Non-local Matching of Superpixel-based Deep Features for Color Transfer

Hernan Carrillo[a], Michaël Clément[b] and Aurélie Bugeau[c]

Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France

Keywords: Superpixels, Attention Mechanism, Color Transfer, High-resolution Features, Non-local Matching.

Abstract: In this article, we propose a new method for matching high-resolution feature maps from CNNs using attention mechanisms. To avoid the quadratic scaling problem of all-to-all attention, this method relies on a superpixel-based pooling dimensionality reduction strategy. From this pooling, we efficiently compute non-local similarities between pairs of images. To illustrate the interest of these new methodological blocks, we apply them to the problem of color transfer between a target image and a reference image. While previous methods for this application can suffer from poor spatial and color coherence, our approach tackles these problems by leveraging on a robust non-local matching between high-resolution low-level features. Finally, we highlight the interest in this approach by showing promising results in comparison with state-of-the-art methods.

1 INTRODUCTION

Non-local operators were introduced in image processing in (Buades et al., 2005) with the so-called Non-local means framework, initially used to filter out image noise by computing a weighted mean of all pixels in an image. Non-local means allow remote pixels to contribute to the filtered response, achieving less loss of details. It was then extended to non-local features matching for super-resolution (Glasner et al., 2009), inpainting (Wexler et al., 2004), or color transfer (Giraud et al., 2017) proving to achieve robust global features similarities.

Non-local ideas have recently been introduced within neural networks architectures (Wang et al., 2018) (Yu et al., 2018). The computation of non-local similarities in neural networks is related to so-called attention mechanisms (Bahdanau et al., 2015). This attention block learns to compute similarities between input embeddings or data sequences. Lately, attention mechanisms have been popularised with the rise of transformers (Vaswani et al., 2017), *i.e.*, end-to-end neural network architectures that include only (self-)attention layers. Transformers compute non-local similarities between multi-level feature maps. This type of architecture succeeds as a state-of-the-art method due to the capacity and flexibility of these attention blocks. The recent work (Wang et al., 2018) has bridged the gap between the self-attention mechanism and non-local means. They stated that the self-attention mechanism captures long-range dependencies between deep learning features by considering all features into the calculation.

Recently, the authors of (Zhang et al., 2019) presented similarity calculation between different feature maps (target and reference images) based on attention mechanism for image colorization. The principal drawback of such mechanism is the complexity of the non-local operation, which has to be done on features with low dimensions due to computational burden. This is known as the quadratic scaling problem. However, low-resolution features usually do not carry sufficient information for calculating precise pixel-wise similarities. For instance, deep features mainly carry high-level semantic information related to a specific application (*i.e.*, classification) that can be less relevant for high-resolution similarity calculation or matching purposes.

In this paper, we compute similarities between high-resolution deep features obtained from pre-trained convolutional neural networks, as this retains rich low-level characteristics. Due to the dimensionality issue, we exploit existing superpixels extractor in order to match these high-resolution features. To

[a] https://orcid.org/0000-0001-6820-004X
[b] https://orcid.org/0000-0002-0899-3428
[c] https://orcid.org/0000-0002-4858-4944

Carrillo, H., Clément, M. and Bugeau, A.
Non-local Matching of Superpixel-based Deep Features for Color Transfer.
DOI: 10.5220/0010767900003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 38-47
ISBN: 978-989-758-555-5; ISSN: 2184-4321

illustrate the interest of this super-features matching operation, we apply it to the problem of color transfer. Color transfer aims at changing the color characteristics of a target image by copying the ones from a reference image. Ideally, the result must reach a visually pleasant image, avoiding possible artifacts or improper colors. It covers various applications in areas such as photo enhancement, films post-production, and artistic design. Transferring the right colors requires computing meaningful similarities between the reference and the target images. These similarities must preserve important textures and structures of the target image. Therefore, we will show that high-resolution features are essential.

The contributions are: 1) we propose the super-features encoding block, which extracts deep feature maps using superpixel decomposition; 2) we propose a robust non-local similarity between super-features using an attention mechanism; and 3) we build upon (Giraud et al., 2017) and include these similarities in a non-local color fusion framework, achieving promising results on several target and reference image pairs.

2 RELATED WORK

2.1 Superpixels

Exploiting superpixel representation allows finding interesting region's characteristics in images, such as color and texture consistency (Achanta et al., 2012). Many advantages can be derived using this type of decomposition, for instance, dimensionality reduction by grouping pixels with similar characteristics (Van den Bergh et al., 2015). Additionally, this compact representation helps to overcome high computational costs on computer vision tasks such as object segmentation (Tighe and Lazebnik, 2010) or object localization (Fulkerson et al., 2009). However, the irregular form of the representation makes its usage difficult in computer vision tasks, especially the ones using deep learning approaches. Nevertheless, some works have proposed some representation to cope with this issue. For instance, (Ihsan et al., 2020) uses a superpixel label map as an input image to a neural network to extract meaningful information for clothing parsing application. (He et al., 2015) presents the SuperCNN as a deep neural network approach for salient object detection. It uses superpixels to describe two 1-D sequences of colors in order to reduce the computational burden. Nonetheless, neither of the existing approaches effectively encodes deep learning features for each superpixel.

2.2 Color Transfer

Transferring the right colors requires computing meaningful similarities between the reference and the target images. These similarities must preserve important textures and structures of the target image. Most works on color transfer have focused on choosing the characteristics on which to compute similarities. These characteristics can be hand-crafted or learned using deep learning methods. The first one extracts image features by relying on manually predefined descriptors (*i.e.*, HOG (Dalal and Triggs, 2005), SIFT (Lowe, 2004)); however there is no guarantee that the descriptors are well suited for the task. The second solves this issue by learning the features from image dataset and leveraging on a training procedure, nonetheless feature dimensionality increases enforcing the usage on low-resolution images. Features similarities can be matched using global information of the images (*i.e.*, color histograms); or local information such as matching small regions on the images (*i.e.*, cluster segmentation, superpixel decomposition). In the literature, color transfer techniques can be classified into three classes: classic global-based methods, classic local-based methods, and deep learning methods.

Global Methods: consider global color statistics without any spatial information. It was initially introduced in (Reinhard et al., 2001) which uses basics statistical tools (*i.e.*, mean, standard deviation) to match target and reference color information. (Pitié and Kokaram, 2007) (Xiao and Ma, 2006) extend color matching on different color spaces to find an optimal color mapping between the images. (Frigo et al., 2015) (Ferradans et al., 2013) propose a global illuminant matching based on optimal transport color transfer for enforcing artifacts-free results. More complex methods such as (Murray et al., 2012) rely on Gaussian Mixture Models to create compressed signatures that ensure a compact representation of color characteristics between images. Nevertheless, as mentioned in (Pitié, 2020), these methods fail to ensure spatial consistency on resulting colors when content change (*i.e.*, transferring day and night images).

Local Methods: relies on spatial color mappings (*i.e.*, segmentation, clustering) to match local regions of the target image and the reference image. (Liu et al., 2016) uses superpixel level style-related and style-independent feature correspondences. (Arbelot et al., 2017) implement a texture-based framework for matching local correspondence. Alternatively, (Tai et al., 2005) uses a probabilistic segmentation in order to impose spatial and color smoothness among local regions. Still, the method does not provide control over the matched superpixel. (Giraud et al., 2017) overcomes this limitation by proposing a constrained approximate nearest neighbor (ANN) patches and a color fusion framework on superpixels. However, in this type of local methods target and reference images requires to share strong similarities.

Deep Learning Methods: brings to the matching semantic-related characteristics from the target image and reference image. Recently (Lee et al., 2020) propose a deep neural network architecture that leverages on color histogram analogy for color transfer. The latter uses target and reference histograms as input to exploit global histogram information over a target input image. (He et al., 2019) relies on semantically meaningful dense correspondence between images. Nonetheless, this type of methods relies on pure semantic features (low-level features), which leads to imprecise results if images from a different scene or instances are used.

3 METHOD

In this section, we present our superpixel based framework to match high-resolution features between two RGB images I_T and I_R of size $\mathbb{R}^{H\times W\times 3}$. In the following, we will refer to I_T as the target image and I_R as the reference image to be consistent with the color fusion application.

3.1 Super-features Encoding

Let f_{T_ℓ} and f_{R_ℓ} be feature maps from a convolutional neural network at layer ℓ of I_T and I_R respectively. In the following, we will consider features coming from pre-trained deep convolutional networks (see Figure 3), but our method could be applied to other types of hand-crafted features. More precisely, we focus on features extracted at the first three layers of a deep network, as they provide a long range of low-level features that suit diverse types of images. These feature maps then have high dimensions, typically the same size as the input image, times C channels with $H\times W\times C$ where $C = 64$, 128 or 256 for example.

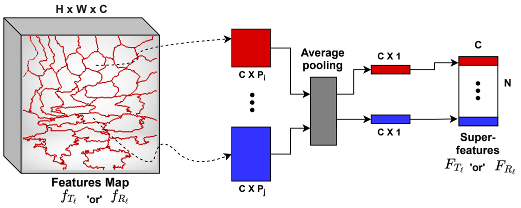


Figure 1: Diagram of our super-features encoding proposal (SFE). This proposal takes as input a feature map of size $H\times W\times C$, in which each superpixel is extracted and encoded in vectors of size $C\times P_i$ pixels. Afterward, the vectors are pooled channel-wise and, finally, stacked in the super-features matrix F with size $C\times N$ number of superpixels.

A critical drawback of using high-resolution features for matching operations is the high computational complexity. Let the number of features in a feature map be $D = H\times W\times C$, then the complexity of the pixel-wise similarity computation is $O\left(D^2\right)$. To solve this quadratic complexity problem, we implement an encoding layer based on superpixel representation. We first generate a superpixel map using a superpixel decomposition algorithm on the initial color images. Let us denote the target superpixel map by S_T, and the reference one by S_R. Each of these maps contains N_T and N_R superpixel respectively with P_i pixels each, where i is the superpixel index. Next, we extract features of size $C\times P_i$ for each superpixel. These extracted features are then pooled spatially by averaging channel-wise and stacked as a matrix of size $C\times N$ called super-features F. Figure 1 illustrates this process. To sum up, the initial feature maps (f_{T_ℓ} and f_{R_ℓ}) pass from size $H\times W\times C$ to super-features encoding (F_{T_l} and F_{R_l}) of size $N_T\times C$ and, $N_R\times C$, making feasible operations such as correlation calculation between large deep neural networks features.

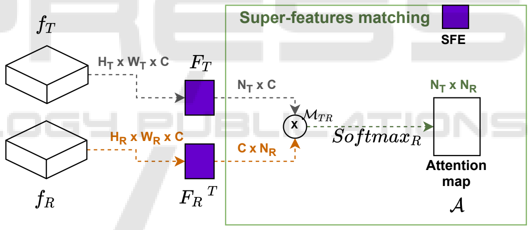


Figure 2: Diagram of our super-features matching (SFM). This layer takes a reference feature map f_R and a target feature map f_T as an input, and outputs an attention map at superpixel level by means of a non-local operation.

3.2 Super-features Matching

Our super-features provide a compact encoding to compute the correlation between high-resolution deep learning features. Here, we take inspiration from the attention mechanism (Zhang et al., 2019) to achieve a robust matching between target and reference super-features. The process is illustrated in Figure 2. Mainly, we exploit non-local similarities between the target and the reference super-features by computing the attention map at layer ℓ as:

$$\mathcal{A}_\ell = \mathrm{softmax}_{\mathrm{R}_\ell}(\mathcal{M}_{T_\ell R_\ell}/\tau). \tag{1}$$

The softmax$_{\mathrm{R}}$ operation normalizes row-wise the input into probability distributions, proportionally to

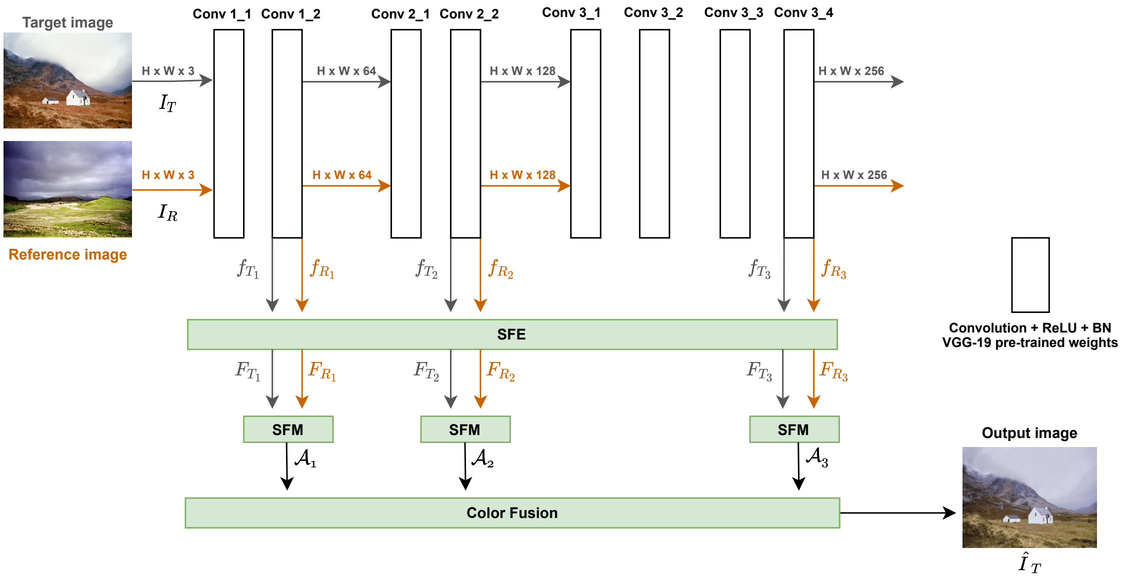


Figure 3: Diagram of our method using the first three levels of a modified VGG-19 architecture as our feature extractor. In our method, we remove max-pooling layers from the baseline VGG-19 architecture to capture similarities between high-resolution feature maps ($H \times W \times C_\ell$). Also, the diagram presents our two new blocks, the super-features encoding block (SFE) and the super-features matching block (SFM).

the number of target superpixels N_R. Then, the final attention map $\mathcal{A}$ is the weighted sum of the attention maps at each layer ℓ:

$$\mathcal{A} = \frac{\sum_{\ell=1}^{3} \omega_\ell \mathcal{A}_\ell}{\sum_{\ell=1}^{3} \omega_\ell}. \tag{2}$$

The matrix $\mathcal{M}_{TR}$ is a correlation matrix between the target and reference super-features and is computed as:

$$\mathcal{M}_{T_\ell R_\ell}(i,j) = \frac{(F_{T\ell}(i) - \mu_{T_\ell}) \cdot (F_{R_\ell}(j) - \mu_{R_\ell})}{\left\|F_{T_\ell}(i) - \mu_{T_\ell}\right\|_2 \left\|F_{R_\ell}(j) - \mu_{R_\ell}\right\|_2}, \tag{3}$$

where μ_T and μ_R are the mean of each super-feature. We found that this normalization keeps correlation values less sensitive to changes on τ for different images. The attention map (1) is the same non-local operator as the one proposed by (Zhang et al., 2019). However, their computation requires low-resolution features due to the inherent quadratic complexity problem (as mentioned in Section 3.1).

We solve this complexity problem thanks to our super-features encoding approach. Let $n = H \times W$ be the number of pixels in an image. Then, the number of features in a deep learning feature map is $D = n \times C$ which translate into a computational complexity of $O(D^2) = O(n^2 C^2)$. In contrast, with our novel super-features encoding, if we set the number of superpixels in the order of $\sqrt{n}$, then instead we rewrite with $D_s = \sqrt{n} \times C$, resulting in $O(D_s^2) = O(n \times C^2)$. As $C \ll n$ can be ignored, we go from a quadratic to a linear complexity operation $O(n)$. As a result, we can incorporate the correlation operation on large deep learning features from both target and reference images. Conversely, (Zhang et al., 2019) can only rely on deep-level features, usually the bottleneck features (*i.e.*, $H/8 \times W/8 \times C$) for similarities calculation.

4 APPLICATION TO COLOR TRANSFER

We now present our color transfer method. It consists of three blocks: 1) super-features encoding (SFE), 2) super-features matching (SFM), and 3) color fusion framework. The process is illustrated in Figure 3.

Our objective is to transfer colors from a reference I_R to a target image I_T. Concretely, this will be done by passing colors from I_R to I_T based on pairwise feature-related similarities.

To match colors at superpixel level, we rely on the attention map $\mathcal{A}$ and the average of each superpixel color. Specifically, we apply our attention map as a soft-weight on the average colors, resulting in a smooth correspondence.

Figure 4 shows a direct super-features matching between the target and reference images from Figure 3. This direct matching uses the weighted average color of its correspondence to replace the target's superpixels colors. Each row depicts the impact of different high-resolution feature maps from the first

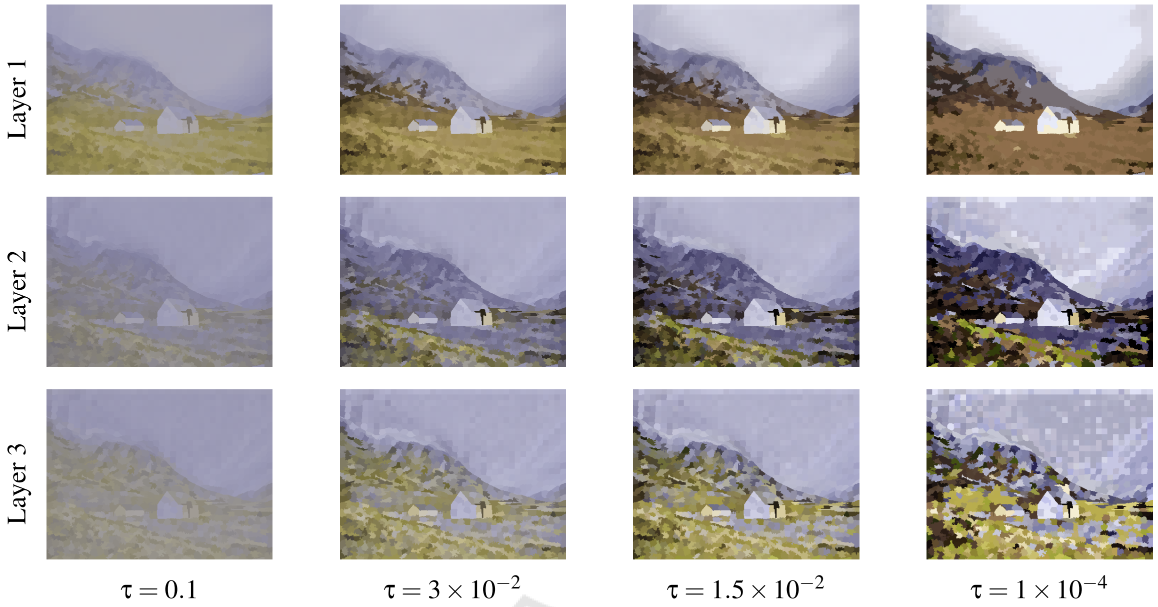

Layer 1
Layer 2
Layer 3
$\tau = 0.1$
$\tau = 3 \times 10^{-2}$
$\tau = 1.5 \times 10^{-2}$
$\tau = 1 \times 10^{-4}$

Figure 4: Direct super-features matching using different τ values. Each of the rows depicts our results for direct matching using super-features from the first, second, and third layers. The use of the first-level features helps to preserve fine details and ensures color consistency. However, the second and third layers bring a more colorful and diverse matching between target and reference super-features.

three levels of a pre-trained VGG-19. The first level (first row) brings fine details and spatial and color consistency onto the direct matching, while deeper features (second and third row) seem more sensitive to color features. This Figure 4 also illustrates (in each column) the influence of temperature τ onto the superpixel attention map. We can see that the probability distribution is over-smoothed (*i.e.*, gray average colors) for larger values of τ (*i.e.*, $\tau = 0.1$), meaning that several reference super-features match one target super-feature. Otherwise, a small τ value results in a hard one-to-one matching between a target and reference super-features (*i.e.*, $\tau = 1 \times 10^{-4}$).

4.1 Color Fusion Framework

Direct superpixel matching by averaging colors is not sufficient to obtain visually satisfying results. Image details are indeed lost at superpixel level (*i.e.*, door, windows, etc., in Figure 4). Therefore we need to transfer color at pixel level from our superpixel matching.

For clarity in further equations, we denote the position and color centroids of a superpixel j in an image I as:

$$\bar{X}(j) = \frac{\sum_{p \in S(j)} p}{P_j}$$

and

$$\bar{I}(j) = \frac{\sum_{p \in S(j)} I(p)}{P_j}$$

respectively, where P_j is the number of pixels in superpixel j.

Inspired by the formulation of (Giraud et al., 2017), we compute the new value $\hat{I}_T(p)$ of each pixel p of the target as a weighted average of reference superpixel representative colors:

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

The weight matrix W depends firstly on the distance between pixel p and all target superpixel as in (Giraud et al., 2017), and secondly, on our attention map:

$$W(p,j) = \sum_{i=1}^{N_T} d(p,i)\mathcal{A}(i,j). \tag{5}$$

The intuition behind the attention map is the addition of more relevant information about reference super-features into the transfer process. The distance between pixel p and superpixel centroids is computed over both positions and colors with a Mahalanobis-like formulation:

$$d(p,i) = e^{\left(-\frac{(V_T(p)-\bar{V}_T(i))^T \Sigma_i^{-1} (V_T(p)-\bar{V}_T(i))}{\sigma_g}\right)}, \tag{6}$$

with position and color vectors being $V(p) = [p, I(p)]$ and $\bar{V}_T(j) = [\bar{X}_T(j), \bar{I}_T(j)]$, and the spatial and colorimetric covariances of pixels in superpixel i:

$$\Sigma_i = \begin{pmatrix} \delta_s^2 \operatorname{Cov}(p) & 0 \\ 0 & \delta_c^2 \operatorname{Cov}(I(p)) \end{pmatrix}. \tag{7}$$

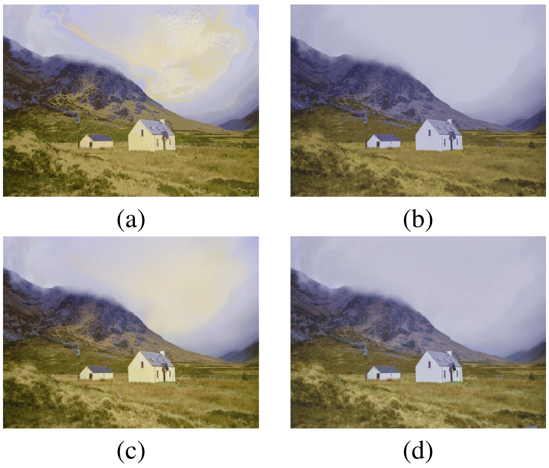


Figure 5: Color fusion framework results. (a) (Giraud et al., 2017) color fusion result. (b) Our color fusion result. (c) (Giraud et al., 2017) result + regrain. (d) Our result + regrain. The regrain algorithm is from (Pitié et al., 2005).

Parameters δ_s and δ_c weight the influence of color and spatial information, respectively.

Finally, as in (Giraud et al., 2017), after color fusion we apply a post-processing step using a color regrain algorithm (Pitié et al., 2005), which eventually matches the color distribution of I_R and the gradient of I_T. Figure 5 presents an example of our color transfer framework compared to the result of (Giraud et al., 2017). Visually, our results present better spatial consistency of colors. For instance, the sky on our results has more natural smooth color transitions compared to non-natural ones with (Giraud et al., 2017) (*i.e.*, yellow to blue).

5 RESULTS

In this section, we first present the implementation details used to validate our method and then provide a detailed qualitative comparison between our results and three state-of-the-art color transfer approaches.

5.1 Implementation Details

Superpixel segmentation is done using the SLIC algorithm (Achanta et al., 2012), in which the number of superpixel depends on the actual size of the image. Experimentally, we set the number of superpixel as $3 \times \sqrt{n}$ where n is the number of pixels in the current image.

To build feature maps, we rely on a modified pretrained VGG-19 (Simonyan and Zisserman, 2015) as our texture and color characteristics extractor, due to its simplicity and its 95.24% classification accuracy on the ImageNet Top-5 classes. The main modification was the removal of max-pooling layers from the first three levels (see Figure 3) as it highly improves matching results compared to using upsampling on max-pooled feature maps at Conv2_2 and Conv3_4 from the baseline VGG-19. Figure 6 exemplifies that matching low-resolution features does not preserve details nor retains color coherence, especially when going deeper into the architecture. Also, note that our approach can work with other types of CNN architectures regardless of their features dimensions.

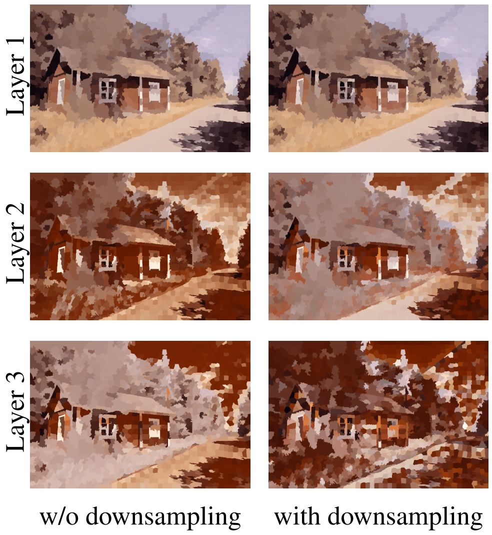


Figure 6: Effect on direct super-features matching (*i.e.*, before color fusion) using high-resolution feature maps (w/o downsampling) and using low-resolution feature maps (with downsampling). Target and reference images are presented in Figure 7.

In order to choose an optimal temperature τ value, we experimented on different images at distinct temperatures. Empirically, we obtain satisfying results using $\tau = 0.015$ and $\omega = 1$. In addition, all experiments have been run with $\delta_s = 10$ and $\delta_c = 0.1$, as recommended by (Giraud et al., 2017) to favor spatial consistency.

5.2 Analysis on Different Layers

Our SFE and SFM blocks support any CNN features map dimensions, so choosing to work with one or coupling many of these features maps depends mostly on the application. In this experiment, we analyze the effects of using separately each of the first three feature map outputs for the color transfer application.

From the different columns of Figure 7 we can retain that each layer focuses on different aspects of the image, resulting in color variations of the same target image. Specifically, in the first row (house image),

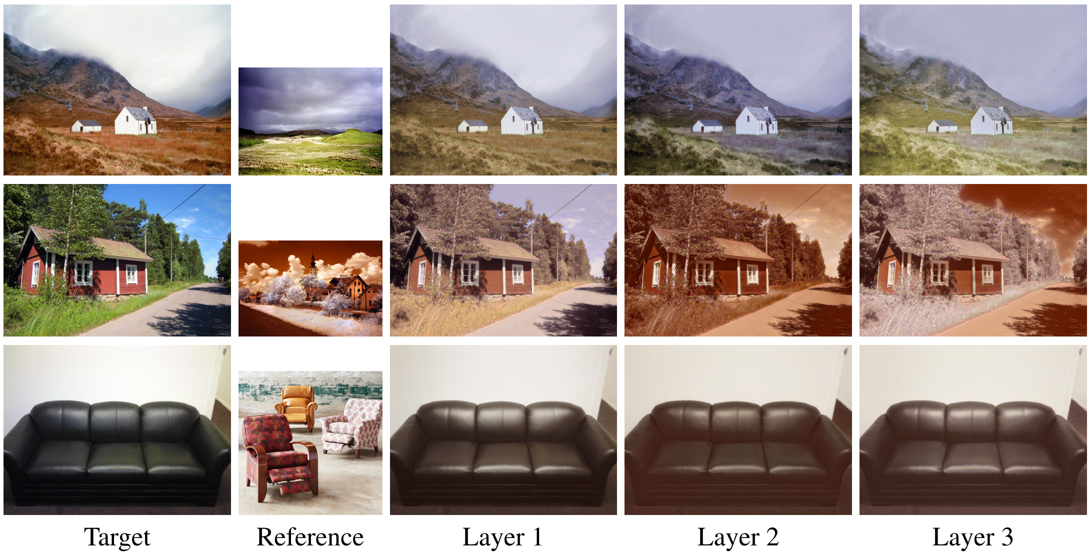


Figure 7: Results of our method using each of the three layers separately with $\tau = 0.015$.

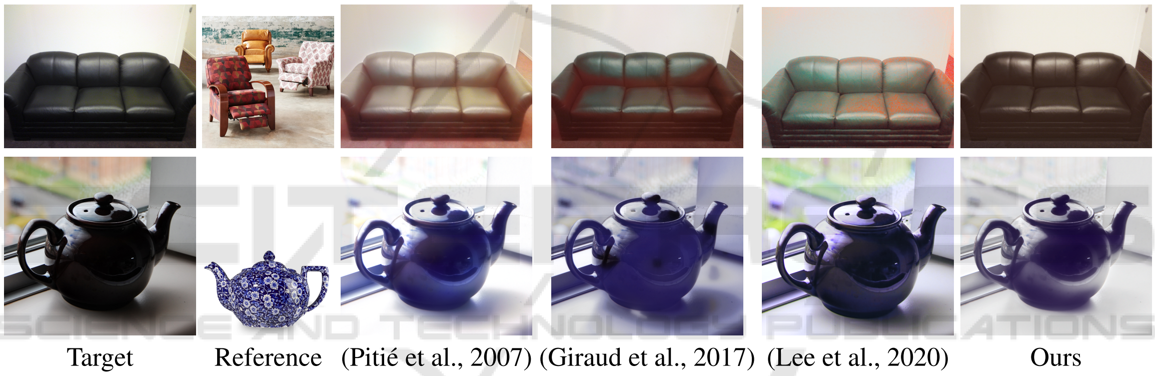


Figure 8: Comparison of color transfer results on indoor images. We compare our method with three different state-of-the-art approaches: (Pitié et al., 2007) color distribution grading, (Giraud et al., 2017) color fusion based on superpixel representation and, (Lee et al., 2020) deep learning-based color histogram analogy.

we can see that the deeper layer (layer 3) focuses on the grass color while the second layer focuses on the mountain color. In the second and third rows, layers 2 and 3 bring stronger colors from the reference image, but the results are still unrealistic. In the first layer, the recovered image seems more natural; however, most of the colors transferred from the reference image are opaque or not transferred.

Finally, we decided to combine all three layers as each of them brings important feature information to achieve pleasant and realistic images for this color transfer application.

5.3 Comparison

We compare our method against three approaches: (Pitié et al., 2007) which proposes an automated color transfer based on color distributions; (Lee et al., 2020) which implements a color transfer approach based on color histogram analogy using a deep neural network; and (Giraud et al., 2017) which implements the color fusion framework by leveraging on its proper superpixel decomposition. All three mentioned approaches have been considered state-of-the-art in color transfer, and have open-source codes for a fair comparison. Each method has been run with its default parameters.

Results comparing the three methods are shown in Figures 8, 9 and 10. Overall, our results (last column) have more visually pleasant colors and consistency in image texture, providing more realistic color transfers with respect to the other methods. Figure 8 shows that our approach correctly matches and transfers natural colors from indoor images, avoiding color bleeding (blue color on the wall) as shown in (Giraud et al., 2017), (Lee et al., 2020) and partially

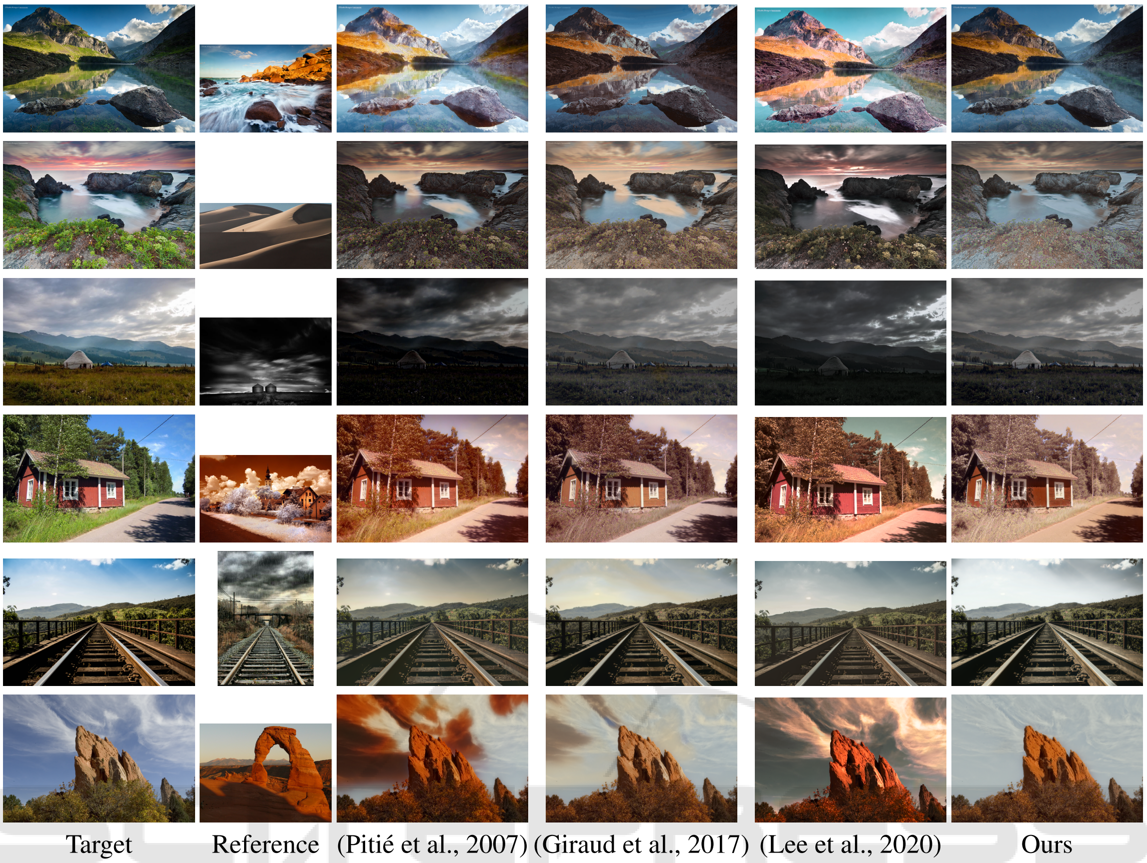
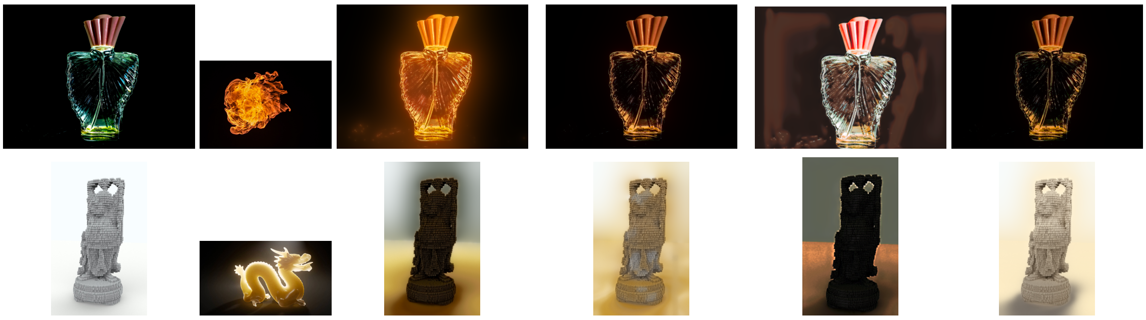


Figure 9: Comparison of color transfer results on outdoor images.

in (Pitié and Kokaram, 2007) results. For outdoor images shown in Figure 9, we observe that (Pitié et al., 2007) and (Lee et al., 2020) can suffer from oversaturation of the illumination on some of their results (first, fourth, sixth images). Although this problem does not appear in (Giraud et al., 2017) some of its results present visible unnatural effects on sky colors such as a halo effect (first image) and yellowish marks (fifth, sixth images). Our results for outdoor images overcome these issues thanks to the robust matching on high-resolution superpixel deep features, which ensures color consistency and spatially coherent colors across the resulting images. Figure 10 presents images with no background (studio shooting like images). In this case, results by (Pitié et al., 2007) and (Lee et al., 2020) show unnatural color effects around the bottle and background. On the other hand, (Giraud et al., 2017) approach and ours achieve pleasant color results without over-saturation nor artifacts on the resulting image. Lastly, our method correctly transfers colors to the statue image resulting in the most visually satisfying and realistic results with regards to all compared methods.

6 CONCLUSION

This paper proposed the novel super-features encoding block (SFE) and the super-features matching (SFM) block that successfully encodes and matches high-resolution deep learning features from different images using superpixel decomposition. We validate these two blocks on the problem of color transfer; for doing that, we update the color fusion framework initially proposed by (Giraud et al., 2017) to consider our attention map, which provides texture and color knowledge from the reference image onto the final color transfer step. Finally, our method achieves more visually consistent and realistic results in comparison to the three state-of-the-art methods considered. Work is underway on applying our new super-features encoding and matching blocks to other image editing applications. Another future line of research aim at including this block in an end-to-end deep learning architecture.



Target Reference (Pitié et al., 2007) (Giraud et al., 2017) (Lee et al., 2020) Ours

Figure 10: Comparison of color transfer results on images with no background.

ACKNOWLEDGEMENTS

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

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.

Arbelot, B., Vergne, R., Hurtut, T., and Thollot, J. (2017). Local texture-based color transfer and colorization. *Computers & Graphics*, 62:15–27.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Buades, A., Coll, B., and Morel, J.-M. (2005). A non-local algorithm for image denoising. In *Conference on Computer Vision and Pattern Recognition*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*.

Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., and Aujol, J.-F. (2013). Regularized discrete optimal transport. In *Scale Space and Variational Methods in Computer Vision*.

Frigo, O., Sabater, N., Demoulin, V., and Hellier, P. (2015). Optimal transportation for example-guided color transfer. In *Asian Conference on Computer Vision*.

Fulkerson, B., Vedaldi, A., and Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision*.

Giraud, R., Ta, V.-T., and Papadakis, N. (2017). Superpixel-based color transfer. In *International Conference on Image Processing*.

Glasner, D., Bagon, S., and Irani, M. (2009). Super-resolution from a single image. In *International Conference on Computer Vision*.

He, M., Liao, J., Chen, D., Yuan, L., and Sander, P. V. (2019). Progressive color transfer with dense semantic correspondences. *ACM Transactions on Graphics*, 38(2).

He, S., Lau, R., Liu, W., Huang, Z., and Yang, Q. (2015). SuperCNN: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision*, 115:330–344.

Ihsan, A., Chu Kiong, L., Naji, S., and Seera, M. (2020). Superpixels features extractor network (SP-FEN) for clothing parsing enhancement. *Neural Processing Letters*, 51:2245–2263.

Lee, J., Son, H., Lee, G., Lee, J., Cho, S., and Lee, S. (2020). Deep color transfer using histogram analogy. *The Visual Computer*, 36(10):2129–2143.

Liu, J., Yang, W., Sun, X., and Zeng, W. (2016). Photo stylistic brush: Robust style transfer via superpixel-based bipartite graph. In *International Conference on Multimedia and Exposition*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Murray, N., Skaff, S., Marchesotti, L., and Perronnin, F. (2012). Toward automatic and flexible concept transfer. *Computers & Graphics*, 36(6):622–634.

Pitié, F. (2020). Advances in colour transfer. *IET Computer Vision*, 14:304–322.

Pitié, F. and Kokaram, A. (2007). The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *European Conference on Visual Media Production*.

Pitié, F., Kokaram, A., and Dahyot, R. (2005). Towards automated colour grading. In *IEEE European Conference on Visual Media Production*.

Pitié, F., Kokaram, A. C., and Dahyot, R. (2007). Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1–2):123–137.

Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *ACM Transactions on Graphics*, 21(5):34–41.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Tai, Y.-W., Jia, J., and Tang, C.-K. (2005). Local color transfer via probabilistic segmentation by expectation-maximization. In *Conference on Computer Vision and Pattern Recognition*.

Tighe, J. and Lazebnik, S. (2010). Superparsing: Scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision*.

Van den Bergh, M., Boix, X., Roig, G., and Van Gool, L. (2015). SEEDS: Superpixels extracted via energy-driven sampling. *International Journal of Computer Vision*, 111(3):298–314.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition*.

Wexler, Y., Shechtman, E., and Irani, M. (2004). Space-time video completion. In *Conference on Computer Vision and Pattern Recognition*.

Xiao, X. and Ma, L. (2006). Color transfer in correlated color space. In *International Conference on Virtual Reality Continuum and its Applications*.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Conference on Computer Vision and Pattern Recognition*.

Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A., and Chen, D. (2019). Deep exemplar-based video colorization. In *Conference on Computer Vision and Pattern Recognition*.

APPENDIX

Figure 11 shows additional resulting images from the three state-of-the-art methods and our method.

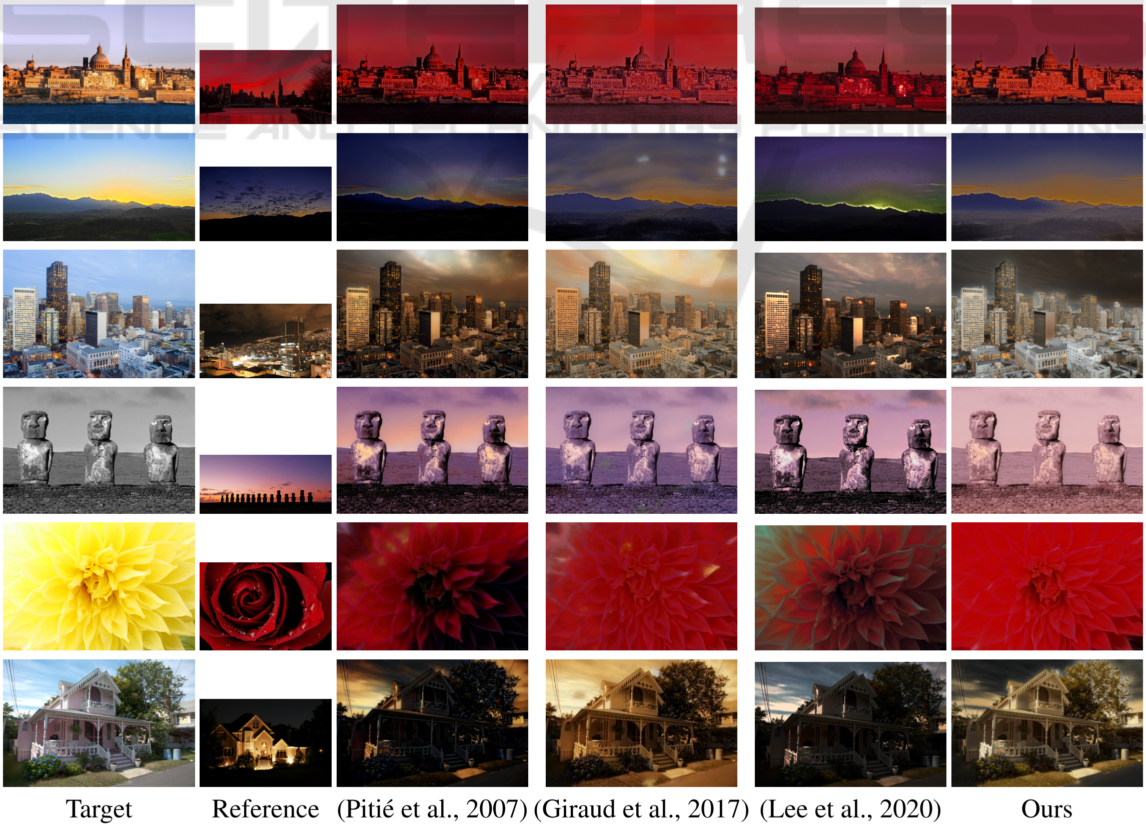


Figure 11: More color transfer results.