Super-attention for exemplar-based image colorization

Hernan Carrillo, Michaël Clément, Aurélie Bugeau

Super-attention for exemplar-based image colorization

Hernan Carrillo[1], Michaël Clément[1], and Aurélie Bugeau[1,2]

[1] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, Talence, France
[2] Institut universitaire de France (IUF)
{hernan.carrillo-lindado, michael.clement, aurelie.bugeau}@labri.fr

Abstract. In image colorization, exemplar-based methods use a reference color image to guide the colorization of a target grayscale image. In this article, we present a deep learning framework for exemplar-based image colorization which relies on attention layers to capture robust correspondences between high-resolution deep features from pairs of images. To avoid the quadratic scaling problem from classic attention, we rely on a novel attention block computed from superpixel features, which we call super-attention. Super-attention blocks can learn to transfer semantically related color characteristics from a reference image at different scales of a deep network. Our experimental validations highlight the interest of this approach for exemplar-based colorization. We obtain promising results, achieving visually appealing colorization and outperforming state-of-the-art methods on different quantitative metrics.

Keywords: Attention mechanism · Colorization · Superpixel.

1 Introduction

Colorization is the process of adding plausible color information to grayscale images. Ideally, the result must reach a visually pleasant image, avoiding possible artifacts or improper colors. Colorization has gained importance in various areas such as photo enhancement, the broadcasting industry, films post-production, and legacy content restoration. However, colorization is an inherently ill-posed problem, as multiple suitable colors might exist for a single grayscale pixel, making it a challenging task. This ambiguity on the color decision usually leads to random choices or undesirable averaging of colors. In order to overcome these issue, several colorization approaches have been proposed, and can be categorized into three types: automatic learning-based colorization, scribble-based colorization, and exemplar-based colorization.

Automatic learning-based methods [1–4] bring fairly good colors to grayscale images by leveraging on color priors learned from large-scale datasets. Nonetheless, this type of methods lacks from user's decision. In scribble-based methods [5–9], the user initially adds color scribbles in different image regions, later propagated to the whole image using similarities between neighboring pixels. The last

group of methods [10–15] transfers color from one (or many) reference color image to the input grayscale, thus biasing the final image's color to the user's preference. However, when semantic similarities do not exist between the reference image and input image, the efficacy of these exemplar-based methods highly decreases. Therefore, there is an interest in adding information from large-scale datasets and associate it with semantic information of a reference image to no longer depend on just naive pixel-wise matching. To deal with the aforementioned challenge, [16] proposed an attention mechanism for image colorization, mainly by calculating non-local similarities between different feature maps (input and reference images). However, attention mechanisms come with a complexity problem, namely a quadratic scaling problem, due to its non-local operation. This is why it has to be applied on features with low dimensions. On the other hand, low-resolution features lack of detailed information for calculating precise and robust pixel-wise similarities. For instance, low-resolution deep features mainly carry high-level semantic information related to a specific application (*i.e.*, segmentation, classification) that can be less relevant for high-resolution similarity calculation or matching purposes.

In this work, we propose a new exemplar-based colorization method that relies on similarity calculation between high-resolution deep features. These features contain rich low-level characteristics which are important in the colorization task. To overcome the complexity issue, we extend to the colorization task the super-attention block presented in [17] that performs non-local matching over high-resolution features based on superpixels. The main contributions made are:

- A new end-to-end deep learning architecture for exemplar-based colorization improving results over state-of-the-art methods.
- A multiscale attention mechanism based on superpixels features for reference-based colorization.
- A strategy for choosing relevant target/reference pairs for the training phase.

2 Related work

Colorization with deep learning. Automatic learning-based colorization methods use large-scale datasets to learn a direct mapping between each of the grayscale pixels from the input image to a color value. In the earliest work [18], a grayscale image is fed to a neural network that predicts UV chrominance channels from the YUV luminance-chrominance color space by a regression loss. In [19] a grayscale image is given as input to a VGG network architecture, and for each of the pixels a histogram of hue and chroma is predicted, which serve as a guide to the final colorization result. Other deep learning architectures have been used, such as Generative Adversarial Networks (GANs). For instance, [4] proposes to couple semantic and perceptual information to colorize a grayscale image. This is done using adversarial learning and semantic hints from high-level classification features. A different approach is proposed by [20] by defining an autoregressive problem to predict the distribution of every pixel's colors conditioned on the previous pixel's color distribution and the input grayscale image, and addresses

the problem using axial attention [21]. Overall, automatic learning-based methods reduce colorization time, in contrast to purely manual colorization; however, this type of methods lacks user-specific requirements.

Exemplar-based methods. Another family of colorization methods is the Exemplar-based methods, which requires a reference color image from which colors can be transferred to a target grayscale image. Early work uses matching between luminance and texture information, assuming that similar intensities should have similar colors [10]. Nonetheless, this approach focuses on matching global statistics while ignoring coherent spatial information, leading to unsatisfactory results. In [11], a variational formulation is proposed where each pixel is penalized for taking a chrominance value from a reduced set of possible candidates chosen from the reference image. In more recent approaches, He *et al.* [13] was the first deep learning approach for exemplar-based colorization. This is done by using the PatchMatch [22] algorithm to search semantic correspondences between reference and target images and leveraging the colors learned from the dataset. The authors of [23], inspired by the style transfer techniques, used AdaIN [24] to initially generate a colorized image to be refined by a second neural network. Finally, [13, 25] use a deep learning framework that leverages on matching semantic correspondences between two images for transferring colors from the reference color image to the grayscale one. However, these methods mostly rely on pure semantic (low-level) features which leads to imprecise colorization results.

Attention layers. Recently, attention mechanisms have been a hot topic in particular with transformer architectures [26]. These networks include (self-)attention layers as their main block. The attention layers learn to compute similarities called attention maps between input data or embedded sequences by means of a non-local matching operation [27, 28]. Attention layers have been use for colorization in some recent works. For example, [20] presents a transformer architecture capable of predicting high-fidelity colors using the self-attention mechanism. Then, [16] extends the attention mechanism for searching similarities between two different feature maps (target and reference) applied to colorization application, achieving interesting results. Lately, [15] proposes an attention-based colorization framework that considers semantic characteristics and a color histogram of a reference image as priors to the final result. Finally, [29] implements the axial attention mechanism to guide the transfer of the color's characteristics from the reference image to the target image. Even though these methods provide promising colorization results, they may suffer from the quadratic complexity problem of attention layers, and therefore can only perform non-local matching at low resolution. In this work, by using attention layers at the superpixels level, we allow this matching to be done at full resolution.

3 Colorization framework

Our objective is to colorize a target grayscale image T by taking into consideration the characteristics of a color reference image R. Let $T_L \in R^{H\times W\times 1}$be

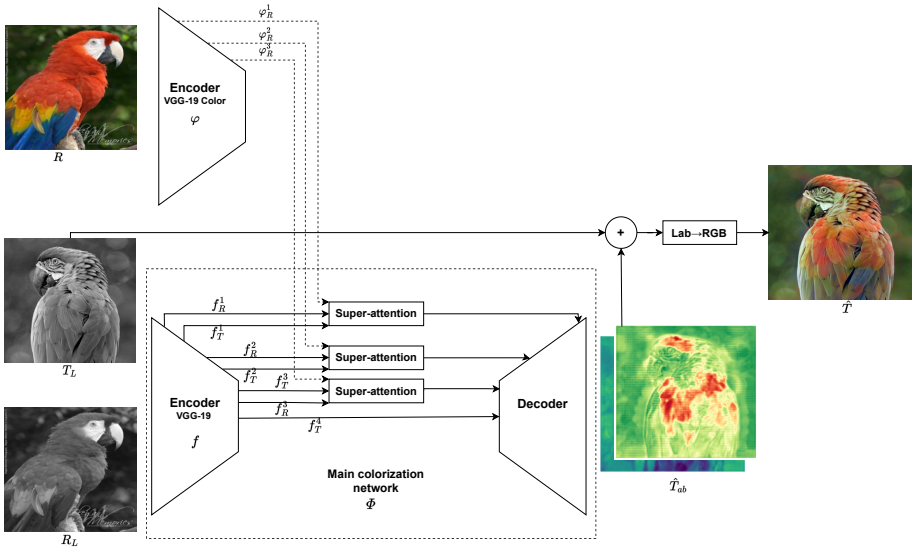


Fig. 1. Diagram of our proposal for exemplar-based image colorization. The framework consists of two main parts: 1) the color feature extractor φ extracts multi-level feature maps φ_R^ℓ from a color reference image R, and 2) the main colorization network Φ, which learns to map a luminance channel image T_L to its chrominance channels $\hat{T}_{ab}$ given a reference color image R. This colorization guidance is done by super-attention modules, which learn superpixel-based attention maps from the target and reference feature maps from distinct levels f_T^ℓ and f_R^ℓ respectively.

the luminance component of the target, specifically the channel L from the color space Lab in CIELAB, and $R_{Lab} \in R^{H \times W \times 3}$ the color reference image in CIELAB color space. In this work, we choose to work with the luminance-chrominance CIELAB color, as it is perceptually more uniform than other color spaces [30].

In order to colorize the input image, we train a deep learning model Φ that learns to map a grayscale image to its chrominance channels (ab) given a reference color image:

$$\hat{T}_{ab} = \Phi(T_L \,|\, R). \tag{1}$$

Our proposed colorization framework is composed of two main parts: an external feature extractor φ for color images, and the main colorization network Φ, which relies on super-attention blocks applied at different levels (see Figure 1). The main colorization network Φ is based on a classical Unet-like encoder-decoder architecture [31], with the addition of the super-attention blocks [17] which allows transferring color hints from the color reference image to the main colorization network. Next, this network predicts the target's chrominance chan-

nels $\hat{T_{ab}}$. And, as a final stage, target luminance and chrominance channels are concatenated into $\hat{T}_{Lab}$ and then converted to the RGB color space $\hat{T}$ using Kornia [32].

3.1 Main colorization network

The main colorization network Φ aims to colorize a target grayscale image based on a reference image when semantic-related content appears, or pulling back to the learned model when this relation is not present in certain objects or regions between the images. The colorization network receives target T_L and reference R_L as input images to obtain deep learning feature maps f_T^ℓ, f_R^ℓ from the ℓ^{th} level of the architecture. In the same sense, the reference color image R is fed to the color feature extractor, which is a frozen pre-trained VGG19 encoder that retrieves multiscale feature maps φ_R^ℓ. Specifically, feature maps are extracted from the first three levels of the encoders. Then, all extracted features are fed to the super-attention blocks, where a correlation is computed between target and reference encoded features. Next, the content is transferred from the reference features to the target by multiplying the similarity matrix and the color reference features. Then, the color features coming from the super-attention modules are transferred to the future prediction by concatenating them to the decoder features. Finally, the decoder predicts the two (ab) chrominance channels $\hat{T}_{ab}$ that are then concatenated to the target luminance channel T_L, then the prediction is converted from CIELAB color space to RGB color space to provide the final RGB image $\hat{T}$. However, this conversion between color spaces arises clipping problems on RGB values as in certain cases the combination of predicted Lab values can be outside the conversion range.

3.2 Super-attention as color reference prior

The super-attention block injects colors priors from a reference image R to the main colorization network Φ. This block relies on multi-level deep features to calculate robust correspondence matching between the target and reference images. Specifically, the super-attention block is divided into two parts: The super-features encoding layer and the super-features matching layer. The super-features encoding layer provides a compact representation of high-resolution deep features using superpixels. For the colorization application, we focus on features maps extracted at the first three levels of the architecture, as they provide a long range of high and low-level features that suit content and style transfer applications [17]. Figure 2 depicts the diagram of the super-attention block where f_T^ℓ, f_R^ℓ and φ_R^ℓ are feature maps from the encoder f and the encoder φ at level ℓ of T_L, R_L and R respectively. This encoding block leverages on superpixels decomposition of the target and reference grayscale images. Each of these decompositions contain N_T and N_R superpixels, respectively, with P_i pixels each, where i is the superpixel index. Finally, the encoding is done by means of a channel-wise *maxpooling* operation resulting in super-features F of size $C \times N$, where $N \ll H \times W$ (a typical superpixel segmentation yields $N \approx \sqrt{H \times W}$

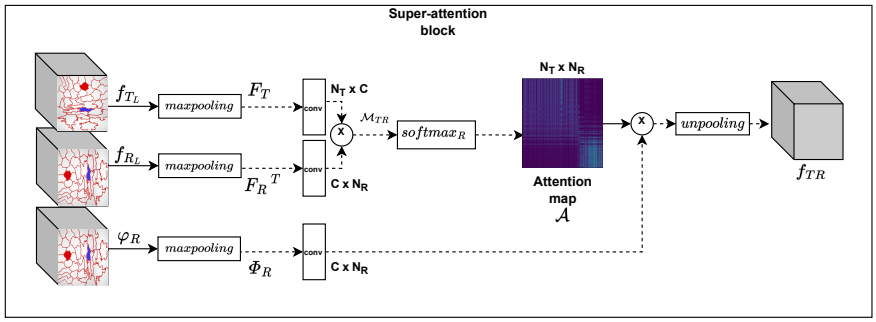


Fig. 2. Diagram of our super-attention block. This layer takes a reference luminance feature map f_R, reference color feature map φ_R and a target luminance feature map f_T, as an input, and learns an attention map at superpixel level by means of a robust matching between high-resolution encoded feature maps.

elements). In summary, this encoding part makes possible operations such as correlation between high-resolution features in our colorization framework.

The super-features matching layer computes the correlation between encoded high-resolution deep-learning features. This layer is inspired by the classic attention mechanism [16] on images to achieve a robust matching between target and reference super-features. However, contrary to [17] our similarity matrix (attention map) is learned by the model. Figure 2 illustrates the process. This layer exploits the non-local similarities between the target F_T and reference F_R super-features, by computing the attention map at layer ℓ as:

$$\mathcal{A}^\ell = \text{softmax}(\mathcal{M}^\ell{}_{TR}/\tau). \tag{2}$$

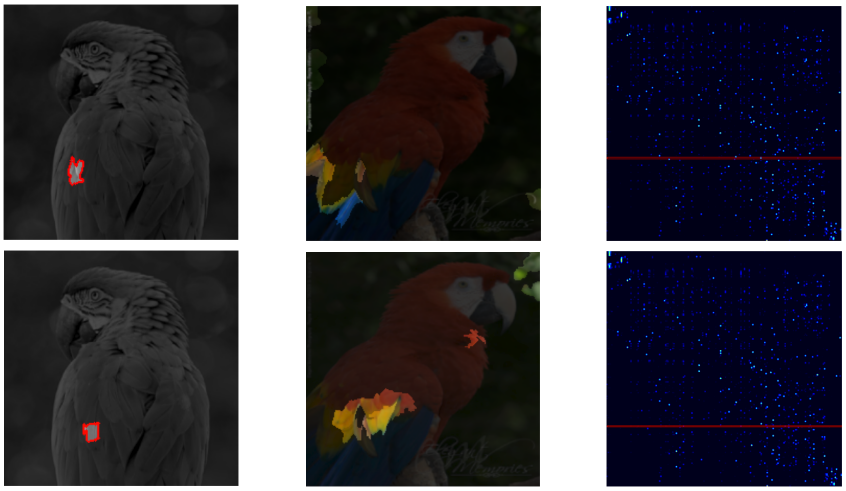
The softmax operation normalizes row-wise the input into probability distributions, proportionally to the number of target superpixels N_R. Then, the correlation matrix $\mathcal{M}_{TR}$ between target and reference super-features reads:

$$\mathcal{M}^\ell_{TR}(i,j) = \frac{\left(F^\ell_T(i) - \mu^\ell_T\right) \cdot \left(F^\ell_R(j) - \mu^\ell_R\right)}{\left\|F^\ell_T(i) - \mu^\ell_T\right\|_2 \left\|F^\ell_R(j) - \mu^\ell_R\right\|_2} \tag{3}$$

where μ_T, μ_R are the mean of each super-feature and i, j are the current superpixels from the target and reference respectively.

In terms of complexity, the super-feature encoding approach allows to overcome the quadratic complexity problem in the computation of standard attention maps. Indeed, instead of computing attention maps of quadratic size in the number of pixels, our super-attention maps are computed on a much smaller number of superpixels (*i.e.*, linear in the number of pixels). More details about this complexity reduction can be found in [17].

To illustrate the use of this block between target and reference images, Figure 3.2 shows some examples of matching using the super-attention block at the



Target Superpixel | Reference superpixel guidance | Attention map

Fig. 3. Example of guidance maps from the super attention mechanism. First column highlights one superpixel in the target; second column, the reference in which the lightness of the superpixels are scaled according to the computed attention map; third column: the attention map with in red the row corresponding to the target superpixel and that is used to generate the second column.

first level of the architecture. Mainly, it shows that for one superpixel from the target feature maps, the learned attention map successfully looks for superpixels on the reference feature maps with similar characteristics to the reference image.

3.3 Training losses

Designing the loss function in any deep learning model is one of the key parts of the training strategy. In the classical case of automatic colorization, one would like to predict $\hat{T}_{ab}$ by reconstructing the colors from the groundtruth image T_{ab}. But, this idea does not work within exemplar-based colorization as the predicted $\hat{T}_{ab}$ colors should take into account color's characteristics from a reference image $\hat{T}_{ab} = \phi(T_L \,|\, R)$. Then, the goal is to guarantee an accurate and well-grounded transfer of color characteristics to the target from the reference. In this work, we propose a coupled strategy of two loss terms, L1 smooth and LPIPS, to help reconstruct the final image.

L1 smooth Loss To close the gap between T_{ab} and $\hat{T}_{ab}$, and to restore the missing color information when reference similarities may not be found, we propose to use a reconstruction term based on Huber loss, also called L1 smooth. This loss is more robust to outliers than the L_2 loss [33], as well as avoiding averaging colors due to the multi-modal ambiguity problem on colorization [13,

16, 14]. Notice that T_{ab} and $\hat{T}_{ab}$ are assumed to be the flattened images into 1D vectors, then the L1 smooth is computed as follows:

$$L_{1_{smooth}} = \begin{cases} 0.5\left(T_{ab} - \hat{T}_{ab}\right)^2, & \text{if } \left|T_{ab} - \hat{T}_{ab}\right| < 1 \\ \left|T_{ab} - \hat{T}_{ab}\right| - 0.5, & \text{otherwise.} \end{cases} \tag{4}$$

VGG-based LPIPS Loss To encourage the perceptual similarity between the groundtruth target image T and the predicted one $\hat{T}$, we use the LPIPS loss [34]. Given a pretrained network, LPIPS is computed as a weighted L_2 distance between deep features of the predicted colorized target image $\hat{T}$ and the groundtruth T. In this work, feature maps f_T^ℓ and $\hat{f}_T^\ell$ are obtained from a pre-trained VGG network, and the loss term is computed as:

$$L_{LPIPS} = \sum_\ell \frac{1}{H^\ell W^\ell} \left\| \omega^\ell \odot \left(f_T^\ell - \hat{f}_T^\ell\right) \right\|_2^2 \tag{5}$$

where H^ℓ (resp. W^ℓ) is the height (resp. the width) of feature map f_T^ℓ at layer ℓ and ω^ℓ are weights for each feature. The features are unit-normalized in the channel dimension. Note that to compute this VGG-based LPIPS loss, both input images have to be in RGB color space and normalized on range $[-1, 1]$. As our initial prediction is in the Lab color space, to apply backpropagation it has to be converted to RGB in a differentiable way using Kornia [32].

Total loss Finally, these two loss terms, $L_{1_{smooth}}$ and L_{LPIPS}, are summed by means of different fixed weights which allow to balance the total loss. The joint total loss used on the training phase is then:

$$L_{total} = \lambda_1 L_{1_{smooth}} + \lambda_2 L_{LPIPS} \tag{6}$$

where λ_1, λ_2 are fixed weights for each of the individual losses.

Notice that some previous exemplar-based colorization methods proposed to add an additional histogram loss to favor color transfer [14, 15, 29]. As we do not want to enforce a complete transfer of all colors from the reference, but only the ones that are relevant with the source image, we have decided not to use it in our model. We provide in supplementary material additional experiments that show that adding this loss can decrease the quality of the results.

3.4 Implementation details

For the training phase, we set the weights of two terms of the loss to $\lambda_1 = 10$, $\lambda_2 = 0.1$. These values were chosen empirically to obtain a good balance between the $L_{1_{smooth}}$ reconstruction term and the LPIPS semantic term. We train the network with a batch size of 8 and for 20 epochs. We use the Adam optimizer with a learning rate of 10^-5 and $\beta_1 = 0.9$, $\beta_2 = 0.99$. Finally, our model is trained with a single GPU NVIDIA RTX 2080 Ti and using PyTorch 1.10.0.

For the super-attention modules, superpixel segmentations are calculated on the fly using the SLIC algorithm [35]. Multiscale superpixels grid are calculated using downsampled versions of the grayscale target T_L and reference R_L images. These downsamplings are performed to match the size of the feature maps from the first three levels of the encoder f.

4 Dataset and references selection

We train the proposed colorization network using the COCO dataset [36]. Unlike ImageNet, this dataset proposes a smaller quantity of images but with more complex scene structures, depicting a wider diversity of objects classes. Additionally, the dataset provides the segmentation of objects, which we rely on for our target/reference images pairs strategy. In detail, this dataset is composed of 100k images for training and 5k images for validation. For the training procedure, we resize the images to the size of 224×224 pixels.

Another key aspect of the training strategy in exemplar-based methods is to find a suitable semantic reference to the target image. To build the target-reference pairs of images, we took inspiration from [13] to design our ranking of reference images. There, they proposed a correspondence recommendation pipeline based on grayscale images. Here, our approach focuses on searching target-reference matches from a wide variety of segmented objects as well as natural scenes images. Our proposal ranks four images semantically related to each target image. First, to increase the variety of reference images within a category, we extract each meaningful object whose size is greater than 10% of the size of the current image. To retrieve the image objects, we use the segmentations provided by the dataset. Second, we compute semantically rich features from the fifth level of a pre-trained VGG-19 [37] encoder φ_T^5 and φ_R^5. Next, for each target, reference images are ranked based on the L_2 distance between these pre-computed features. Finally, during training, target-reference pairs of images are sampled using a uniform distribution with a weight of 0.25 by randomly choosing either itself (*i.e.*, the groundtruth target image is used as the reference) or the other top-4 semantically closest references.

Figure 4 shows examples of target-reference matching based on our proposal. The target images are presented in the first column, and the following columns represent its corresponding reference based on the nearest L_2 distance between feature maps. The second column (Top 1) shows the references most semantically-related to the target, while the last column (Top 4) shows the references the least semantically-related to the target.

5 Experimental validations

In this section, we present quantitative and qualitative experimental results to illustrate the interest of our proposed approach. First, we propose an analysis of our method with an ablation study comparing different architectural choices

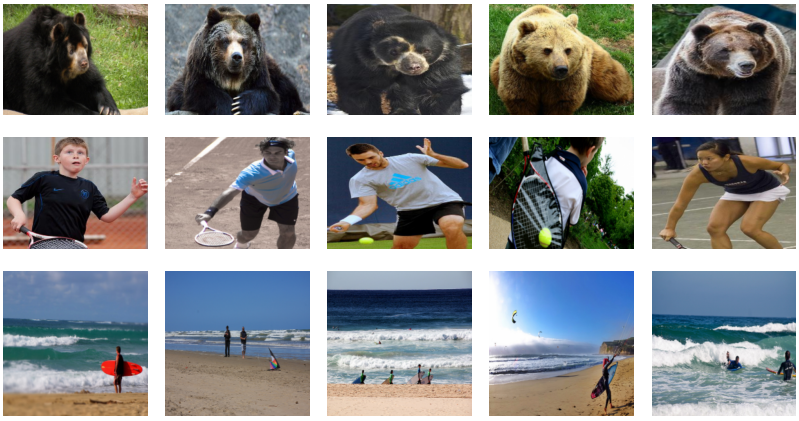


Fig. 4. Illustration of our reference selection method. The first column shows example target images, and the next four columns show the closest references in the dataset, in decreasing order of similarity.

and training strategies. Then, we compare our results to three state-of-the-art exemplar-based colorization approaches.

5.1 Analysis of the method

We start by analyzing quantitatively and qualitatively certain variants of our proposed colorization framework. Within this ablation study, we analyze two variants. The first one is a baseline colorization model without using references. It uses the same generator architecture without any attention layer. The second variant is our framework using a standard attention layer in the bottleneck of the architecture instead of our super-attention blocks. Finally, the third model is our proposed exemplar-based colorization framework, which includes the use of references with super-attention blocks in the top three levels of the network.

Evaluation metrics. To evaluate quantitatively the results of these different methods, we used three metrics. Two metrics that compare the result with the groundtruth color image, and a third metric that compare the prediction of colors with respect to the reference color image. The first one is the structural similarity (SSIM) metric [38], which analyzes the ability of the model to reconstruct the original image content. The second one is the learned perceptual image patch similarity (LPIPS) metric [34] which correlates better with the human perceptual similarity. These first two metrics (SSIM and LPIPS) evaluate the quality of the output colorization compared to the groundtruth. The third metric employed is the histogram intersection similarity (HIS) [39] which is computed between the

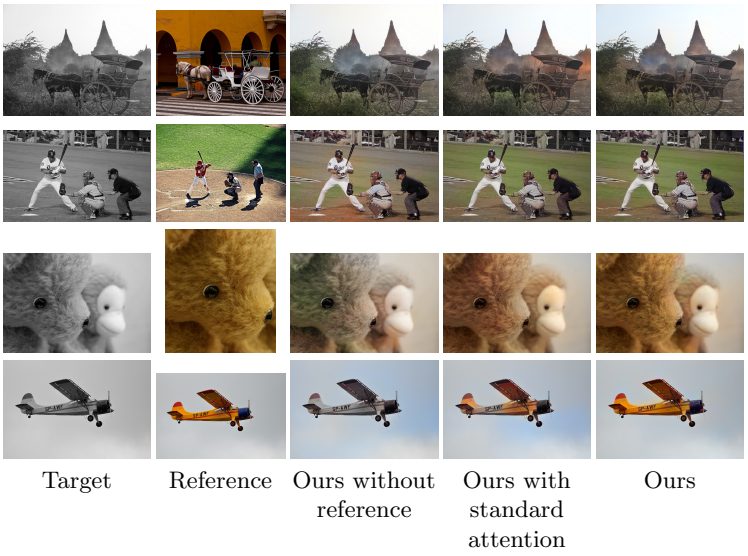


Fig. 5. Colorization results obtained using different variants of our colorization framework. In the first two lines, the chosen color reference is different from the target to be colorized (but semantically similar). In the last two lines, the reference correspond to the actual color version of the target image. These results allow to assess the ability of our method to transfer relevant colors from different types of reference images.

predicted colorization and the reference image. The goal of this third metric is to evaluate if the colors from the reference have been correctly transferred to the prediction. However, unless the groundtruth and reference share the same color distribution, these metrics are inherently contradictory (*i.e.*, good HIS would lead to bad SSIM/LPIPS). In this work, we do not necessarily want to fully transfer the colors from the reference image. Instead, we view the reference as colors hints that can be used by our network to predict a more plausible colorization. Therefore, we consider the HIS between the groundtruth target images and the reference images to be optimal in this sense (*i.e.*, it would be the score obtained with perfect predictions). On the COCO validation set, we obtain this reference, HIS=0.542, by picking the top 1 reference for each groundtruth. In the following, to better assess the quality of our results, we propose to report ΔHIS, that is the absolute difference between the average HIS obtained with our predictions and the reference HIS of 0.542.

Results and ablation study. The results displayed in Table 1 are the averages of the metrics calculated using the evaluation set of 5000 pairs of target/reference images from the COCO validation set. From this Table, we can observe that our full colorization framework achieves the best SSIM and LPIPS scores in comparison with the other variants of the model, suggesting that the use of reference images with super-attention blocks helps in getting better colorization

Table 1. Quantitative analysis of our model. SSIM and LPIPS metrics are calculated with respect to the target groundtruth image. ΔHIS is the absolute difference between the average HIS from the reference and the colorized prediction.

Method	**SSIM↑**	**LPIPS↓**	**ΔHIS↓**
Ours without reference	0.920	0.164	-
Ours with standard attention	0.921	0.172	0.081
Ours	**0.925**	**0.160**	**0.054**

results. We can also notice that the full model achieves a better ΔHIS than the standard attention variant. This suggests that, instead of forcing a global transfer of colors from the references, our model is capable of picking specific and plausible colors from the reference images to generate better colorization results. Note that the ΔHIS is not reported for the first line, as this variant does not use any reference image.

In addition to this quantitative evaluation, in Figure 5 we present a qualitative comparison of our method and its ablation variants. From the results of the first and second row, we can see that the method with standard attention proposes a more global transfer of colors, leading to the appearance of brighter colors related to the reference. However, this also causes unnatural colorization around the carriage in the first example (*i.e.*, green, orange) and on the hand of the baseball player in the second example. On the other hand, our method with super-attention block overcomes this issue and provides a more natural colorization, using specific colors from the reference. The last two lines show the effectiveness of our method when the reference is semantically identical to the target. Indeed, it shows that color characteristics from the reference image are passed with high fidelity, as compared to the variant without reference where it results in opaque colors.

In overall, our framework achieves a better visually pleasant colorization, and the transfer of color characteristics from the reference is done in specific areas of the target image, thanks to our super-attention blocks. When the color reference and the super-attention modules are not included in the framework, we observe a decrease of quality in the resulting colors (*i.e.*, averaging of colors and/or wrong colorization).

5.2 Comparison with other exemplar-based colorization methods

To evaluate the performance of our exemplar-based colorization framework, we compare our results quantitatively and qualitatively with three other state-of-the-art exemplar-based image colorization approaches [13, 15, 29]. In order to fairly compare these methods, we run their available codes for the three approaches using the same experimental protocol and the same evaluation set as in the previous section.

A quantitative evaluation is proposed to compare the three state-of-the-art methods and our framework. For comparing the methods, we again use SSIM, LPIPS, and ΔHIS. As shown in Table 2 our colorization framework preserves

Fig. 6. Comparison of our proposed method with different reference-based colorization methods: Deep Exemplar [13], Just Attention [15] and XCNET [29].

stronger perceptual structural information from the original target image with respect to all three state-of-the-art methods. LPIPS score measures the perceptual similarity between colorized results and target groundtruth. Our method surpasses [29], [15], [13] significantly. Finally, our method achieves a smaller ΔHIS with respect to all compared state-of-the-art methods.

Figure 6 shows colorization results from [29], [15], [13] and our approach. For the first two images, our proposal produces a more visually pleasant colorization and provides more natural colors than the other methods. In contrast, the results for the first two images of [15] and [29] shows a high amount of color bleeding, mainly on top of the baseball player and on the background, as their approach captured global color distribution from the reference image. For the fourth image, methods [15] and [29] failed to transfer the color information from the red jacket. Conversely, [13] and our approach did transfer the jacket's color. Next,

Table 2. Quantitative comparison with three state-of-the-art exemplar based-colorization methods. Ours correspond to our full model with references and super-attention blocks.

Method	**SSIM** ↑	**LPIPS** ↓	**ΔHIS**↓
XCNET [29]	0.867	0.270	0.139
Deep Exemplar [13]	0.894	0.241	0.127
Just Attention [15]	0.896	0.239	0.125
Ours	**0.925**	**0.160**	**0.054**

in the results for the fifth image, we observe unnatural colors as the green water for methods [13] and [29]. In contrast, our method and [15] achieves a natural colorization by transferring colors from the reference. Finally, for the results of the last image, [13] encourages the transfer of the colors from the reference image better than the other methods; however, the final colorization results seem unnatural. For the same image, the colorization results on method [29] and ours seem to be the right balance between color transfer and naturalness from the learned colorization model.

6 Conclusion

In this paper, we have proposed a novel end-to-end exemplar-based colorization framework. Our framework uses a multiscale super-attention mechanism applied on high-resolution features. This method learns to transfer color characteristics from a reference color image, while reducing the computational complexity compared to a standard attention block. In this way, we coupled into one network both semantic similarities from a reference and the learned automatic colorization process from a large dataset. Our method outperforms quantitatively state-of-the-art methods, and achieves qualitatively competitive colorization results in comparison with the state-of-the-art methods.

The transfer of colors which serve as a colorization guidance in the exemplar-based methods plays a key relevance on the final results. In our method, we introduce semantically-related colors characteristics from the reference by means of the super-attention block. This block let us find correspondences at different levels of the architecture, resulting in rich low-level and high-level information. Yet we suspect that concatenating the retrieved features of our attention mechanism do not completely enforce a strong guidance from reference color and future work should concentrate on a scheme for finer transfer. One solution we will consider is adding segmentation masks as in [3] inside our model.

Finally, another future line of research aims at coupling scribbles [40, 9, 5] as user hints with our reference-based colorization framework.

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

1. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision. (2016)
2. Deshpande, A., Lu, J., Yeh, M.C., Chong, M.J., Forsyth, D.: Learning diverse image colorization. In: Conference on Computer Vision and Pattern Recognition. (2017)
3. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: Conference on Computer Vision and Pattern Recognition. (2020)
4. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: Winter Conference on Applications of Computer Vision. (2020)
5. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Transactions on Graphics (2004)
6. Huang, Y.C., Tung, Y.S., Chen, J.C., Wang, S.W., Wu, J.L.: An adaptive edge detection based colorization algorithm and its applications. In: ACM International Conference on Multimedia. (2005)
7. Qu, Y., Wong, T.T., Heng, P.A.: Manga colorization. International Conference on Computer Graphics and Interactive Techniques (2006)
8. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics (2017)
9. Zhang, L., Li, C., Simo-Serra, E., Ji, Y., Wong, T.T., Liu, C.: User-guided line art flat filling with split filling mechanism. In: Conference on Computer Vision and Pattern Recognition. (2021)
10. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. ACM Transactions on Graphics (2002)
11. Bugeau, A., Ta, V.T., Papadakis, N.: Variational exemplar-based image colorization. IEEE Transactions on Image Processing (2014)
12. Chia, A.Y.S., Zhuo, S., Gupta, R.K., Tai, Y.W., Cho, S.Y., Tan, P., Lin, S.: Semantic colorization with internet images. ACM Transactions on Graphics (2011)
13. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics (2018)
14. 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
15. 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)
16. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: Conference on Computer Vision and Pattern Recognition. (2019)
17. Carrillo, H., Clément, M., Bugeau, A.: Non-local matching of superpixel-based deep features for color transfer. In: International Conference on Computer Vision Theory and Applications. (2022)
18. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: International Conference on Computer Vision. (2015)
19. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European Conference on Computer Vision. (2016)
20. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: International Conference on Learning Representations. (2021)

21. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
22. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (2009)
23. Xu, Z., Wang, T., Fang, F., Sheng, Y., Zhang, G.: Stylization-based architecture for fast deep exemplar colorization. In: Conference on Computer Vision and Pattern Recognition. (2020)
24. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: International Conference on Computer Vision. (2017)
25. Iizuka, S., Simo-Serra, E.: DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. In: ACM Transactions on Graphics. (2019)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017)
27. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Conference on Computer Vision and Pattern Recognition. (2018)
28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Conference on Computer Vision and Pattern Recognition. (2018)
29. 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)
30. Connolly, C., Fleiss, T.: A study of efficiency and accuracy in the transformation from rgb to cielab color space. IEEE Transactions on Image Processing (1997)
31. 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)
32. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: An open source differentiable computer vision library for PyTorch. In: Winter Conference on Applications of Computer Vision. (2020) 3674–3683
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition. (2018)
35. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2274–2282
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. (2014) 740–755
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. (2015)
38. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)

39. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition. (2017)
40. Heu, J., Hyun, D.Y., Kim, C.S., Lee, S.U.: Image and video colorization based on prioritized source propagation. In: International Conference on Image Processing. (2009)