Exemplar-based image colorization using object-guided attention

Hernan Carrillo, Michaël Clément, Aurélie Bugeau

Exemplar-based image colorization using object-guided attention

Hernan Carrillo, Michaël Clément and Aurélie Bugeau

***Abstract*—Exemplar-based image colorization is a challenging task that involves adding color to a grayscale image using a reference color image. The goal is to preserve the semantic content of the target image while also incorporating the color style of the reference image. However, results from previous methods are still unsatisfactory for real-world applications. One of the reasons is that they are inefficient at exploiting semantic color information, mainly when two or more objects are presented in the target or reference images. In this work, we propose a novel end-to-end deep learning framework for exemplar-based colorization that integrates user-provided object masks. We aim to guide the colorization on specific and meaningful objects rather than a full reference image. Our framework consists of an encoder-decoder generator architecture. The core module of the encoder is our proposed masked super-attention. This multiscale object-specific attention mechanism improves the ability to transfer color characteristics from the user's selected objects. In addition, we introduce a strategic method for selecting pertinent target/reference image pairs at the object-level. To comprehensively evaluate the effectiveness of our proposed approach, we conduct a complete evaluation of both full-level and object-level images. Finally, our framework achieves colorful and visually pleasant colorization and surpasses state-of-the-art methods on different quantitative metrics.**

***Index Terms*—Colorization, attention mechanism, segmentation.**

I. Introduction

COLORIZATION is assigning plausible colors to grayscale images, aiming to produce visually appealing images while avoiding unwanted artifacts or incorrect colors. This application is used in many fields, including restoration of legacy photos/videos, broadcasting, film post-production, and animation. However, current processes are often time-consuming and tedious, as they highly depend on manual intervention from the artist. Automating the colorization process can greatly improve workflow for artists, but it is challenging due to its inherent ambiguity. This is because several plausible colors can be assigned to the same gray pixel of an image, depending on various factors such as complex structures on the image. Therefore, there is no unique correct solution, and user input is often required to achieve satisfactory results.

Several colorization approaches have been proposed in the literature and can be classified into three types: scribble-based colorization, exemplar-based colorization, and automatic colorization methods without interaction. Scribble-based methods [1], [2] require users to manually assign colors to specific pixels based on the semantics and luminance of the patch, which is both time-consuming and difficult for those without artistic sensibility. Exemplar-based methods [3]–[6] use grayscale and similar reference images to output a colorized image based on chrominance information. However, the process can be time-consuming, and the quality of results depends heavily on the reference image chosen. Automatic colorization methods without interaction [7], [8] leverage a large-scale image database to train a neural network to predict colors for the target image automatically. However, the process is uncontrollable and does not allow for customization. Furthermore, satisfactory results can only be achieved if similar objects are included in the image database.

To address the weaknesses identified in the previous methods, we combine exemplar-based and learning approaches. We propose a guided attention mechanism using a segmentation map with an exemplar-based colorization method to better guide the colorization on specific, meaningful objects rather than a whole reference image. We suggest using segmentation masks to enhance the quality of image colorization in the following ways. First, the segmentation masks identify visually significant regions within the image. This allows the colorization framework to prioritize important objects rather than, for example, the background. Typically, objects of interest are more colorful, while backgrounds tend to be dominated by green and blue hues, such as sky, trees, and water. Our approach decreases the probability that the framework is biased toward the background colors. Secondly, segmentation masks help localize specific objects, highlighting semantically relevant regions with distinct boundaries. The previous is advantageous for colorization networks as it reduces color bleeding artifacts. Finally, adding segmentation as input is relatively easy for the user.

In this paper, we present a novel approach to object-specific exemplar-based colorization. Our approach builds upon the super-attention block introduced in [9], which leverages skip connections to transfer semantically related color characteristics from a reference image across various scales of a deep neural network. In this work, we extend the application of the super-attention block by incorporating an object-specific guidance mechanism and examining its impact on the encoder part of the architecture. Our main contributions are as follows:

- We develop a new end-to-end deep learning framework for exemplar-based colorization capable of incorporating object masks provided by the user.
- We integrate the super-attention block within the encoder of the network architecture instead of within the skip

H. Carrillo, M. Clément and A. Bugeau are with Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, Talence, France (emails: hernan.carrillo-lindado, michael.clement, aurelie.bugeau@labri.fr). A. Bugeau is also with Institut universitaire de France (IUF), France.

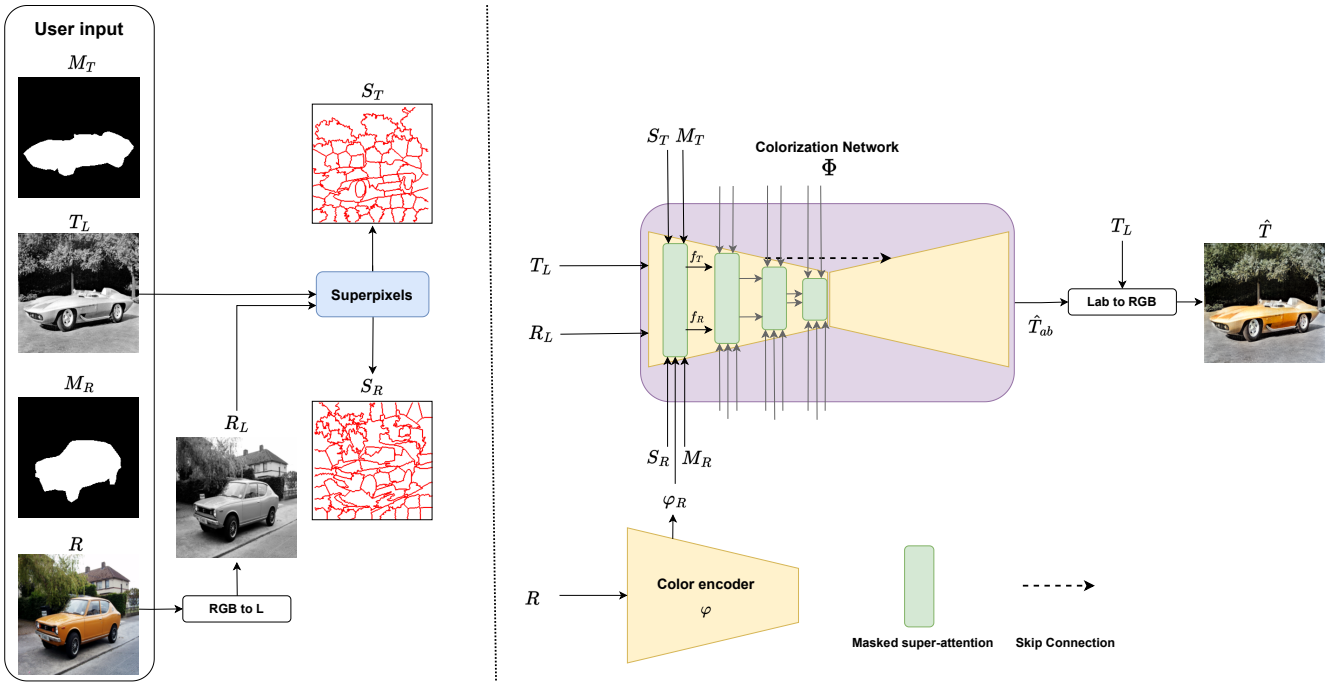

Fig. 1. Our proposal. The colorization framework includes two main parts that jointly handle the semantic correspondence and chromatic propagation between the input images. First, The color feature extractor φ extracts multi-level feature maps from a color reference image R. The main colorization network Φ that learns to map a luminance channel image T_L to its chrominance channels $\hat{T}_{ab}$ given color characteristics from image R. The colorization guidance is done by the masked super-attention modules. These attention layers receive feature maps from distinct levels f_T, f_R and their respective superpixels grids S and a segmentation mask M from the target and reference images.

connections as in [9].

- We propose a multiscale object-specific attention mechanism that utilizes masked superpixel features for reference-based colorization, called masked super-attention.
- We leverage a strategy for selecting relevant object pairs in target/reference images.
- We conduct a complete and comprehensive evaluation of our approach at both the full-image level and object-level images, comparing it to state-of-the-art methods.

II. Related works

Automatic colorization methods without interaction. These colorization techniques use large datasets to learn how to map each grayscale pixel in an input image to a specific color value. The earliest work in this area [10] proposes feeding to a neural network a grayscale image and predicting UV chrominance channels using a regression loss from the YUV luminance-chrominance color space. Another method proposed by Larsson et al. [7] used a VGG network architecture to predict a histogram of hue and chroma for each pixel, providing guidance for the final colorization result. Other deep learning approaches, such as Generative Adversarial Networks (GANs), have also been employed. For example, by Vitoria et al. [8] combined semantic and perceptual information through adversarial learning and high-level classification features to colorize grayscale images. In contrast, [11] used axial attention to predict the distribution of each pixel colors based on the previous pixel color distribution and the input grayscale image in an autoregressive framework. Overall, previous automatic colorization methods without interaction reduce colorization time but lack user-specific requirements when compared to purely manual colorization.

Exemplar-based colorization. These methods transfer color information from a reference image, which can either be provided by the user or proposed by a recommendation system, to the grayscale target image. Welsh et al. [12] matched luminance and texture information to transfer color information between images. Many extensions of this method have been proposed in the literature [3], [13], [14]. While providing satisfactory results, they all highly depend on the reference provided, and they cannot generate new colors if not present. To address this limitation, combining learning-based methods and exemplar-based approaches can overcome this issue. In recent years, He et al. [4] proposed a fully automatic image colorization system that used an end-to-end neural network to calculate the similarity between the reference image and the target image before color transfer. Their image retrieval algorithm also automatically suggests reference images by analyzing luminance and semantic features to reduce manual work further. After, Yi et al. [6] and Lu et al. [5] propose an end-to-end colorization network that exploits semantic correspondences between two images based on a gated attention mechanism. Improvement over [6] was suggested in Blanch et al. [15] where they introduced the axial attention mechanism for guiding the transfer of color attributes from the reference image to the target image. Recently Carrillo et al. [9] implemented the super-attention

blocks that enable the transfer of semantically related color characteristics from the reference image at various scales of a deep network. And finally, [16] design a framework that supports colorization in multiple modalities, both unconditional and conditional approaches such as stroke, exemplar, text, and their combinations, achieving state-of-the-art results. However, their technique cannot compensate for incorrect colors in less semantically significant areas or differentiate less semantic portions with identical local textures.

Segmentation. Image segmentation involves dividing an image into multiple regions or segments, each of which corresponds to a specific object. These approaches range from classical methods such as thresholding [17], [18], clustering [19], and edge detection [20] to deep learning-based methods. They have been successfully applied to image generation [21], image-to-image translation [22], [23], and semantic image synthesis [24]. Segmentation has been used for the cartoon colorization task, starting by Sykora et al. [25], which presents an exemplar-based colorization technique that uses unsupervised image segmentation joined with patch-based sampling to transfer colors from a reference colorized image. Extensions have been made to natural image colorization. Irony et al. [26] proposed a method for colorizing grayscale images using a segmented reference image. It considers the higher-level context of each pixel, resulting in colorization with a higher degree of spatial consistency through the mean-shift segmentation algorithm [27]. Later, Gupta et al. [28] used superpixels to improve the colorization process by speeding up the task and increasing spatial coherence, which was further improved in [29] by taking into account intensity, texture, and semantic features. Recent approaches such as Zhao et al. [30] coupled neural networks and pixel-level object semantics to guide colorization and mitigate the context confusion issues. Recently, Su et al. [31] proposed a method to improve image colorization with multiple objects, it uses an object detector to extract object instances, then employs a neural network to capture object-level features for later combining them with full-image features using a fusion module to predict accurate colors. Previous methods leverage fully automatic colorization methods with segmentation, meaning human intervention is unavailable for the colorization or segmentation tasks. The previous leads to issues where the automatic segmentation mask is inaccurate, or none of the objects were identified correctly, causing visible artifacts such as washed-out colors or bleeding across object boundaries.

III. Colorization framework

Our objective is to add feasible colors to a grayscale image using a color reference image. We aim to apply reference colors to semantically related content in the target image while creating a plausible colorization for regions or objects without such relationships. This goal poses two challenges. First, measuring the semantic connection between reference and target images is particularly challenging when the reference and the target images are partly semantically different. Secondly, even if we have good similarity metrics, selecting appropriate reference colors and effectively propagating them through the target image remains a difficult task.

We propose an end-to-end colorization network framework to address the previous two challenges. This framework includes two main parts that jointly handle the semantic correspondence and chromatic propagation between the input images. By doing so, we can break down the colorization task into two distinct subproblems instead of a highly complex one. Then, an external feature extractor is designed to extract color features from the reference color image. The main colorization network uses the original super-attention modules [32] in combination with our proposal on masked features at various levels of the encoder to guide the final colorization. The main colorization network uses a traditional encoder-decoder architecture similar to Unet [33], incorporating our proposed superpixel-level masked attention blocks. These blocks enable the transfer of color characteristics from the reference image to the main colorization network, allowing a more accurate and robust colorization. An overview of our proposal is depicted in Figure 1.

Our approach uses the CIELAB color space, taking a grayscale target image T and a color reference image R. Precisely, we extract the luminance component $T_L \in R^{H\times W\times 1}$ of the target image, which is represented by channel L from the CIELAB color space. The color reference image $R_{Lab} \in R^{H\times W\times 3}$ is also represented in the same color space. In this study, we opted for the luminance-chrominance CIELAB color space as it is more perceptually uniform than other color spaces [34]. Our framework predicts the target chrominance channels $\hat{T}_{ab}$, and concatenates it with the target luminance T_L for retrieving the complete image in the LAB color space $\hat{T}_{Lab}$. Next, it converts this result to the RGB color space $\hat{T}$.

For training, we use a two-phase sequential approach that involves first, training the framework without any segmentation (i.e., just pairs of full target-reference images), and then, pre-loading weights from the previous step and re-training with the masked super-attention block.

A. Colorization network

The main colorization network Φ aims to colorize a grayscale target image based on a reference image, transferring semantic-related color content where similarities exist and relying on the learned model when there is a lack of this information between the images. The users input to the colorization network the target image T_L and mask M_T and reference image R_L and mask M_R, which are processed to obtain deep learning feature maps f_T^ℓ and f_R^ℓ from the ℓ^{th} level of the network architecture. To learn to extract specific color features from the reference image R, we use a VGG19 [35] encoder pre-trained on ImageNet [36]. The color feature extractor φ (see Figure 1) retrieves multiscale feature maps φ_R^ℓ. The extracted features from the target and reference images are then fed to our proposed attention blocks (see Section III-B), where a correlation is computed between the masked features of the target and reference images. Next, the network relies on attention maps to transfer the content from the reference to the target image. The color features generated from the masked super-attention blocks are then introduced to the main

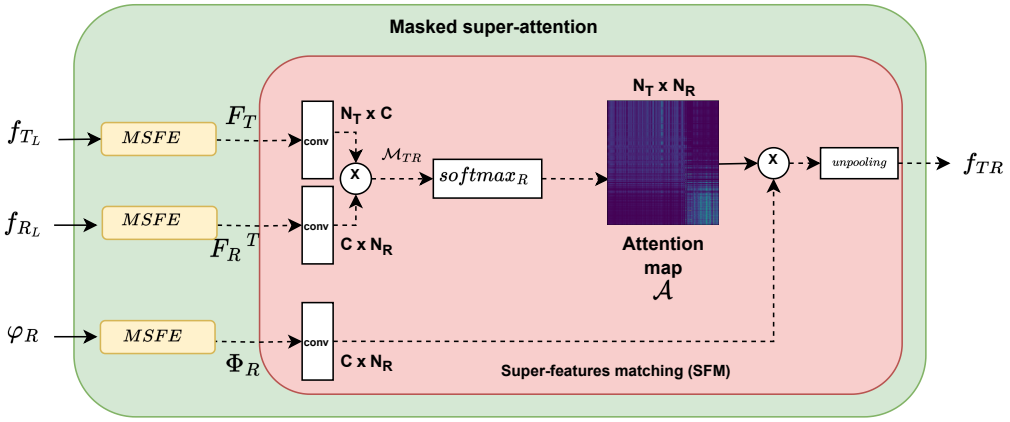

Fig. 2. Overview of our masked super-attention layer. Given a reference luminance feature map, denoted as f_R, a reference color feature map represented as φ_R, and a target luminance feature map called f_T with their respective reference mask M_R and target mask M_T. Through a robust matching process between high-resolution encoded feature maps, this layer learns an attention map at the level of superpixels

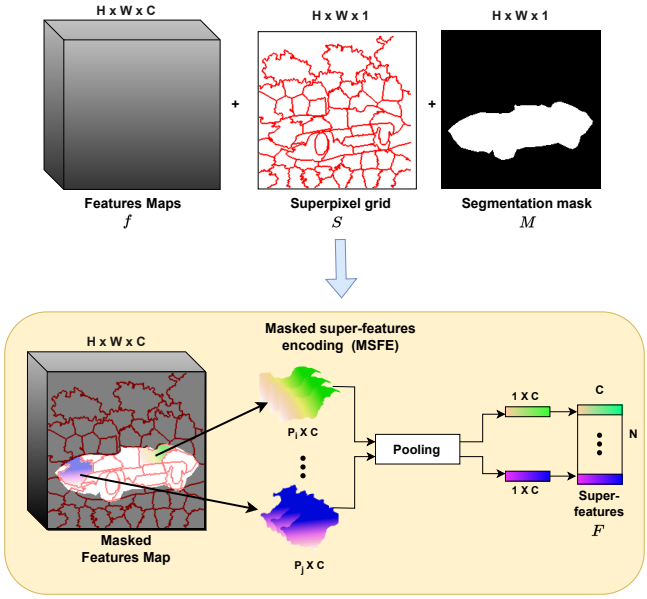

Fig. 3. Diagram of our masked super-features encoding proposal (MSFE). This encoding block takes as input a feature map of size $H \times W \times C$, where each superpixel that belongs inside the mask M_s is extracted and encoded in vectors of size $C \times P_i$ pixels. Afterward, the vectors are pooled channel-wise and, finally, stacked in the super-features matrix F with size $C \times N$ number of superpixels.

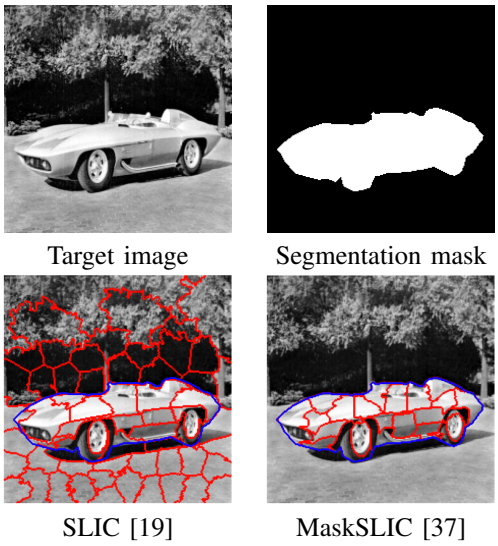

Fig. 4. Example of superpixel algorithm on a gray-scale image. The superpixel grids are generated using the SLIC and the MaskSLIC algorithm, which is a region-based image segmentation algorithm. The SLIC algorithm divides the full image into a set of small, non-overlapping regions. On the other hand, the MaskSLIC divides only the regions inside a segmentation mask.

colorization network encoder. Finally, the decoder predicts the two chrominance channels $\hat{T}_{ab}$.

B. Masked Super-attention

In addition to colorizing grayscale images from full reference images, our colorization framework can also be used to colorize specific objects within an image. This is done using a segmentation mask M_s to identify the object of interest. Once the object of interest has been specified, a super-attention mechanism [9] is applied to focus on the most essential features in that region. Mainly, this masked super-attention mechanism, learns to find similar object-to-object characteristics between a reference and a target image.

Our masked super-attention block is a novel way to add controlled color information from a reference image to the main colorization network Φ. This is achieved by comparing the features of the target and reference images at multiple levels. The masked super-attention block has two parts: the masked super-features encoding layer (MSFE) and the super-features matching layer (SFM). The MSFE creates a compact representation of the high-resolution deep features using superpixels constrained to a segmentation mask. The SFM layer then matches these compact representations to find the most similar features between the target and reference object images. In the MSFE, we use features from all four levels of the architecture, as these features provide a broad range of high-level and low-level characteristics that are well-suited for content and style applications [32]. Figure 2 shows the diagram of our masked 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. Figure 3 exemplifies the encoding process of the MSFE block where superpixels are used to represent the target and reference images into smaller regions. Each of these smaller regions inside the mask M_s contains N_T and N_R superpixels, respectively, with P_i pixels each, where i is the superpixel index. We apply a channel-wise masked pooling operation on the chosen superpixels to perform the encoding. This results in super-features F with dimensions $C \times N$, where N is significantly smaller than $H \times W$. Our masked-super-attention block is inspired by the super-attention module [9], Which guides the colorization, considering the global context of a full reference image. However, our masked super-attention module focuses on specific object-to-object feature maps, helping the network guide the colorization to a particular structure the user decides. This guidance is done by multi-level feature correlations between the target F_T and reference F_R masked super-features by computing the attention map at layer ℓ as:

$$\mathcal{A}^{\ell} = \mathrm{softmax}(\mathcal{M}^{\ell}{}_{TR}/\tau). \tag{1}$$

The $\mathrm{softmax}$ 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:

$$\mathcal{M}^{\ell}_{TR}(i,j) = \frac{\left(F^{\ell}_T(i) - \mu^{\ell}_T\right) \cdot \left(F^{\ell}_R(j) - \mu^{\ell}_R\right)}{\left\|F^{\ell}_T(i) - \mu^{\ell}_T\right\|_2 \left\|F^{\ell}_R(j) - \mu^{\ell}_R\right\|_2} \tag{2}$$

where μ_T, μ_R are the mean of each super-feature and i, j are the current superpixels that belongs to the segmentation mask M from the target and reference respectively.

Figure 5 illustrates attention maps $\mathcal{A}_\ell$ at all four levels ℓ of the architecture encoder. These attention maps depict the similarity between specific characteristics of an object in the target image and another in the reference image. This shows that the learned masked attention map can find relevant superpixels in the reference feature maps with similar characteristics to the target superpixel.

Masked super-attention vs. original super-attention: Masked super-attention can be seen as a generalization of the super-attention from [9]. To retrieve this global attention, we can simply apply the block without an object mask.

C. Loss function

In the context of automatic colorization, a common approach involves predicting the colors ($\hat{T}_{ab}$) by reconstructing them from the ground truth image (T_{ab}). However, this approach can be insufficient in exemplar-based colorization because the main idea is to use the colors from the reference image R to fully or partially colorize the grayscale image T_L. Therefore, the predicted colors $\hat{T}_{ab}$ need to incorporate the color information from a reference image, that is $\hat{T}_{ab} = \phi(T_L|R)$. The objective is to ensure a precise and consistent transfer of color characteristics from the reference to the target.

Our method proposes a combined strategy using two loss terms: L1 smooth and LPIPS [38]. The L1 smooth helps to ensure that the predicted colors are smooth and gradual as well as helping to address the multi-modal ambiguity problem in colorization [4], [5], [39]. This problem refers to the fact that there can be multiple possible colorizations for a given grayscale image. And the LPIPS loss encourages the network to generate perceptually plausible images. These terms are essential in reconstructing the final image accurately. The joint total loss used in the training phase is then:

$$L_{total} = \lambda_1 L_{1_{smooth}} + \lambda_2 L_{LPIPS} \tag{3}$$

In this equation, λ_1 and λ_2 represent the predetermined weights for each individual loss component.

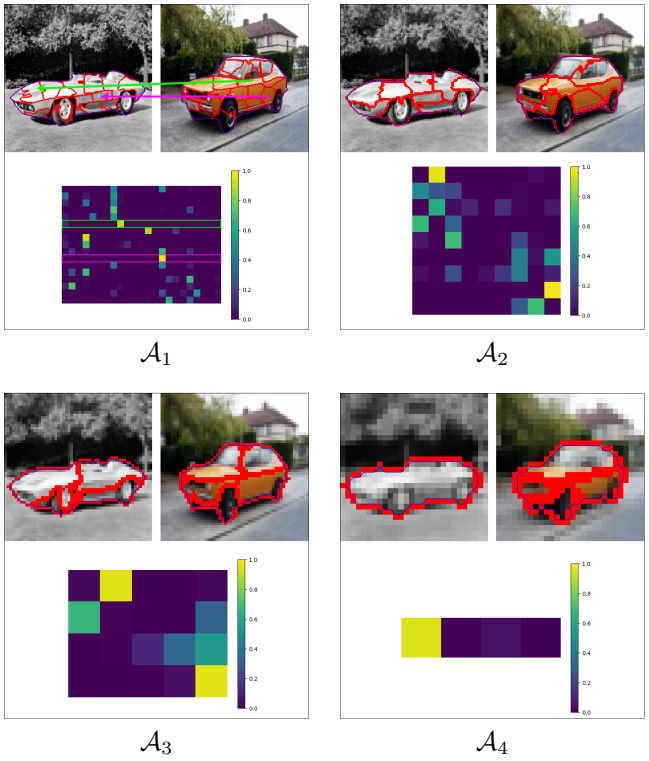

Fig. 5. Example of masked super-attention mechanism guidance. Each bounding box presents a target and a reference image with a superpose grid of superpixel using MaskSLIC [37] and its respective attention map $\mathcal{A}_\ell$. In the first bounding box, we can see two arrows that point to which superpixels from the reference image correspond to its similarities in the target image, in addition to its similarity score in the attention map where rows refer to the target superpixels and columns to reference superpixels.

IV. Dataset and Reference Selection

Our framework was trained on COCO dataset [40]. This dataset exhibits images with complex scene structures and diverse object classes. Additionally, it provides object segmentation information, which we later use in our strategy of pairing object-specific target and reference images. For our training, we use two different splits of the dataset. First, a full image-level split, which consists of 100k images for training and 5k images for validation. Second, an object-level split consists of 25k object images and their segmentation mask for training and 1k images for testing. We resized the images to a standardized size of 224×224 pixels during the training process.

Another essential aspect of the training strategy of exemplar-based methods is the identification of an appropriate semantic reference for the target image. For searching pairs between target and reference images in the full image-level split, we use Carrillo et al. [9] approach to match several reference images with each target one. In [9], five reference images are ranked regarding semantic similarity using a pre-trained VGG-19 and a L2 distance. However, we found that after top-3, images do not convey significant semantic relevance with respect to the target image. We therefore keep only top-3 target images and complete this set with two additional pseudo-synthetic reference images. These two images are obtained by appearance and spatial transformation on the current target image using the Thin Plate Spline (TPS) [41], [42], a non-linear spatial transformation operator.

For the second split, we search pairs at the object level between target and reference images. First, we do a local search in each class to find meaningful objects whose size is larger than a percentage of the actual image. This is because image features are downsampled at each of the four levels of the architecture, and then small objects will not introduce meaningful characteristics to the attention calculation. For this, we set the percentage empirically to 30%. Therefore, doing superpixels on a smaller threshold (smaller objects) would not represent the actual object well in the architecture lower levels. Knowing the object class, we randomly sampled three reference images from this class, and additionally, we applied TPS transformation on the target object to finish a top-5 reference object images.

Finally, during training and for both splits, target-reference pairs of images were sampled using a uniform distribution with a weight of 0.25. This was accomplished by randomly selecting either the three semantically closest reference images or the two synthetic references.

V. Experiments

A. Implementation details

In this paper, we implement an Unet-like generator architecture for our main colorization network Φ where, for each of the levels in the encoder, we introduce our masked super-attention block. Both the main model Φ and color encoder φ are jointly trained. We employed the Adam optimizer to optimize both networks with a learning rate of $10^{-}5$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. The training was conducted in two phases. First was a full image-level phase where we used the original super-attention without providing any segmentation mask for 40 epochs. The second phase uses our masked super-attention to introduce object-specific characteristics to the network for 10 epochs. Throughout both phases, a batch size of 8 was used. In order to balance the losses, we set the coefficients for each loss function as follows: $\lambda_1 = 2$ and $\lambda_2 = 0.15$. The training process was performed on a single GPU, specifically, the NVIDIA RTX 2080 Ti, and PyTorch 1.30.0 was used as the programming framework.

We took inspiration from the original super-attention module [9] when developing our masked super-attention approach. The classic super-attention module employs the SLIC algorithm [19] to calculate superpixel segmentation on the full image level. However, our masked super-attention leverages an object-specific segmentation mask for the calculation. We then employ MaskSLIC [37] to compute both the target and reference masked superpixel grid. Figure 4 shows an example of the difference between SLIC and MaskSLIC. Note that any superpixel segmentation method could be used.

B. Metrics details

To quantitatively evaluate the results, we used five metrics. Three of the metrics compare the results with the ground-truth color image, while the two other metrics compare the prediction of colors with respect to the reference color image.

Structural similarity (SSIM) [43]. This metric analyzes the ability of the model to reconstruct the original image color and texture.

Learned perceptual image patch similarity (LPIPS) [38]. The goal is to measure the perceptual similarity between the predicted and ground-truth images.

Learned perceptual image patch similarity w.r.t reference (LPIPS$_R$) [44]. This metric, also known as contextual loss, measures the perceptual similarity between non-aligned images, in this case, the predicted and the reference images.

Fréchet Inception Distance (FID) [45]. This metric measures the similarity between the distribution of features extracted from a set of predicted images and the distribution of features extracted from a set of ground-truth images.

Fréchet Inception Distance infinity (FID$_\infty$) [46]. Chong et al. [46] show that the bias in the FID metrics depends on the particular model being evaluated, so a specific model may get a better score than another simply because the bias term is smaller. The number of samples heavily influences this effect. More precisely, FID is linear to $1/N$, where N is the number of generated samples. In [46], they propose a method for extrapolating the FID scores to obtain an effectively bias-free estimate of scores computed with an infinite number of samples called FID infinity. Their method involves randomly sampling images from a generated dataset of size N in k intervals, each containing N_{itv} images. They calculate a FID score for each of these intervals and perform linear regression on these k data points to determine the bias-free FID metric, denoted as $FID_{N_{itv}}$. This metric is particularly useful for comparing our test set at the object level, especially when our current split comprises only $N = 1k$ images and is susceptible to this bias issue. In detail, we let $k = 15$ as the default value in their metric. In addition, we choose to calculate FID_{300} and FID_{600} as they are sufficient to know the true tendency of the metric. Finally, To ensure robustness and reliability in our results, since the metric relies on randomly sampled intervals from the test set, we evaluate FID_{300} and FID_{600} ten times. The final results are obtained by calculating the average and standard deviation across these ten evaluations.

Histogram intersection similarity (HIS) [47]. This metric evaluates the similarity of the global color distributions of the two images. This metric becomes contradictory if the ground-truth and reference have different color distributions. In other words, a good histogram intersection similarity (HIS) score between the predicted and reference image would lead to

Fig. 6. Results of using the original super-attention [9] on skip-connection and our implementation with super-attention module in the encoder. Our proposal is more effective at transferring color characteristics between the reference and the target.

TABLE I
QUANTITATIVE ANALYSIS BETWEEN SUPER-ATTENTION IN THE SKIP CONNECTION [9] AND SUPER ATTENTION IN THE ENCODER AT FULL IMAGE LEVEL.

	Full image split				
	$\hat{T}$ - GT target			$\hat{T}$ - Reference	
Method	**SSIM**↑	**LPIPS**↓	**FID**↓	**LPIPS$_R$**↓	**Δ HIS**↓
Super-attention [9]	**0.92**	**0.14**	11.24	2.14	0.23
Ours w/o seg	0.91	**0.14**	**9.20**	**2.01**	**0.18**

poor scores in terms of structural similarity (SSIM), learned perceptual image patch similarity (LPIPS), and the Fréchet Inception Distance (FID). We consider the reference image as color guidance to our network in generating a more plausible and realistic colorization. Thus, we regard the HIS score between the ground-truth target images and the reference images as the optimal score in this context, representing what would be achieved with perfect predictions. Then, the equation used for calculating ΔHIS:

$$\Delta HIS = \mid \mathcal{T}_{\text{hist}}\left(T_H, R_H\right) - \mathcal{T}_{\text{hist}}\left(\hat{T}_H, R_H\right) \mid, \tag{4}$$

where $\mathcal{T}_{\text{hist}}$ refers to histogram intersection metric [48], and T_H, $\hat{T}_H$, and R_H represent the chrominance histogram calculated in the ab space for the target ground-truth, predicted image, and reference image, respectively.

The metrics SSIM, LPIPS, and FID evaluate the quality of the output colorization compared to the ground-truth. The other two metrics, HIS and LPIPS$_R$, are computed between the predicted and reference colors images. The final results for these five metrics are the averages calculated using either the full-image evaluation set from the COCO validation set or the subset of object-level within the same dataset. We strongly believe that these five metrics provide a comprehensive measurement of the quality of the output colorization at the full-level image and the object-level image.

C. Analysis on super-attention at the encoder

We conduct qualitative and quantitative evaluations to inspect the super-attention effectiveness in the encoder rather than in the skip-connections as presented in [9]. For our proposal with the original super-attention in the encoder, we named it "our proposal without segmentation", as we used the super-attention and not the masked super-attention. And the approach using the super-attention in the skip-connection is called super-attention [9]. In terms of quantitative results, Table I shows that our proposal, without segmentation, retrieves a better FID score than using the original super-attention in the skip connection. However, in the other two metrics that compare the target with respect to the ground-truth (LPIPS and SSIM), our super-attention in the encoder got worse results than [9]; this happened because the super-attention in the encoder uses sequentially the outputs of high and low deep-features features of the encoder which push the transfer of features from the reference to the target image. For the metrics comparing the results with respect to the reference image, our proposal achieves better results in LPIPS$_R$ and ΔHIS, which means that results using super-attention in the encoder present more semantics characteristics from the reference image, as well as better similarity on the global tone from the reference image than the super-attention approach [9].

Figure 6 compares these two versions qualitatively. In the case of the super-attention in the skip connection [9], it produces washed-out and opaque colors, which means that colors expected from the reference images are not being truly

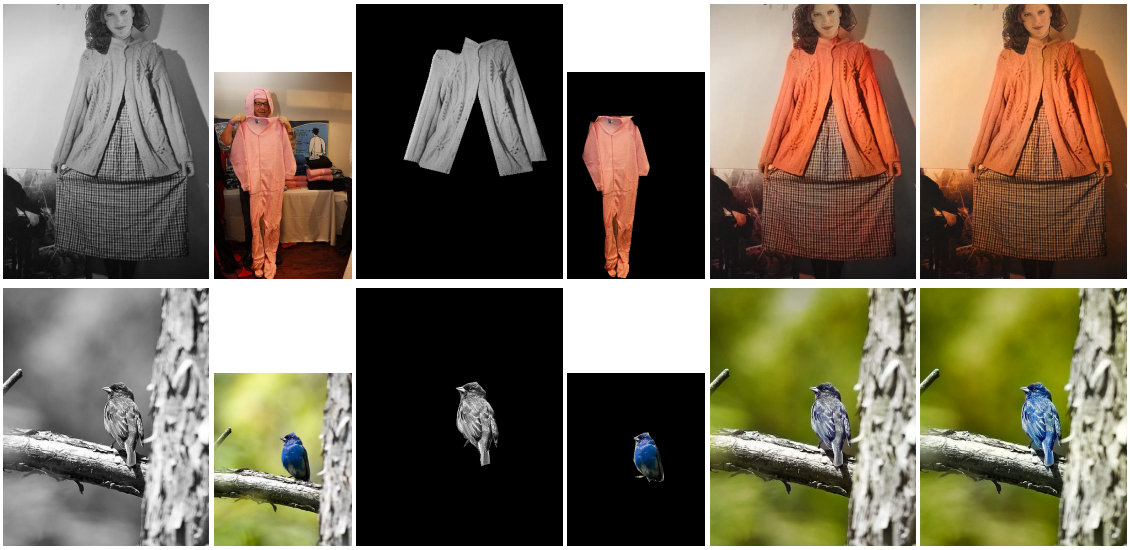

Fig. 7. Comparison of the proposal with and without fine-tuning at object-level. The first four columns show the inputs to our framework, which the user provides. The fifth column shows the results of our framework without object-level learning. This means we only use our masked super-attention module without additional training on object-specific images. The last column shows the results of our full method, which includes both the masked super-attention module and fine-tuning on object-level images

TABLE II
QUANTITATIVE RESULTS ON FINE-TUNING AT OBJECT-LEVEL. OURS W/O FINE-TUNE. CORRESPONDS TO OUR PROPOSAL WITHOUT FINE-TUNING ON OBJECT IMAGES BUT USING MASKED SUPER-ATTENTION. OURS CORRESPONDS TO OUR FULL MODEL WITH MASKED SUPER-ATTENTION MODULE AND AFTER FINE-TUNING WITH OBJECT-RELATED IMAGES.

	Object level split					
	$\hat{T}$ - GT target				$\hat{T}$ - Reference	
Method	**LPIPS**↓	**FID**↓	**FID**$_{\infty\_300}$ ↓	**FID**$_{\infty\_600}$ ↓	**LPIPS**$_R$↓	**Δ HIS**↓
Ours w/o fine-tune.	0.17	32.80	10.87 ± 0.29	4.02 ± 0.33	**1.92**	0.18
Ours	**0.15**	**30.45**	**6.80 ± 0.28**	**2.67 ± 0.32**	2.04	**0.17**

transferred but average to the target image. On the other hand, placing the super-attention in the encoder forces vivid and visually pleasant transfer of color characteristics between reference and target images.

D. Fine tuning at object-level

This subsection evaluates the benefits of training our proposal with object-level images. We conducted experiments on two versions of our model using the same architecture and pre-trained at the full-image level (first training phase). The first uses the masked super-attention module without further fine-tuning or training on object-specific images. The second is our full proposal, including the masked super-attention and training phases with full and object-level images. We compared the two models quantitatively and qualitatively.

As shown in Table II, our full proposal achieved better LPIPS and FID results, meaning that stronger perceptual similarities are retained between the colorized results and the target ground truth. In metrics concerning the references, such as LPIPS$_R$, the approach without fine-tuning achieves better results showing that the masked super-attention can transfer meaningful semantic characteristics from the reference object to target object. Finally, our full method achieves smaller ΔHIS, meaning that the global tone of the reference object is also well transferred.

In addition to the quantitative evaluation, Figure 7 provides a qualitative comparison of the two approaches. In the first row, the goal is to colorize the woman's sweater in the target image with the color characteristics of the pink pajamas in the reference image while ensuring that the rest of the image is properly colorized. The result achieved by our proposal without leaning at the object level shows a correct transfer of color within the sweater; however, the face shows a not visually pleasant grayish color with slight color bleeding on the wall. Our full proposal overcomes previous aspects; however, we got a yellowish tonality in the back wall. For the last image, the goal is to transfer the bird's blue in the reference images to the bird in the image in the grayscale. The result of our proposal, without learning at the object level, presents a fairly good transfer of colors from the global tones within the object, as we can see that a mix of dark blue is transferred to the target image. However, our full proposal shows a more colorful colorization with a brighter blue, nearly as the one in the reference mask image.

The results showed that our full proposal achieved the best performance in terms of both quantitative and qualitative metrics with respect to our proposal without fine-tuning object-specific images. This suggests that by fine-tuning the masked super-attention results gains more spatial consistency in colors between the object and the image background, resulting in

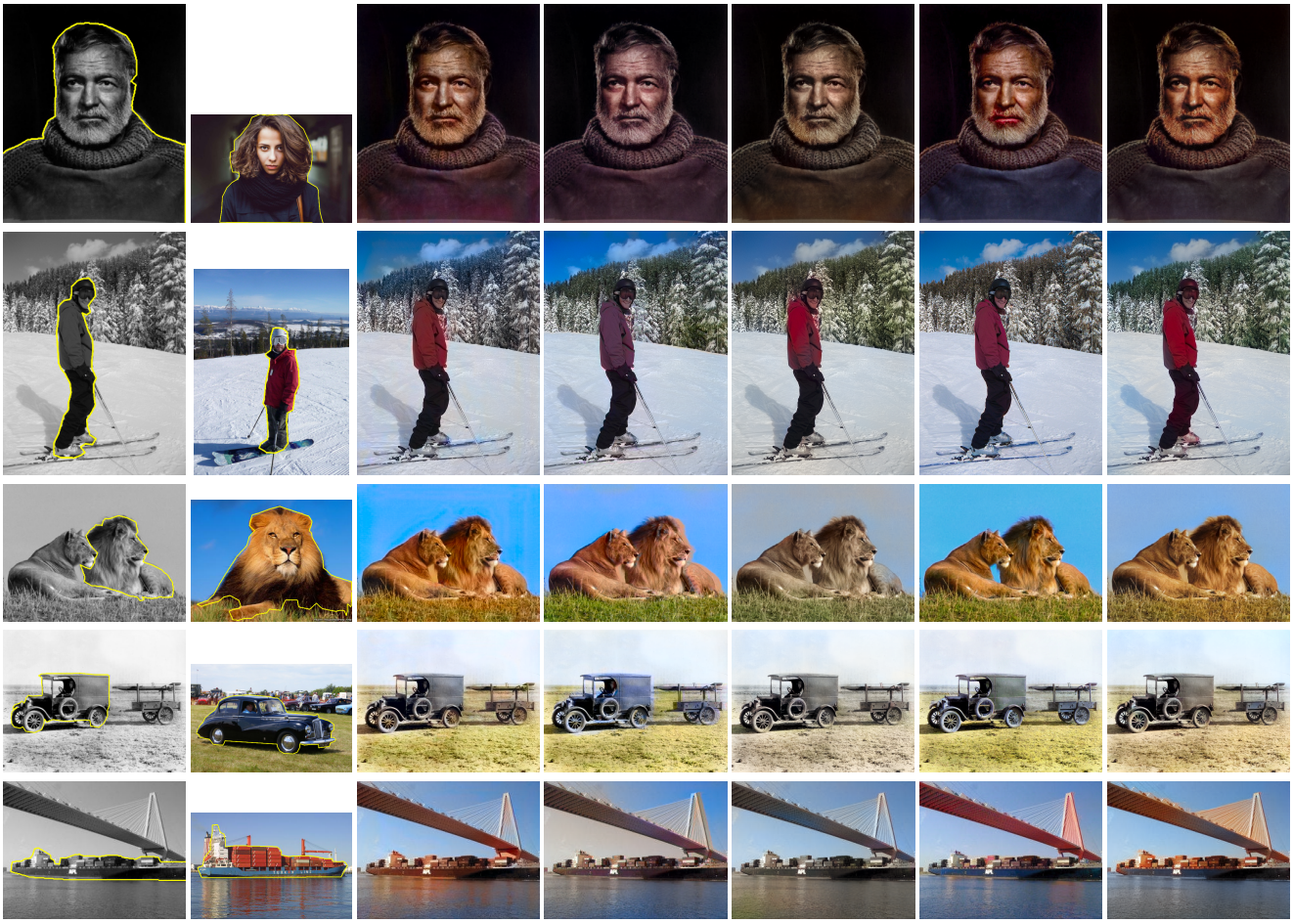

Fig. 8. Comparison of our proposed method with different reference-based colorization methods. Target and reference images are the only input to the SOTA methods: Gray2colorNet [5], Just Attention [6], Super-attentionT [9], Unicolor [16]. For our full method, in addition to the target and reference images, users can also provide an object segmentation mask. The yellow contours in both columns of images indicate the object segmentation mask.

TABLE III
COMPARATIVE EVALUATION AT FULL IMAGE LEVEL.

	Full image split				
	$\hat{T}$ - GT target			$\hat{T}$ - Reference	
Method	**SSIM↑**	**LPIPS↓**	**FID↓**	**LPIPS$_R$↓**	**Δ HIS↓**
Gray2ColorNet [5]	0.89	0.24	22.04	2.17	0.34
Just attention [6]	0.90	0.23	16.80	2.23	0.26
Super-attention [9]	**0.92**	**0.14**	11.24	2.14	0.23
Unicolor [16]	0.88	0.22	9.70	2.11	0.21
Ours w/o seg	0.91	**0.14**	**9.20**	**2.01**	0.18

more naturalness in the image.

E. Comparison with state-of-the-art

We evaluate the performance of our framework, and for that, we compared our results quantitatively and qualitatively with four other state-of-the-art exemplar-based image colorization approaches Gray2colorNet [5], Just attention [6], Super-attention [9] and Unicolor [16]. To ensure a fair comparison, we ran the available codes for the four approaches using the same experimental protocol and the same evaluation set for all the methods.

Color transfer at full image level. As shown in Table III our proposal without object-specific segmentation obtains the best LPIPS, FID, and LPIPS$_R$. The latter means that our framework retains strong perceptual information not only from the original target image but as well as from the reference color image. For the SSIM metric, ours achieve competitive results with respect to super-attention [9] and surpasses all four other methods. Finally, our method achieves a smaller ΔHIS with respect to all compared state-of-the-art methods. This indicates that rather than forcing to transfer all colors from the reference images, our model has the ability to selectively choose specific colors from the references. As a result, it can generate natural

TABLE IV
COMPARATIVE EVALUATION AT OBJECT-LEVEL.

	Object level split					
	$\hat{T}$ - GT target				$\hat{T}$ - Reference	
Method ↑	**LPIPS**↓	**FID**↓	**FID**$_{\infty\_300}$ ↓	**FID**$_{\infty\_600}$ ↓	**LPIPS**$_R$↓	**Δ HIS**↓
Gray2ColorNet [5]	0.18	35.85	19.04 ± 0.31	18.40 ± 0.26	2.88	0.21
Just attention [6]	0.19	39.65	15.34 ± 0.23	14.76 ± 0.45	2.58	0.18
Super-attention [9]	0.17	32.61	8.22 ± 0.28	4.28 ± 0.38	2.14	0.20
Unicolor [16]	0.23	32.40	7.64 ± 0.15	3.57 ± 0.32	**1.87**	**0.16**
Ours	**0.15**	**30.45**	**6.80 ± 0.28**	**2.67 ± 0.32**	2.04	**0.16**

colorization results.

Color transfer over object. Table IV shows the comparison of the four evaluation metrics for each of the different methods. It is important to note that these metrics were calculated on object-specific images. Instead of doing calculations on the entire predicted image, we cropped the specific object for which color transfer was desired and measured each of the metrics on this object image. As for the metrics used, we employed LPIPS, FID, LPIPS$_R$, and HIS as mentioned before. In addition, we utilized both the classic FID method and the unbiased FID$_\infty$ for this purpose. In terms of metrics our full proposal retrieves more perceptually semantically characteristics at the object level than other fourth methods. From all the variants of FID calculations, our method manages to well retain similar characteristics distribution from the ground-truth images. LPIPS$_R$ metric measures how well the model transfers perceptual characteristics from the reference image to the target one. In this case, our method surpasses all state-of-the-art approaches. Finally, in terms of ΔHIS our full method successfully achieves smaller results with these metrics in comparison with all state-of-the-art methods as well as our method without segmentation part. This demonstrates the capability to transfer color characteristics from an object reference to specific regions on the target object.

Figure 8 shows results of image colorization from five different methods: [5], [6], [9], [16], and our full method. For the first two images, our proposal produces more visually pleasant and natural colorization than the other four methods. The results from [5], [6], and [9] fail to transfer the blue from the woman's sweater. Additionally, [6] and [9] produce washed-out colors, making the head and clothes having the same color in the first image. For the fourth image, [5] and [16] show a high amount of color bleeding, mainly on the car. This color bleeding also appears on the small trailer and in its background. In contrast, our proposal and [5] shows the right balance between transferring colors between objects and colorizing the background. Finally, for the last image, [6], [9], and [16] struggle to find the correct colors to transfer. The first method transfers a red color, the second method transfers a blue color, and the third method transfers an average of colors. Our proposal and [16] correctly transfer vivid colors from the reference image, especially from the ship in the reference image. Our proposal and [16] shows the right balance between a colorful colorization and the naturalness from the learned colorization model.

VI. CONCLUSION

In this paper, we have introduced a novel end-to-end deep learning framework for exemplar-based colorization, which stands out for its ability to incorporate user-provided object masks. Our proposed masked super-attention provides visually pleasant and spatially consistent results with vivid colors. We performed a comprehensive evaluation, which includes both full-image and object-level metrics, outperforming quantitatively four state-of-the-art methods. However, we believe there is room for improvement, particularly regarding the layers where the masked super-attention module is applied. A possible solution involves calculating the attention maps on upsampled low-level features and re-weighing them with all attention maps. Another future line of work could be to study the clipping problem arising from passing from Lab to RGB spaces when the combination of predicted Lab values falls outside the conversion range. One solution could be using an oblique projection [49] in the final part of our model.

ACKNOWLEDGMENTS

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] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Transactions on Graphics*, 06 2004.
[2] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *International Conference on Computer Graphics and Interactive Techniques*, 2006.
[3] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Transactions on Image Processing*, 2014.
[4] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics*, 2018.
[5] P. Lu, J. Yu, X. Peng, Z. Zhao, and X. Wang, "Gray2colornet: Transfer more colors from reference image," in *ACM International Conference. on Multimedia*, 2020.
[6] W. Yin, P. Lu, Z. Zhao, and X. Peng, "Yes, "attention is all you need", for exemplar based colorization," in *ACM International Conference. on Multimedia*, 2021.
[7] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*, 2016.
[8] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *Winter Conference on Applications of Computer Vision*, 2020.
[9] H. Carrillo, M. Clément, and A. Bugeau, "Super-attention for exemplar-based image colorization," in *Asian Conference on Computer Vision*, 2022.
[10] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *International Conference on Computer Vision*, 2015.
[11] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," in *International Conference on Learning Representations*, 2021.
[12] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Transactions on Graphics*, 2002.

[13] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," *ACM Transactions on Graphics*, 2008.
[14] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," *ACM Transactions on Graphics*, 2011.
[15] M. G. Blanch, I. Khalifeh, A. Smeaton, N. E. Connor, and M. Mrak, "Attention-based stylisation for exemplar image colourisation," in *IEEE International Workshop on Multimedia Signal Processing*, 2021.
[16] Z. Huang, N. Zhao, and J. Liao, "Unicolor: A unified framework for multi-modal colorization with transformer," *ACM Transactions on Graphics*, vol. 41, 2022.
[17] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, 2004.
[18] A. Anjos and H. Shahbazkia, "Bi-level image thresholding - a fast method," in *International Conference on Bio-inspired Systems and Signal Processing*, 2008.
[19] 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*, 2012.
[20] T. Lindeberg and M.-X. Li, "Segmentation and classification of edges using minimum description length approximation and complementary junction cues," *Computer Vision and Image Understanding*, 1997.
[21] K. K. Singh, U. Ojha, and Y. J. Lee, "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *Conference on Computer Vision and Pattern Recognition*, 2019.
[22] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Conference on Computer Vision and Pattern Recognition*, 2019.
[23] S. Ma, J. Fu, C. W. Chen, and T. Mei, "Da-gan: Instance-level image translation by deep attention generative adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2018.
[24] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Conference on Computer Vision and Pattern Recognition*, 2018.
[25] D. Sýkora, J. Buriánek, and J. Žára, "Unsupervised colorization of black-and-white cartoons," in *International symposium on Non-photorealistic animation and rendering*, 2004.
[26] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Eurographics conference on Rendering Techniques*, 2005.
[27] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 2002.
[28] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *ACM International Conference. on Multimedia*, 2012.
[29] B. Li, F. Zhao, Z. Su, X. Liang, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Transactions on Image Processing*, 2017.
[30] J. Zhao, L. Liu, C. G. M. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," 2018.
[31] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Conference on Computer Vision and Pattern Recognition*, 2020.
[32] 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 Applications*, 2022.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015.
[34] C. Connolly and T. Fleiss, "A study of efficiency and accuracy in the transformation from rgb to cielab color space," *IEEE Transactions on Image Processing*, 1997.
[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition*, 2009.
[37] B. Irving, "maskslic: regional superpixel generation with application to local pathology characterisation in medical images," *arXiv preprint arXiv:1606.09518*, 2016.
[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Conference on Computer Vision and Pattern Recognition*, 2018.
[39] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *Conference on Computer Vision and Pattern Recognition*, 2019.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.
[41] F. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.
[42] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Conference on Computer Vision and Pattern Recognition*, 2020.
[43] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.
[44] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *European Conference on Computer Vision*, 2018.
[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, 2017.
[46] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Conference on Computer Vision and Pattern Recognition*, 2020.
[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2017.
[48] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.
[49] P. Fabien, A. Jean-François, B. Aurélie, and T. Vinh-Thong, "Luminance-hue specification in the RGB space," in *Scale Space and Variational Methods in Computer Vision*, 2015.

BIOGRAPHY SECTION

Hernan Carrillo received the M.Sc degree in signal and image processing from the Ecole centrale de Nantes, France in 2020. Since, he is pursuing his Ph.D at the University of Bordeaux and attached to the Laboratoire Bordelais de Recherche en Informatique to the image and signal processing department. His research areas includes computer vision, image processing, deep-learning.

Michaël Clément is associate professor of computer science at Bordeaux INP and researcher at LaBRI since 2018. He received the PhD degree in computer science from Université Paris Descartes in 2017. His research interests include pattern recognition, computer vision and image analysis, with a focus on structural representations of images for different applications (image colorization, medical imaging, mechanical structures, etc.).

Aurélie Bugeau is Professor in Computer Science at the University of Bordeaux (FR) and a researcher at LaBRI. In 2022, she has been distinguished as a junior member of the Institut Universitaire de France. She received the PhD degree in Signal Processing in 2007. Her research interests include patch-based and deep learning-based methods for image processing and analysis. She has been regularly involved in works on colorization, inpainting and restoration since 2008. She now also studies environmental impacts of ICT.