Non-local matching of superpixel-based deep features for color transfer and colorization

Roxane Leduc, Hernan Carrillo, Nicolas Papadakis

▶ **To cite this version:**

Roxane Leduc, Hernan Carrillo, Nicolas Papadakis. Non-local matching of superpixel-based deep features for color transfer and colorization. 2024. hal-04394848

HAL Id: hal-04394848

https://cnrs.hal.science/hal-04394848

Preprint submitted on 15 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Non-local matching of superpixel-based deep features for color transfer and colorization

Roxane Leduc[1], Hernan Carrillo[2], and Nicolas Papadakis[3]

[1]INSA Rouen, France
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[1]Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France

January 15, 2024

Abstract

In this article, we give a thorough description of the algorithm proposed in [H. Carrillo, M. Clément and A. Bugeau, Non-local matching of superpixel-based deep features for color transfer, 2021] for color transfer by relying on a robust non-local correspondence between low-level features at high resolution. An adaptation of this method to colorization process is also described. We highlight the overall relevant results obtained with this technique for both applications and also show its limitations.

1 Introduction

This article deals with both color transfer, the process of changing the color distribution of a target image based on a reference image, and colorization, the process of digitally applying color to grayscale images. One practical application of these techniques is, for example, to provide filmmakers with a quasi-automatic tool that makes their tasks less time-consuming and tedious than today's professional software. While both color transfer and colorization can sometimes suffer from poor spatial and color consistency, the method described in the studied paper [3] addresses these issues by relying on robust non-local matching between low-level features at high resolution. The non-local concept has already been used extensively in computer vision, notably by [2] to improve the performance of digital image denoising methods.

The general idea behind the method proposed by [3] is the following. A superpixel segmentation of the target and reference images is first realized using the SLIC algorithm [1]. Next deep feature maps, which are abstract and semantic representations of an image obtained using a pre-trained deep neural network, are extracted at a superpixel level. Subsequently, a non-local correspondence between superpixels of both images is established using an attention mechanism on the deep features. Global relationships between superpixel are taken into account thanks to this non-local correspondence step, that does not include any additional training. Once the non-local correspondence is established, the pixel-level colors are transferred using a weighted average that takes into account the previously computed attention map between superpixels. All in all, the unsupervised method of [3] is able to transfer the color of the reference image to the target image, while respecting the structure of the target and maintaining a limited computation time for efficient image or video processing.

In this paper, we propose an online implementation of this color transfer method and an extension to the colorization problem. While converting a color image to a grayscale image is

a standard task, the reverse operation is a complex problem since no information about the colors to be added is known *a priori*. This task is classically performed by users, based on their expertise or artistic experience to add hues to monochromatic images. We show that it is possible to automate this colorization process thanks to recent advances in machine learning and neural networks.

The organization of the paper is as follows. Section 2 gives a detailed explanation of the non-local matching technique proposed in [3]. We then study the application of the method to color transfer in section 3. The proposed extension to colorization is finally explored in section 4.

2 Non-local matching method

In this section, we present the method proposed by [3] for non-local matching of super-pixel-based deep features between two RGB images I_T and I_R. In what follows, we will keep the notations of the reference article: I_T will be the target image and I_R the reference one. We first describe in section 2.1 the extraction of super-pixels features, called super-features. The matching process between super-features is then detailed in section 2.2.

2.1 Super-Features Encoding (SFE)

The encoding of the super-features F of an image I takes place in three stages: 1) super-pixel decomposition, 2) extraction of deep features using convolutional neural networks, 3) channel averaging process to obtain super-features.

2.1.1 Super-pixel segmentation

Firstly, two super-pixel maps are generated using a super-pixel decomposition algorithm on the target and reference images. A super-pixel is a group of connected pixels that share common characteristics such as similarity of color or intensity and spatial proximity. Super-pixels are commonly used to speed up the execution of image processing algorithms and, in some cases, to improve results.

We use the SLIC algorithm [1] which is an adaptation of the k-means clustering algorithm for image segmentation purpose. SLIC segments an image into superpixels using both color and spatial position information. It associates neighboring pixels based on their similarity and updates the superpixel centers until convergence is reached. Rather than simply decimating the image to reduce the amount of information, segmenting into superpixels provides a set of regions of interest to process, without reducing the amount of raw information in the image. Figure 1 illustrates this process. This steps produces two super-pixels maps S_T and S_R that respectively contain N_T and N_R super-pixels. We also denote as P_i the number of pixels contained in the i-th super-pixel $S_T(i)$.

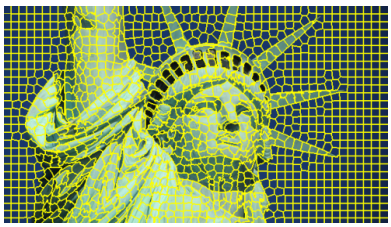


Figure 1: Visualisation of a superpixel segmentation S obtained with the SLIC algorithm [1].

2.1.2 Deep features

Deep learning features are obtained from a pre-trained convolutional network applied on the reference and target images I_R and I_T. We use here the first three layers of a modified VGG-19 architecture [5] as a feature extractor. These first three layers provide a long range of low-level features that suit diverse types of images.

In each case, after the convolution step and the application of a ReLU function, activations are then batch-normalized before moving on to the next layer. This stabilizes activation values, reduces covariation effects between different activations and improves learning convergence. We remove max-pooling layers from the baseline VGG-19 architecture.

This step thus produces $l = 3$ (one for each layer) feature maps f_{Tl} and f_{Rl}, composed of C = 64, 128 then 256 channels. The spatial dimension of the feature maps f_{Tl} (resp. f_{Rl}) correspond to the one of the target image I_T (resp. reference image I_R).

Remark. *This approach takes into account features derived from VGG-19, a pre-trained deep convolutional networks particularly effective for processing high-dimensional data such as images or videos. The method can nevertheless handle any other handcrafted or learned features.*

2.1.3 Average pooling

Finally, we apply average pooling to the deep features using super-pixel maps S_T and S_R. The features f_T and f_R of the pixels inside each superpixel are averaged per channel and then stacked in the form of matrices known as *super-features*. We therefore obtain three super-feature maps $F_{T_l} \in \mathbb{R}^{N_T \times C}$, for the target image, with l index ranging from 1 to 3 corresponding to the information from the first three layers of VGG-19 convolutional neural network and N_T the number of super-pixel of the target image. In the same way, for the reference image, we obtain three super-feature maps $F_{R_l} \in \mathbb{R}^{N_R \times C}$, $l = 1 \cdots 3$. The overall super-features encoding is illustrated in Figure 2.

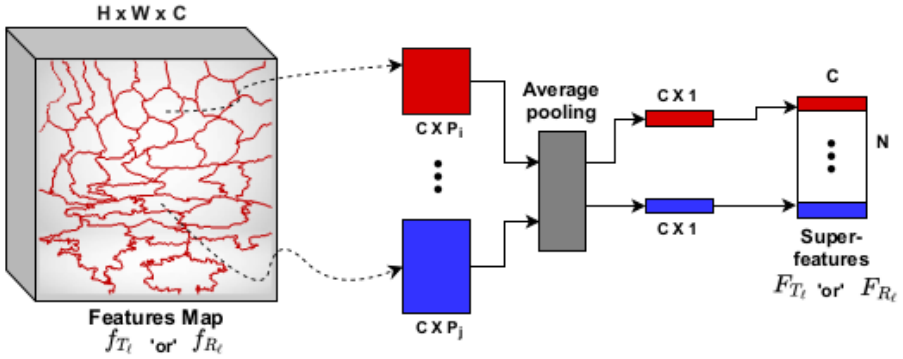


Figure 2: Super-features encoding of [3]. Let H and W be the image size in number of pixels, C the number of channels of the different VGG-19 layers used, N_R (resp. N_T) the total number of reference (resp. target) super-pixels, and P_i the number of pixels in the i-th superpixel. The process takes as input a feature map of size $H \times W \times C$, in which each super-pixel is extracted and encoded in vectors of size $C \times P_i$. Afterward, the vectors are pooled channel-wise and, finally, stacked in the respective super-features matrices F_{T_l} (resp. F_{R_l}) of size $C \times N_R$ (resp. $C \times N_T$).

2.2 Super-Features Matching (SFM)

As stated above, super-features provide a compact encoding for calculating the correlation between deep features. In order to achieve a robust correspondence between the super-features of the target named F_{T_l} and the super-features of the reference named F_{R_l}, the authors were inspired by the attention mechanism described in [7].

Attention mechanisms [6] were popularized with the rise of transformers, a type of neural network architecture that has revolutionized the field of natural language processing and other sequence processing tasks. The attention mechanism allows the model to focus on specific parts of the input when generating output. Rather than processing the whole sequence at once, transformers use attention to give variable importance to each element of the sequence depending on its context.

The idea of the Super-Features Matching (SFM) process is to exploit the non-local similarities between the super-features of the images by calculating the attention map at layer l as follows:

$$A_l(i,.) = \text{softmax}\left(\frac{M_{T_l R_l}(i,.)}{\tau}\right), \qquad i = 1 \cdots N_T, \tag{1}$$

where, for each superpixel i of the target and j of the reference, we define:

$$M_{T_l R_l}(i,j) = \frac{(F_{T_l}(i) - \mu_{T_l})^T (F_{R_l}(j) - \mu_{R_l})}{\|F_{T_l}(i) - \mu_{T_l}\|_2 \, \|F_{R_l}(j) - \mu_{R_l}\|_2}. \tag{2}$$

In this equation, $M_{T_l R_l}$ corresponds to the correlation matrix between the super-features of the target $F_{T_l}(i)$ and the reference $F_{R_l}(j)$, computed using the mean values $\mu \in \mathbb{R}^C$ over the N_T (resp N_R) super-feature values. The choice of the normalization through correlation is motivated by the use of a global temperature parameter $\tau > 0$ to process all components $M_{T_l R_l}$ in the same way. The softmax operation in expression (2) is realized with respect to the second dimension of the matrix, so that $\sum_j A_l(i,j) = 1$ for all super-pixels $i = 1 \cdots N_T$ of the target image; while $A_l(i,j) \geq 0$ for all $i = 1 \cdots N_T$ and $j = 1 \cdots N_R$.

The final attention map A is the weighted sum of the attention maps for each layer, divided by the sum of the weights:

$$A(i,j) = \frac{\sum_{l=1}^{l=3} w_l A_l(i,j)}{\sum_{l=1}^{l=3} w_l}, \tag{3}$$

where all weights w_l are set to 1 in our experiments. The value of the attention map $A(i,j)$ can be understood as a measure of the influence of the super-pixel $S_R(j)$ of the reference image I_R for the processing of the super-pixel $S_T(i)$ of the target image I_T.

3 Color transfer

In this section, we focus on transferring the color of a reference image I_R to a target image I_T. To that end, we follow the color fusion framework initially proposed in [4] which uses the attention maps to obtain the new color for each pixel of the target image.

3.1 Method

Color transfer aims at changing the colors of pixels in the target image I_T using the color palette of the reference image I_R. To that end, the method [3] extends the color fusion framework initially proposed in [4]. The process leverages on the attention map A provided by SFE-encoding and SFM-matching, which encodes semantic correspondences between both images. For all pixels inside a super-pixel $S_T(i)$, $i = 1 \cdots N_T$, the attention map $A(i,j)$ is used as a weight to balance the importance of the color to transfer from the reference super-pixel $S_R(j)$.

Denoting as $\overline{I}_R(j)$ the mean color value of pixels belonging to the super-pixel j of the reference image I_R, the output of the color transfer process on the target image I_T at pixel p

is obtained as

$$\widehat{I}_T(p) = \frac{\sum_{j=1}^{N_R} W(p,j)\overline{I}_R(j)}{\sum_{j=1}^{N_R} W(p,j)}, \tag{4}$$

where $W(p,j) = \sum_{i=1}^{N_T} d(p,i)A(i,j)$ and $d(p,i)$ depends on the distance between the pixel p and the center $\bar{p}_i = (\sum_{p \in S_T(i)} p)/P_i$ of the i-th super-pixel of the target image. This weight is calculated using a Mahalanobis-type formula:

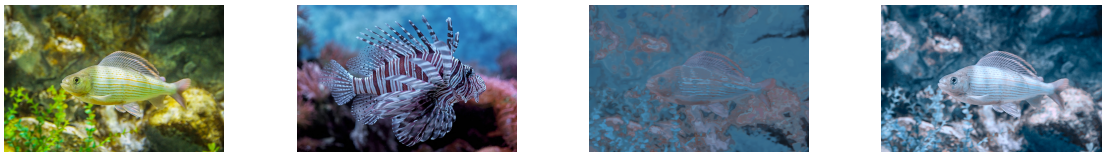
$$d(p,i) = e^{-\frac{(V_T(p)-\overline{V}_T(i))^T \Sigma_i^{-1} (V_T(p)-\overline{V}_T(i))}{\sigma_g}} \tag{5}$$

with $V_T(p) = [p, I_T(p)]$ being the vector describing the position and the color of pixel p, and $\overline{V}_T(i) = [\bar{p}_i, \overline{I}_T(i)]$ being the average vector describing the position and color centroids of superpixel $S_R(i)$. The spatial and colorimetric covariances of pixels belonging to the superpixel $S_R(i)$ are computed as:

$$\Sigma_i = \begin{pmatrix} \delta_s^2 [Cov(p)]_{p \in S_R(i)} & 0 \\ 0 & \delta_c^2 [Cov(I_T(p)]_{p \in S_R(i)} \end{pmatrix}, \tag{6}$$

where the parameters δ_s^2 and δ_c^2 respectively weight the influence of spatial and color information.

In order to optimize computational performance, an initial resizing (*down-sizing*) was applied to the I_R and I_T input images. To guarantee optimum visual quality at the end of the algorithm, we introduce a conversion to the CIELAB color space. We start by converting $\hat{I}_T$ from RGB to CIELAB color space. We thus obtain the 3 channels *Lab* representation $(\hat{L}, \hat{a}, \hat{b})$ for the colorized image $\hat{I}_T$. We then isolate its chrominance channels $\hat{a}$ and $\hat{b}$ and use them for the resizing operation (*up-sizing*). Next, we concatenate the luminance channel of the original I_T image (also converted into CIELAB space to retrieve the luminance channel) with the chrominance channels a and b of $\hat{I}_T$. This process replaces the luminance channel of the resulting image $\widehat{I}_T$ with the original gray level of I_T. This is an essential step to maintain the structural information present in I_T. Finally, we convert back the resulting image from CIELAB color space to the RGB color space, giving us the final synthesized image. As illustrated in Figure 3, this post-processing makes the colorized images more realistic.



(a) (b) (c) (d)

Figure 3: (a) Target; (b) Reference; (c) Resizing without any prior CIELAB conversion; (d) Resizing combined with CIELAB conversion.

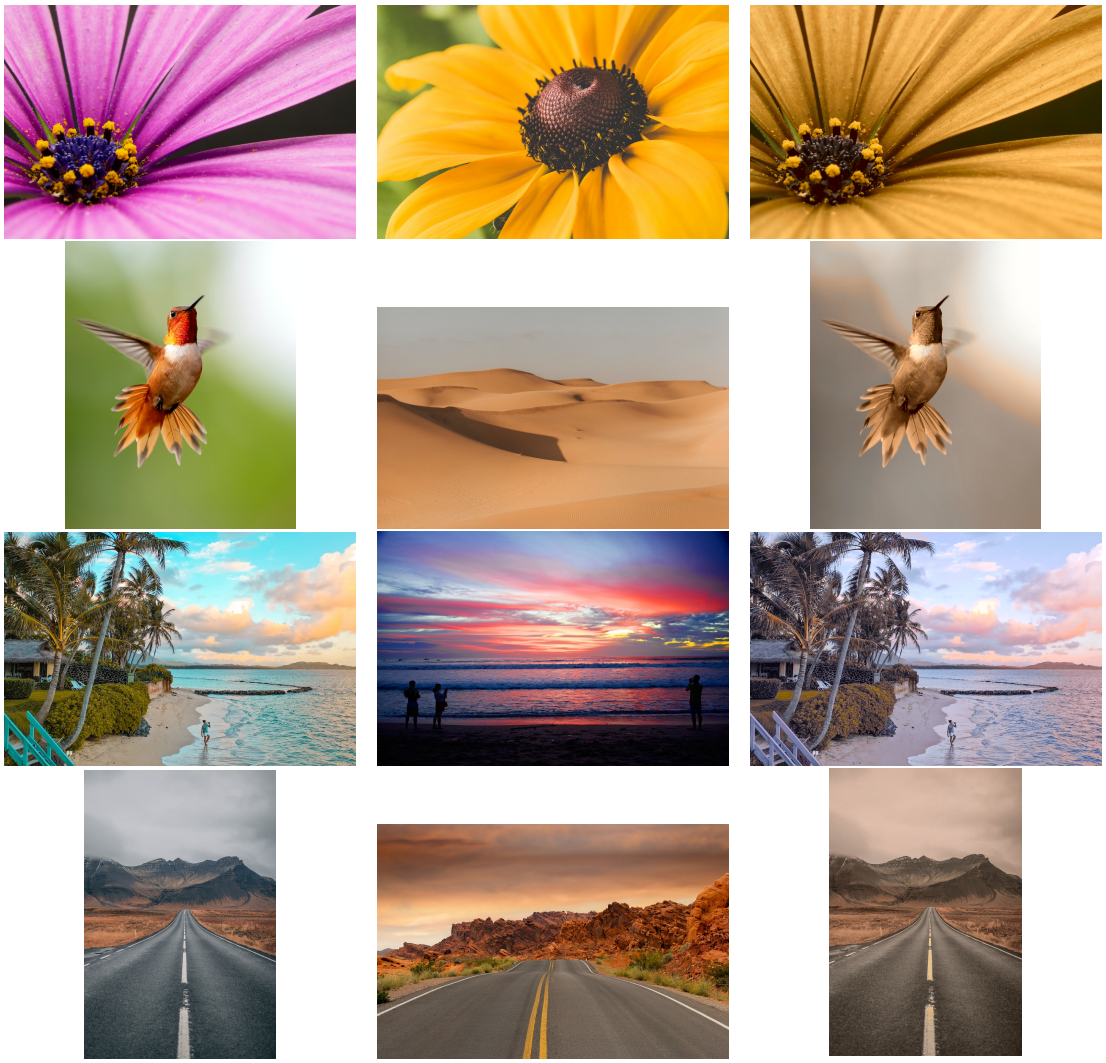
3.2 Experimental study

Using the online demonstration, the color transfer method [3] can be applied on any pairs of target and reference images. In this section we carry out a series of experiments to describe the influence of the different parameters.

We first run the companion demonstration code on four pairs of real images[1] to illustrate the pros and cons of this technique applied to color transfer. For the purposes of this demonstration, we have arbitrarily assumed that the number of superpixels is equal to $3 \times \lfloor \sqrt{H \times W} \rfloor$. We then analyze the influence of the different τ, δ_s and δ_c parameters of the method.

[1] Real means no computer-generated images.

As illustrated in Figure 4, the color transfer method can provide plausible results for images containing a single object of interest (flower and bird examples) or more complex scenes (beach example). In the case of the road against a mountain backdrop – see the 4th row of the figure, the correspondence between super-pixels is questionable (as an example, the color of the broken line in the middle of the road has remained unchanged).

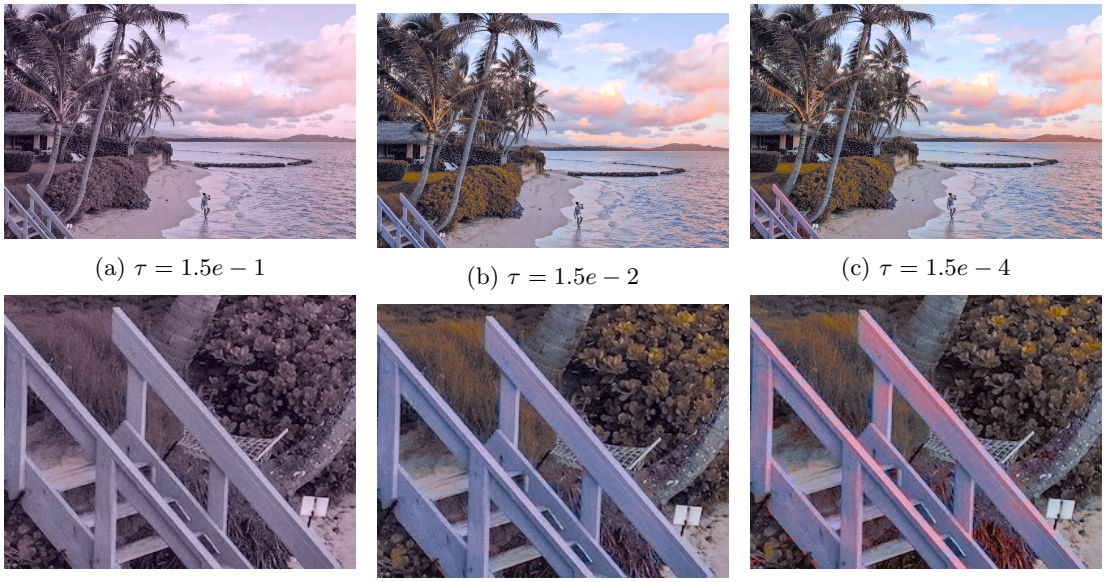


(a) Targets I_T (b) References I_R (c) Results $\hat{I}_T$

Figure 4: Application of color transfer to 4 different image series.

τ-parameter sensitivity analysis. We now evaluate the impact of the main temperature parameter τ of equation (1), that balances the weights in the attention map A. Figure 5 illustrates the color transfers obtained for the beachfront image and different values of τ. As stated in [3], our experiments suggest that a value $\tau = 1.5e-2$ gives satisfactory visual results of a large range of images.

As the value of τ increases (for example, $\tau = 1.5e{-}1$), the softmax operation of expression (1) makes the probability distribution $A(i,.)$ more uniform. This implies an important mixing of the colors of all superpixels $S_R(j)$, which lead to drab colors in the synthesized image (Figure 5a. On the other hand, when τ decreases, there is a one-to-one matching between a target super-feature and a reference super-feature. For $\tau = 1.5e{-}4$, this results in non-uniform color transfer at the bottom left part of the image displayed in Figure 5f.



(a) $\tau = 1.5e-1$ (b) $\tau = 1.5e-2$ (c) $\tau = 1.5e-4$

(d) $\tau = 1.5e-1$ (e) $\tau = 1.5e-2$ (f) $\tau = 1.5e-4$

Figure 5: Influence of the temperature value τ on the color transfer.

δ_s and δ_c-parameters sensitivity analysis. We now evaluate the impact of the δ_c and δ_s-parameters of equation (6), that weight the influence of color and spatial information. Figure 6 illustrates the color transfers obtained for the road image and different values of δ_s and δ_s. As stated in [3], our experiments suggest that values $\delta_s = 10$ and $\delta_c = 0.1$ gives satisfactory visual results of a large range of images. Indeed, with a preponderance of color information δ_c and less importance given to spatial information δ_s, we observe an inadequate distribution of colors, particularly at the foot of the mountain. In addition, the edges of the road are poorly colored, with dull, lackluster hues that do not faithfully reflect the chromatic palette of the original image.

4 Colorization

We now propose an extension of the method [3] to colorization. In this setting, the target image I_T is a grayscale image, whereas the reference one I_R is a color image.

4.1 Method

The colorization technique we propose consists of 4 steps: 1) transformation of the reference color image into a grayscale one I_{R_g}; 2) computing SFE-encoding and SFM-correspondences between super-features of grayscale images I_T and I_{R_g}; 3) synthesis of the colorized target image $\hat{I}_T$, using attention maps and the original colors of the source image I_R; 4) post-processing in CIELAB color space, by mixing the original luminance information in the grayscale image I_T with the chrominance channels of $\hat{I}_T$.

From colors to grayscale. The transformation of the RGB color image $I_R = (R, G, B)$ into a grayscale image is done using the standard weighted average from PAL or NTSC models: $I_{R_g} = 0.299R + 0.587G + 0.114B$.

Super-features encoding and matching. We apply the SFE-encoding to both the grayscale image I_{R_g} and the target one I_T. Next we perform the SFM-matching with a slight modifica-

(a) $\delta_c = 10$ $\delta_s = 0.1$ (b) $\delta_c = 0.1$ $\delta_s = 10$

(c) $\delta_c = 10$ $\delta_s = 0.1$ (d) $\delta_c = 0.1$ $\delta_s = 10$

Figure 6: Influence of the δ_s and δ_c-parameters on the color transfer.

tion of the attention map computation described in section 2.2. The merging of the attention maps corresponding to the 3 VGG-19 layers is here realized before the softmax operation:

$$A(i,.) = \text{softmax}\left(\frac{\sum_{l=1}^{l=3} w_l M_{T_l R_{gl}(i,.)}}{\tau}\right), \qquad i = 1 \cdots N_T. \tag{7}$$

In our experiments, this change appeared useful to avoid the transfer of drab colors.

Color fusion framework. The colorized image $\hat{I}_T$ is obtained by tracing back the original colors of the superpixels in the reference image I_R, as detailed in equation (4) of section 3.

Post-processing step. As explained in section 3, we are going to perform a conversion in CIELAB space in order to improve the visual rendering during resizing, which reduces computation time. We will concatenate the luminance channel of the target grayscale image I_T with the chrominance channel of the resulting image $\hat{I}_T$. Figure 7 illustrates this process that produces results that are more consistent, more perceptually faithful and less sensitive to variations in brightness.

(a) (b) (c) (d)

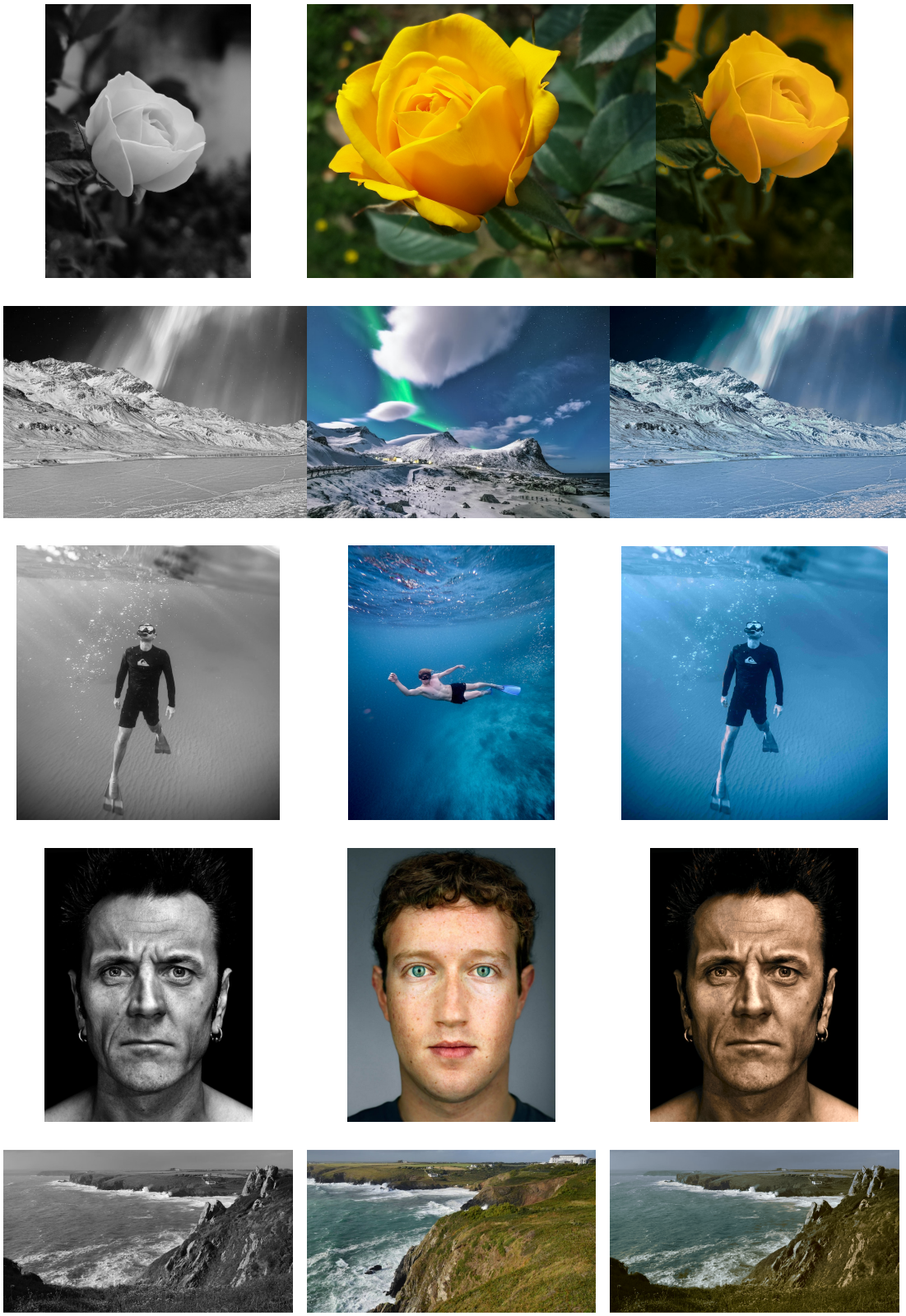
Figure 7: (a) Target; (b) Reference; (c) Resizing without any prior CIELAB conversion; (d) Resizing combined with CIELAB conversion.

4.2 Experimental study

We present in Figure 8 colorization results obtained with the companion demonstration code applied on four real images. When the source and target images are carefully selected, so that they have a particularly high degree of similarity, the visual results are relevant, with realistic colorization (see for instance the flower and the coastline). The colorization process nevertheless tends to reproduce the most predominant hue in the reference image (dipper and human face examples).

τ-parameter sensitivity analysis. We evaluate the influence of the temperature parameter τ in equation (7) on the flower example. As illustrated in Figure 9, when $\tau = 1.5e-2$, the color palette of the reference image I_R is better represented in the colorized image. On the other hand, increasing the value as $\tau = 1.5e-1$ results in a decrease in the diversity of colors present in the final image.

δ_s and δ_c-parameters sensitivity analysis. We now evaluate the impact of the δ_c and δ_s-parameters of equation (6), that weight the influence of color and spatial information. Figure 10 illustrates the colorization process obtained for the image and different values of δ_s and δ_c. Again, our experiments suggest that values $\delta_s = 10$ and $\delta_c = 0.1$ gives satisfactory visual results of a large range of images. Indeed, when the emphasis is placed primarily on color information while neglecting spatial information, the final rendering proves less satisfactory and lacks coherence. Close examination of the image reveals an inconsistent distribution of colors, marked by noticeable and distinct variations in different places. This non-homogeneity creates contrasting areas where hues appear to diverge significantly, introducing visual irregularities within the image.



(a) Targets in grayscale (b) References (c) Results

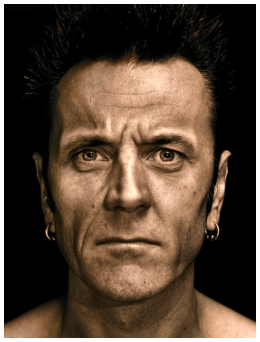
Figure 8: Application of colorization to 4 different image series.

(a) $\tau = 1.5e-1$

(b) $\tau = 1.5e-2$

Figure 9: Influence of the temperature value τ in the colorization results.

(a) $\delta_c = 10 \qquad \delta_s = 0.1$



(b) $\delta_c = 0.1 \qquad \delta_s = 10$

Figure 10: Influence of the δ_s and δ_c-parameters in the colorization results.

5 Conclusion

This paper presents a framework for color transfer and colorization of a target image using the color information contained in a reference image. The process is based on non-local matching of deep features extracted from superpixels through an attention mechanism. Experimental results demonstrate the effectiveness of this technique, that manages to preserve the fine details, textures and structures of the target images, while producing consistent and plausible color synthesis using the information from the reference image.

Acknowledgment

This study has been carried out with financial support from the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01).

Image Credits

 Jeremy bishop (Pexels)[2]

 Stein Egil Liland (Pexels)[2]

 Frans Van Heerden (Pexels)[2]

 Style transfer for headshot portraits (MIT Open Access Articles)[3]

 Style transfer for headshot portraits (MIT Open Access Articles)[3]

 Rennon kiefer (Pexels)[2]

 Nikhil Singh Rajput (Pexels)[2]

 Pixabay (Pexels)[2]

 James Lee (Pexels)[2]

 Ylanite Koppens (Pexels)[2]

 Jess Loiterton (Pexels)[2]

 Pixabay (Pexels)[2]

 Sebastian Palomino (Pexels)[2]

 Pixabay (Pexels)[2]

 Frank Cone (Pexels)[2]

 Vlada Karpovich (Pexels)[2]

 Slimmars (Pexels)[2]

 Slimmars (Pexels)[2]

 Denys Razumovskyi (Pexels)[2]

 Zoosnow (Pexels)[2]

[2] https://www.pexels.com/fr-fr/discover/

[3] https://dspace.mit.edu/bitstream/handle/1721.1/100018/Durand_Style.pdf

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, nov 2012.

[2] A. Buades, B. Coll, and J.-M. Morel. A Non-Local Algorithm for Image Denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65, 2005.

[3] H. Carrillo, M. Clément, and A. Bugeau. Non-local Matching of Superpixel-based Deep Features for Color Transfer. In *International Conference on Computer Vision Theory and Application (VISAPP'22)*, pages 38–47, 2022.

[4] R. Giraud, V.-T. Ta, and N. Papadakis. Superpixel-based color transfer. In *IEEE International Conference on Image Processing (ICIP'17)*, pages 700–704, 2017.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen. Deep exemplar-based video colorization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, pages 8052–8061, 2019.