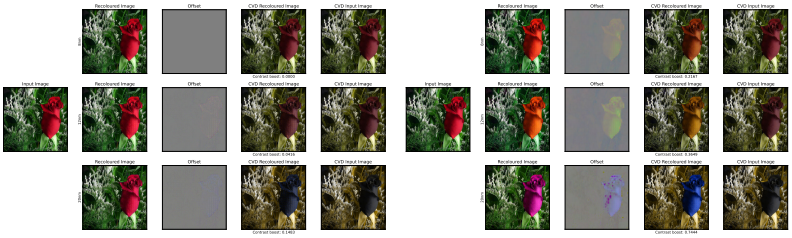
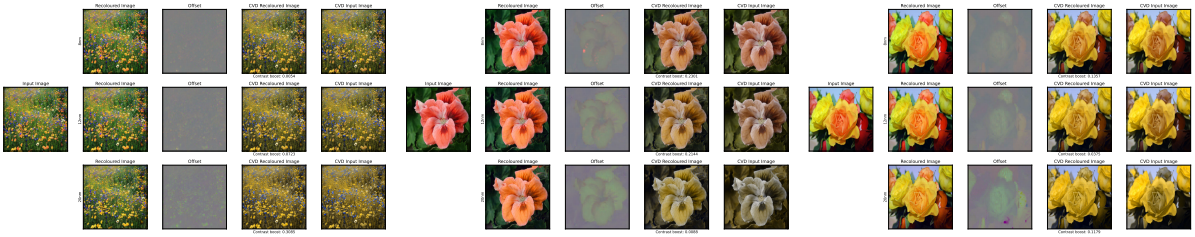


Figure 3. Example image using a single full resolution offset image (left). The result shows blocky artifacts and the metric score is dramatically lower than that of the full model (right)

Figure 4 shows example outputs with a lower resolution of size 64×64 offset image. Each input image is of size 256×256, so a single pixel in the offset image is scaled up to a patch of size 4×4 in the output image. In Example 1 (Fig. 4d) the result has a comparable metric score to that of the full model (Fig. 4a), which indicates that a patch size of 4×4 is sufficient to boost the high frequency detail present in the image. The other two examples (Fig. 4e and Fig. 4f) show diminished contrast improvement, while also containing larger, more uniform regions which are boosted in the full model outputs (Fig. 4b and Fig. 4b). Together, this suggests that the patch size from the upscaled offset images has an impact on the size of the regions which will be contrast-boosted. The implication is that boosting regions of varying sizes requires a combination of varying patch sizes.

3.1.3 Limiting attention layers

Figure 5 shows some example images using only the first attention layer, rather than the four used in the full model. Visually, the images show more muted changes than the baseline method. Since we have not analysed which image features each attention layer corresponds to, we cannot state why this occurs. Existing work inspecting the component filters of Vision Transformers [6] find that deeper attention layers consider larger receptive fields, so it is possible that considering longer-range dependencies in the input image leads to better recolouring outcomes.



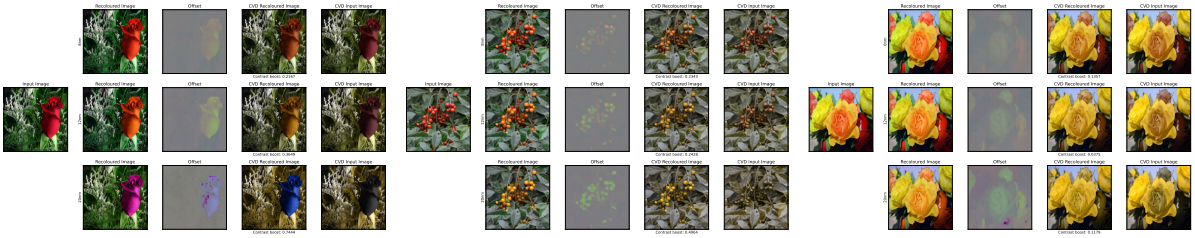
(a) Example 1 (b) Example 2 (c) Example 3

Multiple resolution offsets

(d) Example 1 (e) Example 2 (f) Example 3

Single resolution offset

Figure 4. Examples using an offset image of size 64×64.



(a) Example 1 (b) Example 2 (c) Example 3

Four attention layers

(d) Example 1 (e) Example 2 (f) Example 3

Single attention layer

Figure 5. Examples using only the first attention layer in the loss function. The contrast improves over the raw input, but the changes are more muted than when the full four layers are included.

4. Additional results

This section provides further examples of images adapted for colour vision deficient viewers, with each deficiency type in a separate section. All example images show the input image and its adaption for mild, moderate, and high severity.

4.1. Full results

Dataset	CVD type	Severity	Ours	Swin ViT [1]	Halo-Free [2]
Abstract Art	D	0.4	0.2499	**0.5251**	0.1538
		0.6	0.3185	**0.7677**	0.2455
		1.0	0.4432	**0.9585**	0.3298
	P	0.4	0.2383	**0.4411**	0.2053
		0.6	0.3012	**0.6627**	0.4327
		1.0	0.4334	0.8554	**0.8783**
	T	0.4	**0.3359**	-	0.2298
		0.6	0.2873	-	**0.3266**
		1.0	0.2857	-	**1.0074**
Flowers	D	0.4	0.1391	0.3898	**0.3904**
		0.6	0.1337	0.5866	**0.6375**
		1.0	0.2518	0.7497	**0.9333**
	P	0.4	0.0553	**0.3847**	0.2394
		0.6	0.0039	**0.5542**	0.4369
		1.0	0.0192	0.6733	**0.7666**
	T	0.4	**0.3247**	-	0.1020
		0.6	0.2774	-	**0.3899**
		1.0	0.3391	-	**1.7209**
Places365	D	0.4	0.4072	**0.7224**	0.0894
		0.6	0.5556	**1.0488**	0.2170
		1.0	0.7883	**1.3071**	0.3435
	P	0.4	0.4096	**0.5360**	0.1432
		0.6	0.5916	**0.7854**	0.4069
		1.0	0.8855	**0.9992**	0.9238
	T	0.4	**0.3257**	-	0.0996
		0.6	**0.2490**	-	0.1198
		1.0	0.1002	-	**0.5590**
WikiArt Abstract	D	0.4	0.2093	**0.4251**	-0.1321
		0.6	0.2119	**0.5831**	-0.1713
		1.0	0.2799	**0.6930**	-0.1913
	P	0.4	0.1885	**0.3707**	-0.0894
		0.6	0.2232	**0.5300**	-0.0117
		1.0	0.3245	**0.6620**	0.3048
	T	0.4	**0.3380**	-	0.2012
		0.6	0.2882	-	**0.3278**
		1.0	0.2232	-	**0.8649**
WikiArt Landscape	D	0.4	0.3691	**0.4416**	-0.4335
		0.6	0.3578	**0.6041**	-0.5655
		1.0	0.4498	**0.7107**	-0.6280
	P	0.4	**0.4464**	0.4169	-0.4627
		0.6	0.5166	**0.5905**	-0.5116
		1.0	**0.7562**	0.7200	-0.2033
	T	0.4	**0.4980**	-	0.2791
		0.6	0.4724	-	**0.5911**
		1.0	0.4459	-	**1.6902**
WikiArt Still Life	D	0.4	0.3933	**0.4812**	-0.3636
		0.6	0.3983	**0.6504**	-0.4621
		1.0	0.4697	**0.7649**	-0.5356
	P	0.4	**0.4274**	0.3634	-0.3019
		0.6	0.5084	**0.5208**	-0.2625
		1.0	**0.7248**	0.6498	0.0604
	T	0.4	**0.4090**	-	0.2095
		0.6	**0.3492**	-	0.3318
		1.0	0.2034	-	**0.8471**

Table 1. Contrast boost for different datasets for different CVD types at mild, moderate, and high severity

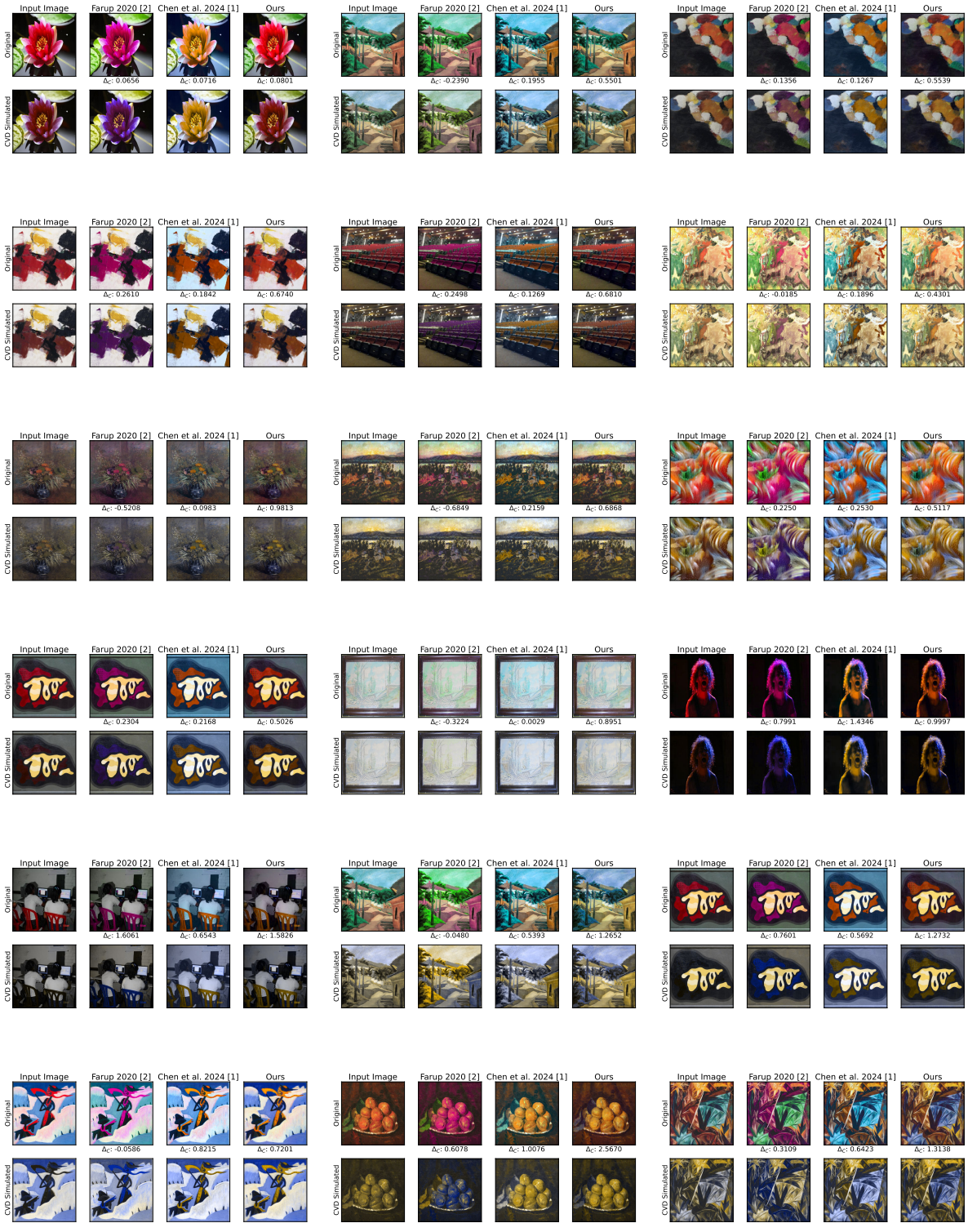
Dataset	CVD type	Severity	Ours	Swin ViT [1]	Halo-Free [2]
Abstract Art	D	0.4	**0.0436**	0.1484	0.0780
		0.6	**0.0564**	0.1484	0.0984
		1.0	**0.0694**	0.1484	0.1249
	P	0.4	**0.0480**	0.1691	0.0764
		0.6	**0.0645**	0.1691	0.0986
		1.0	**0.0818**	0.1691	0.1304
	T	0.4	**0.0222**	-	0.0300
		0.6	**0.0282**	-	0.0424
		1.0	**0.0534**	-	0.0854
Flowers	D	0.4	**0.0491**	0.1785	0.0876
		0.6	**0.0689**	0.1785	0.1104
		1.0	**0.0951**	0.1785	0.1394
	P	0.4	**0.0558**	0.2091	0.0855
		0.6	**0.0766**	0.2091	0.1104
		1.0	**0.1000**	0.2091	0.1454
	T	0.4	**0.0222**	-	0.0327
		0.6	**0.0273**	-	0.0496
		1.0	**0.0577**	-	0.1025
Places365	D	0.4	**0.0534**	0.1469	0.0611
		0.6	**0.0700**	0.1469	0.0770
		1.0	**0.0873**	0.1469	0.0976
	P	0.4	0.0560	0.1482	**0.0529**
		0.6	0.0757	0.1482	**0.0680**
		1.0	0.0956	0.1482	**0.0895**
	T	0.4	**0.0255**	-	0.0272
		0.6	**0.0313**	-	0.0392
		1.0	**0.0592**	-	0.0800
WikiArt Abstract	D	0.4	**0.0391**	0.1290	0.0691
		0.6	**0.0482**	0.1290	0.0875
		1.0	**0.0588**	0.1290	0.1113
	P	0.4	**0.0416**	0.1532	0.0668
		0.6	**0.0547**	0.1532	0.0865
		1.0	**0.0700**	0.1532	0.1149
	T	0.4	**0.0219**	-	0.0291
		0.6	**0.0276**	-	0.0417
		1.0	**0.0487**	-	0.0845
WikiArt Landscape	D	0.4	**0.0287**	0.0890	0.0429
		0.6	**0.0343**	0.0890	0.0535
		1.0	**0.0409**	0.0890	0.0675
	P	0.4	**0.0323**	0.1122	0.0449
		0.6	**0.0405**	0.1122	0.0578
		1.0	**0.0517**	0.1122	0.0768
	T	0.4	**0.0208**	-	0.0259
		0.6	**0.0247**	-	0.0363
		1.0	**0.0396**	-	0.0744
WikiArt Still Life	D	0.4	**0.0318**	0.1099	0.0463
		0.6	**0.0383**	0.1099	0.0584
		1.0	**0.0472**	0.1099	0.0743
	P	0.4	**0.0354**	0.1299	0.0457
		0.6	**0.0451**	0.1299	0.0592
		1.0	**0.0586**	0.1299	0.0788
	T	0.4	**0.0202**	-	0.0251
		0.6	**0.0240**	-	0.0357
		1.0	**0.0376**	-	0.0745

Table 2. Offset energy for different datasets for different CVD types at mild, moderate, and high severity

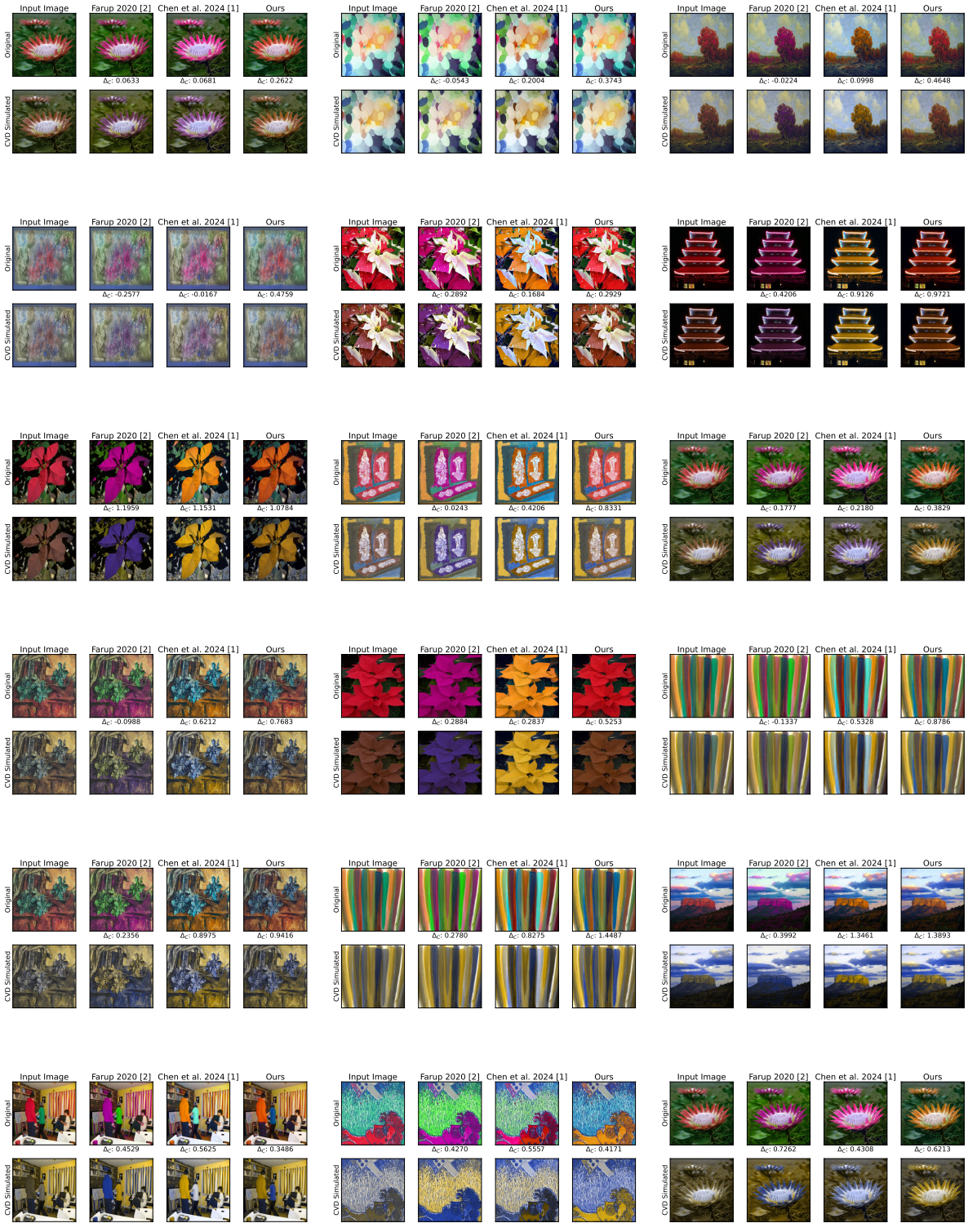
Dataset	CVD type	Severity	Ours	Swin ViT [1]	Halo-Free [2]
Abstract Art	D	0.4	**0.2499**	0.1668	0.0504
		0.6	**0.3185**	0.3092	0.1080
		1.0	0.4432	**0.4661**	0.1435
	P	0.4	**0.2383**	0.1337	0.0883
		0.6	**0.3012**	0.2686	0.2561
		1.0	0.4334	0.4473	**0.5748**
	T	0.4	**0.3359**	-	0.2075
		0.6	**0.2873**	-	0.2567
		1.0	0.2857	-	**0.7281**
Flowers	D	0.4	0.1391	0.1148	**0.2083**
		0.6	0.1337	0.2369	**0.4066**
		1.0	0.2518	0.4107	**0.6654**
	P	0.4	0.0553	0.1164	**0.1512**
		0.6	0.0039	0.2256	**0.3250**
		1.0	0.0192	0.3616	**0.6276**
	T	0.4	**0.3247**	-	0.0821
		0.6	**0.2774**	-	0.2394
		1.0	0.3391	-	**1.0349**
Places365	D	0.4	**0.4072**	0.2918	0.0377
		0.6	**0.5556**	0.5348	0.1769
		1.0	0.7883	**0.8210**	0.3039
	P	0.4	**0.4096**	0.2346	0.0992
		0.6	**0.5916**	0.4508	0.4470
		1.0	0.8855	0.7242	**1.0612**
	T	0.4	**0.3257**	-	0.1156
		0.6	**0.2490**	-	0.1295
		1.0	0.1002	-	**0.5066**
WikiArt Abstract	D	0.4	**0.2093**	0.1383	-0.1324
		0.6	0.2119	**0.2291**	-0.1532
		1.0	0.2799	**0.3290**	-0.1603
	P	0.4	**0.1885**	0.1120	-0.1246
		0.6	**0.2232**	0.2068	-0.0662
		1.0	0.3245	**0.3308**	0.1626
	T	0.4	**0.3380**	-	0.1907
		0.6	**0.2882**	-	0.2616
		1.0	0.2232	-	**0.5962**
WikiArt Landscape	D	0.4	**0.3691**	0.1507	-0.3311
		0.6	**0.3578**	0.2438	-0.4059
		1.0	**0.4498**	0.3410	-0.4131
	P	0.4	**0.4464**	0.1292	-0.3860
		0.6	**0.5166**	0.2272	-0.4064
		1.0	**0.7562**	0.3519	-0.1598
	T	0.4	**0.4980**	-	0.2633
		0.6	**0.4724**	-	0.4544
		1.0	0.4459	-	**0.9804**
WikiArt Still Life	D	0.4	**0.3933**	0.1514	-0.3027
		0.6	**0.3983**	0.2443	-0.3588
		1.0	**0.4697**	0.3522	-0.3944
	P	0.4	**0.4274**	0.1043	-0.2943
		0.6	**0.5084**	0.1892	-0.2593
		1.0	**0.7248**	0.3066	0.0070
	T	0.4	**0.4090**	-	0.1976
		0.6	**0.3492**	-	0.2465
		1.0	0.2034	-	**0.4679**

Table 3. Scaled contrast boost for different datasets for different CVD types at mild, moderate, and high severity

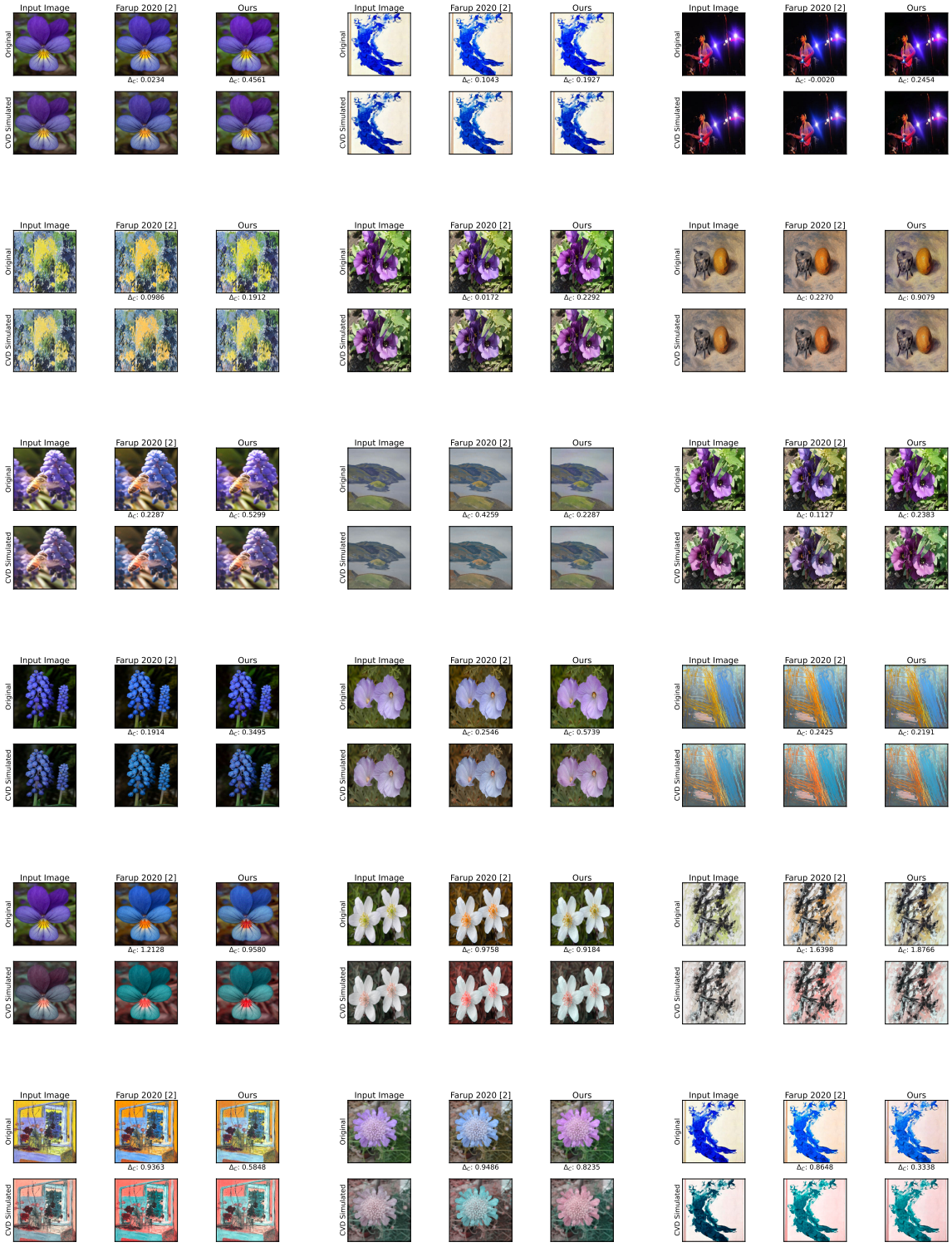
4.2. Protanopic examples



4.3. Deuteranopic examples



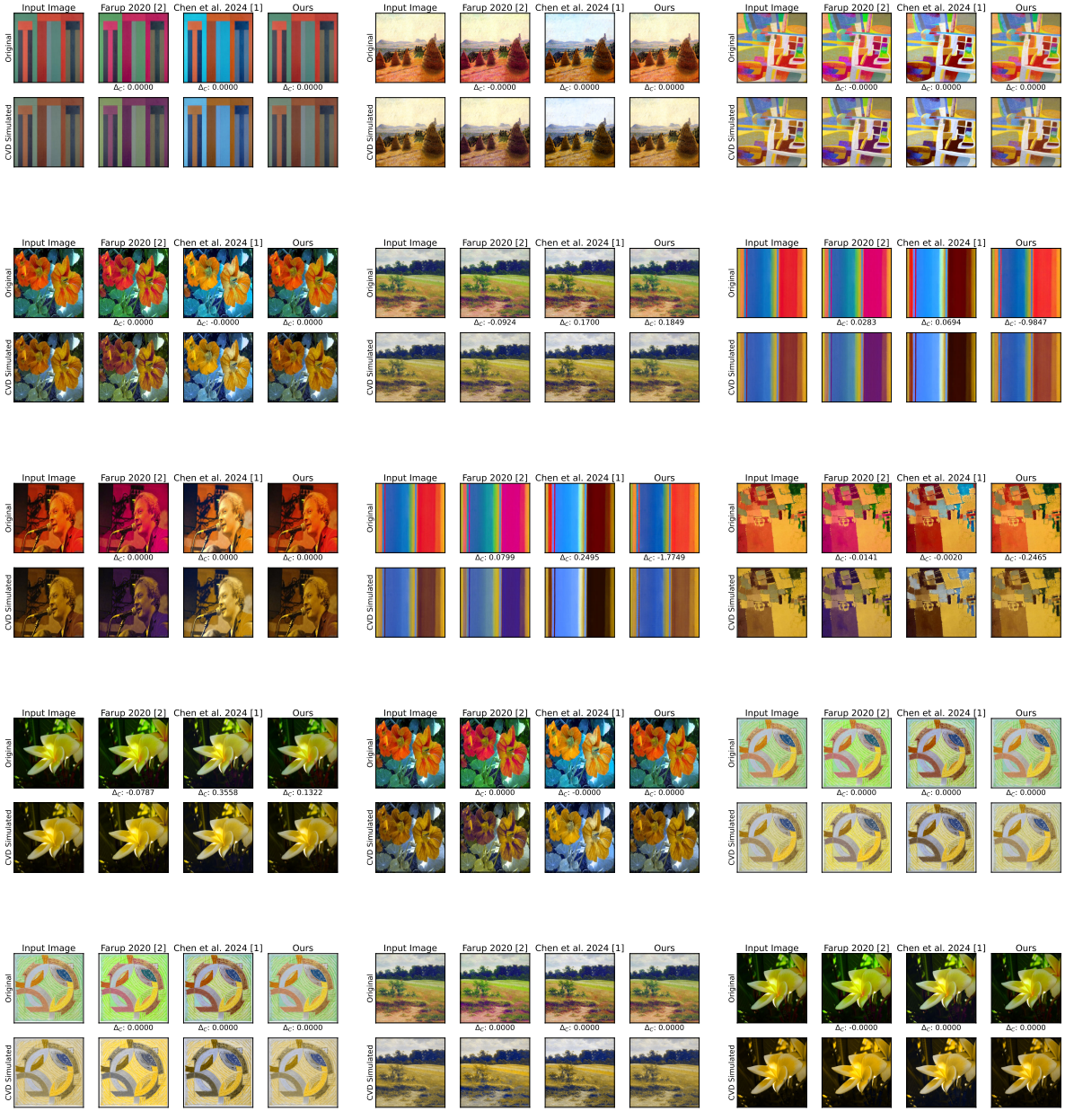
4.4. Tritanopic examples

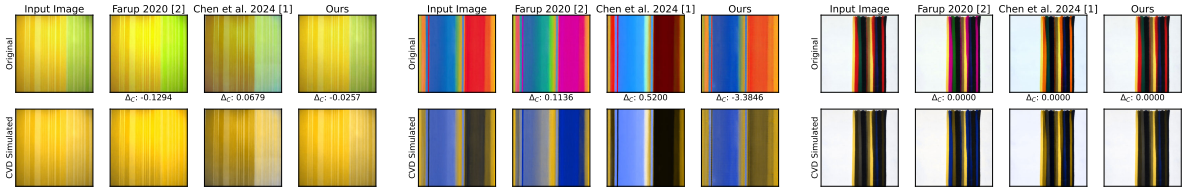


5. Unmodified image examples

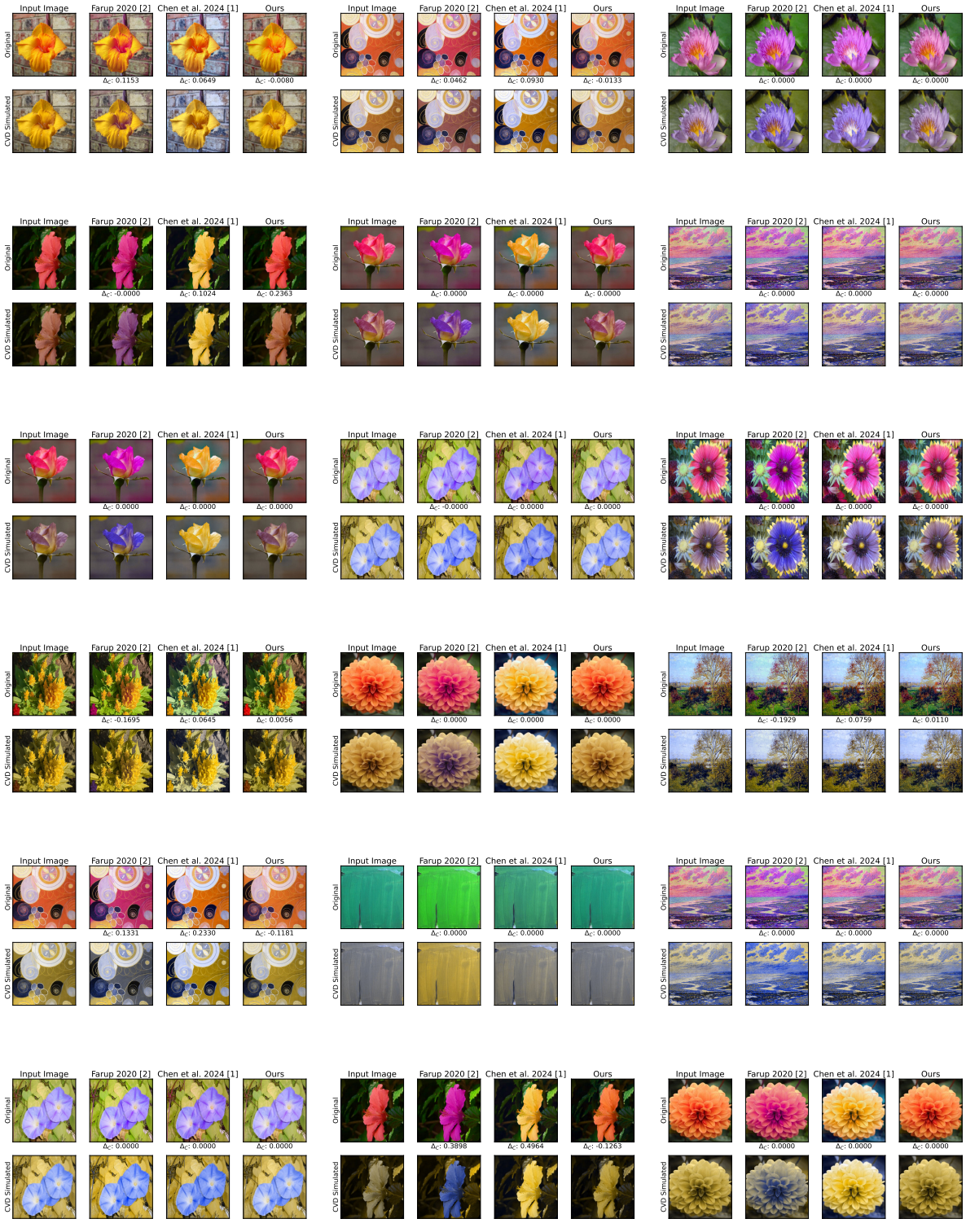
This section contains further examples of images which have not been modified, or modified to a negligible degree by the recolouring algorithm. As with the modified examples, images are divided into deficiency type and each image is shown at different severity levels.

5.1. Protanopic examples

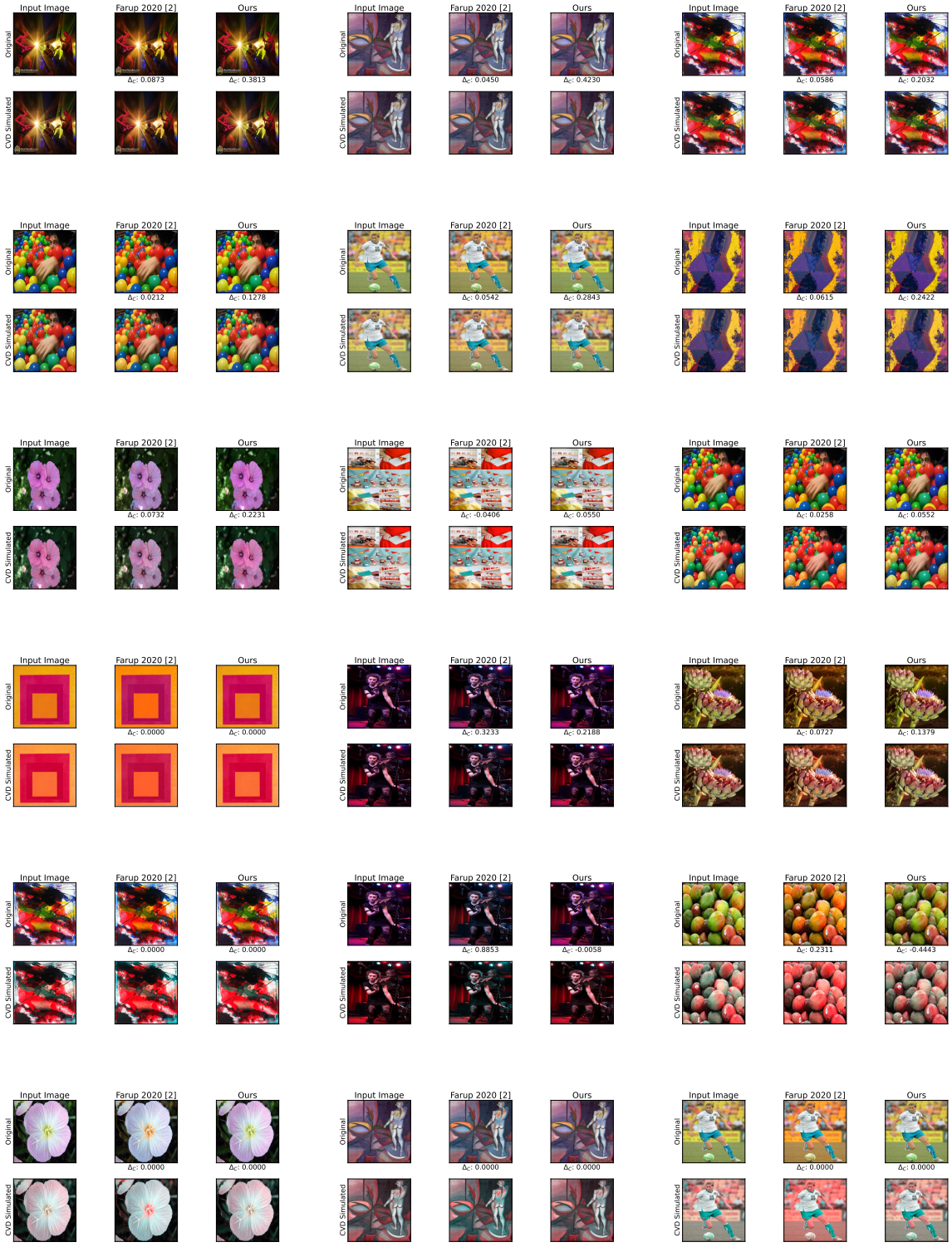



Input Image
Farup 2020 [2]
Chen et al. 2024 [1]
Ours
Original
CVD Simulated
Δ_C: -0.1294
Δ_C: 0.0679
Δ_C: -0.0257
Δ_C: 0.1136
Δ_C: 0.5200
Δ_C: -3.3846
Δ_C: 0.0000
Δ_C: 0.0000
Δ_C: 0.0000

5.2. Deuteranopic examples



5.3. Tritanopic examples



References

[1] 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. 2, 6, 7, 8

[2] Ivar Farup. Individualised halo-free gradient-domain colour image daltonisation. *Journal of Imaging*, 6(11):116, 2020. 6, 7, 8

[3] 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

[4] Noburu Ohta and Alan R. Robertson. *Evolution of CIE Standard Colorimetric System*, chapter 6, pages 175–228. John Wiley Sons, Ltd, 2005. 2

[5] Noburu Ohta and Alan R. Robertson. *Uniform Color Spaces*, chapter 4, pages 115–151. John Wiley Sons, Ltd, 2005. 2

[6] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 4