Supplementary materials: Super-attention for exemplar-based image colorization

Hernan Carrillo[1 [0000−0001−6820−004X]], Michaël Clément[1 [0000−0002−0899−3428]], and Aurélie Bugeau[1 [0000−0002−4858−4944]]

Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France
{hernan.carrillo-lindado, michael.clement, aurelie.bugeau}@labri.fr

1 Architecture details

Figure 1 details the complete architecture. Our implementation is based on a U-net like encoder-decoder architecture [1], the encoder part is a pre-trained VGG19 without the final dense layers, and the decoder is the mirror architecture of the encoder but without pre-train. Each level has two to three convolutional layers with ReLU as its activation function and a batch normalization layer. The architecture uses max-pooling as a downsampling operator and transposed convolution for upsampling. For each skip connection, we add the super-attention block between the encoder and the decoder for the first three levels. For detailed information about size of output resolution of our framework, see Table 1 and Table 2.

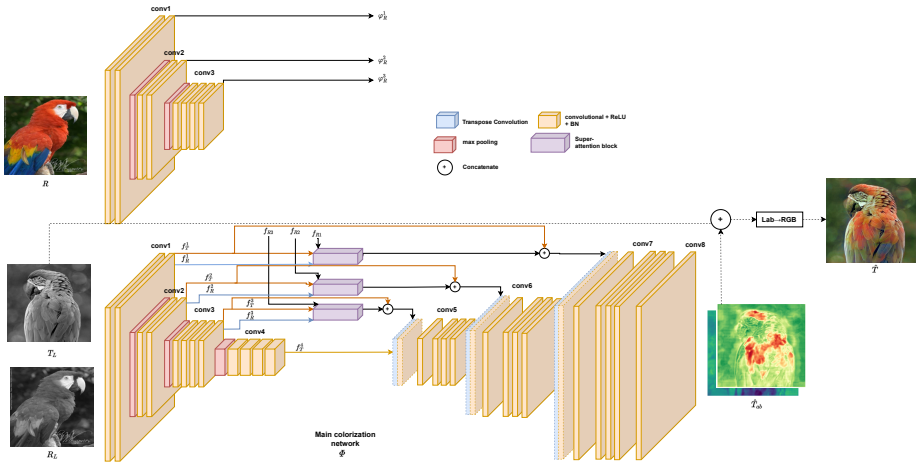

Fig. 1. Detailed architecture of our colorization pipeline.

Table 1. Detailed architecture and output resolution for each block.

Layer type	Output resolution
Input	3 x H x W
Conv1 + Max-pooling	64 x H/2 x W/2
Conv2 + Max-pooling	128 x H/4 x W/4
Conv3 + Max-pooling	256 x H/8 x W/8
Conv4 + Conv. Transpose (I)	512 x H/4 x W/4
Conv5 + Conv. Transpose (II)	256 x H/2 x W/2
Conv6 + Conv. Transpose (III)	64 x H x W
Conv8	C x H x W

Table 2. Detailed architecture and output resolution for super-attention blocks.

Layer type	Output resolution
Super-attention 1	64 x H x W
Super-attention 2	128 x H/2 x W/2
Super-attention 3	256 x H/4 x W/4

2 Comparisons with non-learning based methods

This section presents additional experimental results. Figure 2 shows the comparison between our method and two state-of-the-art non-learning based methods: Welsh *et al.* [2] and Pierre *et al.* [3]. In general, [2] and [3] present more unrealistic colorization results and, in certain cases, evident color bleeding over the images. The reason is that both methods rely on patch matching, and semantic characteristics of images are not taken into account. For instance, no good patch correspondences are found in the reference image for the giraffe from the first image or the coat from the fourth image. Conversely, our method uses semantic features, which let us retrieve content not presented in the reference image, such as the color of the egg from the last image.

3 Results on archive images

Colorizing archive images is still a challenging task for all methods because of the difference in quality between legacy black and white images and modern images. In Figure 3 we show a comparison between state-of-the-art-methods, the previously presented two methods [2] and [3], and three deep-learning exemplar-based methods [4], [5] and [6] on archive black and white images.

4 Adding histogram loss

A way to help reference-based colorization is to favor the transfer of color histogram from the reference to the target image. While we do not want a complete

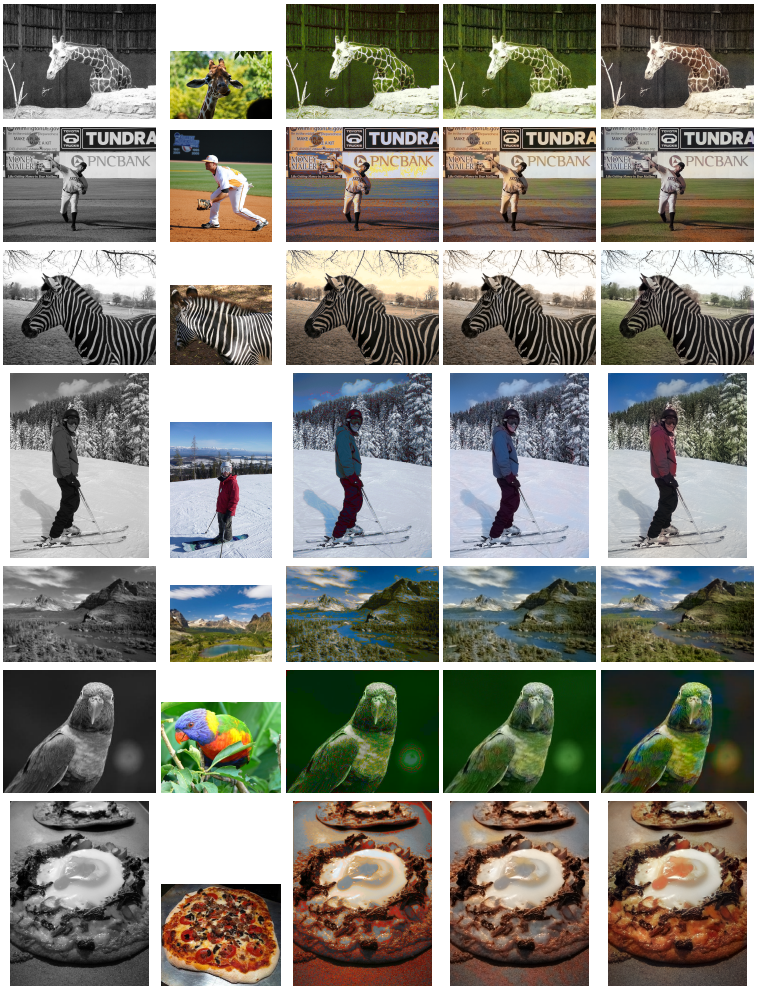

Fig. 2. Comparison of our proposed method with two non-learning reference-based Welsh *et al.* [2] and Pierre *et al.* [3].

histogram transfer, as reference and target images are usually not similar, encouraging histogram transfer may ensure the presence of most colors from the reference into the final color image.

In the deep learning literature, [7] proposed to rely on an additional histogram loss L_{hist} to train their model. This loss forces the predicted color histogram to be similar to the one of the reference image. In [7], the authors proposed a differentiable way to compute the color histogram from the output chrominance channels. This method relies on a set of discretized chrominance bins and bilinear

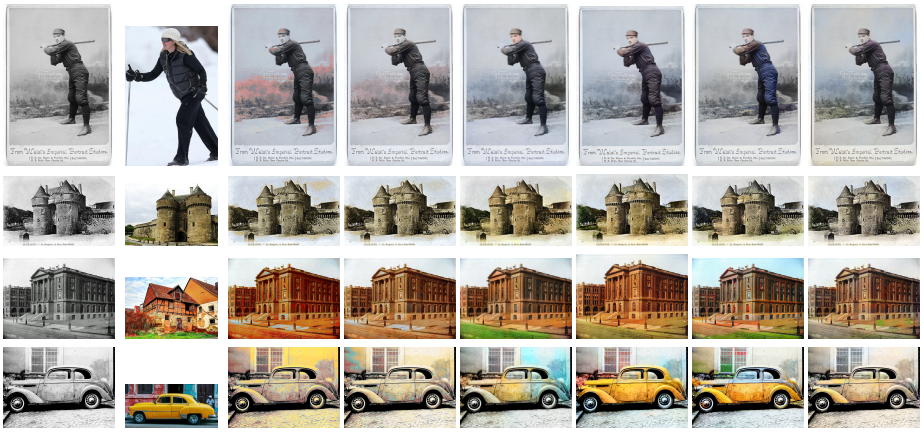

Fig. 3. Comparison of our proposed method on archive images with: two non-learning reference-based Welsh *et al.* [2] and Pierre *et al.* [3]; and three deep-learning exemplar-based methods Deep Exemplar [4], Just Attention [5] and XCNET [6].

Table 3. Quantitative analysis of our model. The metrics SSIM and LPIPS are calculated w.r.t the target groundtruth image, and the HIS metric is calculated w.r.t the ab channel's histogram from the reference image.

	Comparison with groundtruth		Histogram transfer	
Model	**SSIM ↑**	**LPIPS ↓**	**HIS**	**ΔHIS ↓**
Ours without reference	0.920	0.164	-	-
Ours with histogram loss	0.901	0.187	0.786	0.244
Ours	**0.925**	**0.160**	0.596	**0.054**

interpolation. We experimented the same approach to compute color histograms. The histogram loss is then defined as the symmetric $\mathcal{X}^2$ distance [8], between the reference color histogram and the one from our prediction:

$$L_{hist} = 2\sum_{q=1}^{Q} \frac{\left(\mathcal{H}_{\hat{T}_{ab}}(q) - \mathcal{H}_{R_{ab}}(q)\right)^2}{\mathcal{H}_{\hat{T}_{ab}}(q) + \mathcal{H}_{R_{ab}}(q) + \epsilon} \tag{1}$$

where ϵ prevents division by zero and q represents the histogram bins. In our experiments, we set $\epsilon = 1e^{-5}$ and $Q = 441$.

For quantitative evaluation we consider the addition of histogram loss in our model. The results of this evaluation are shown in Table 3. As expected, the addition of histogram loss to our framework leads to much higher raw HIS score, suggesting that more colors from the reference are transferred. However, this comes at the cost of a loss of performance in the two other reconstruction metrics, namely SSIM and LPIPS, in comparison with other variants. Besides,

Fig. 4. Comparison of results obtained using different variants of our colorization framework.

we recall that the goal of this exemplar-based method is not necessarily force the transfer of colors of the reference, but to allow the main model to use these colors as hints to facilitate the final colorization. In this sense, we finally compare the results with ΔHIS, which is the difference between HIS and the reference HIS score of 0.542 (obtained by computing HIS between groundtruth targets and references). Our interpretation is that histogram loss induces a stronger, global transfer of colors between reference and target images, which is not desirable for our application, while without this term it encourages a more specific color transfer.

This interpretation are confirmed by the qualitative results presented in Figure 4. From these results, we can observe that the histogram loss variant leads to more vivid but unrealistic colors, as well as additional color bleeding.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. (2015)
2. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. ACM Transactions on Graphics (2002)
3. Pierre, F., Aujol, J.F., Bugeau, A., Papadakis, N., Ta, V.T.: Luminance-chrominance model for image colorization. SIAM Journal on Imaging Sciences (2015)
4. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics (2018)

5. Yin, W., Lu, P., Zhao, Z., Peng, X.: Yes, "attention is all you need", for exemplar based colorization. In: ACM International Conference on Multimedia. (2021)
6. Blanch, M.G., Khalifeh, I., Smeaton, A., Connor, N.E., Mrak, M.: Attention-based stylisation for exemplar image colourisation. In: IEEE International Workshop on Multimedia Signal Processing. (2021)
7. Lu, P., Yu, J., Peng, X., Zhao, Z., Wang, X.: Gray2colornet: Transfer more colors from reference image. In: ACM International Conference on Multimedia. (2020) 3210–3218
8. Puzicha, J., Hofmann, T., Buhmann, J.: Non-parametric similarity measures for unsupervised texturesegmentation and image retrieval. (1997)