Yes, "*Attention Is All You Need*", for Exemplar based Colorization

Wang Yin[1], Peng Lu[1*], Zhaoran Zhao[1] and Xujun Peng[2]
[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]ISI, University of Southern California, Marina del Rey, CA, USA
{yinwang,lupeng,zhaozhaoran}@bupt.edu.cn,xpeng@isi.edu

Figure 1: Colorization results outputted by the state-of-the-art approaches. The input and reference images are semantic related in the first and second rows. In the third and forth rows, the input and reference images are semantic independent. Note that the reference image is from palette dataset [1] in the forth row, which has no semantic meaning.

ABSTRACT

Conventional exemplar based image colorization tends to transfer colors from reference image only to grayscale image based on the semantic correspondence between them. But their practical capabilities are limited when semantic correspondence can hardly be found. To overcome this issue, additional information, such as colors from the database is normally introduced. However, it's a great challenge to consider color information from reference image and database simultaneously because there lacks a unified framework to model different color information and the multi-modal ambiguity in database cannot be removed easily. Also, it is difficult to fuse different color information effectively. Thus, a general attention based colorization framework is proposed in this work, where the color histogram of reference image is adopted as a prior to eliminate the ambiguity in database. Moreover, a sparse loss is designed to guarantee the success of information fusion. Both qualitative and quantitative experimental results show that the proposed approach achieves better colorization performance compared with the state-of-the-art methods on public databases with different quality metrics.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**.

KEYWORDS

Image understanding, GAN, Colorization

ACM Reference Format:
W. Yin, P. Lu, Z. Zhao, X. Peng. 2021. Yes, "*Attention Is All You Need*", for Exemplar based Colorization. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475385

* Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475385

1 INTRODUCTION

Colorization, which assigns colors to grayscale images to improve their visual qualities and makes them more plausible, is ill-posed and inherently ambiguous because multiple colors might be suitable for a single pixel in the input grayscale image. For example, a balloon can be colored in blue, red, or other colors without damaging the overall style of the image. Under this ambiguity circumstance, many approaches have been proposed to achieve better colorization performance, which can be roughly categorized as three types according to different types of color information sources: automatic colorization, user guided colorization and exemplar based colorization.

The early fully automatic colorization methods [3–6, 17, 18, 24, 35] can provide a reasonable coloring suggestion for grayscale image by using the color knowledge from the large scale data. However, most of them ignore the users' aesthetic feelings and preferences which are generally required for color designing tasks.

The user-guided colorization requires users to provide color strokes for particular objects in the grayscale image, where those colors are propagated based on similarity metrics between adjacent pixels or regions [13, 20, 22, 25]. Benefiting from the development of deep learning, several colorization approaches [26, 36] employing the colors learned from database along with the users' interactions are proposed. However, their efficiency are limited by extensive human efforts with professional knowledge.

In exemplar based colorization, early methods [2, 6, 9, 14, 32] required the semantic correspondence between the grayscale and reference images. Based on this correspondence, the colors can be transferred from reference image to grayscale image. But its effectiveness is low when the semantic correspondence doesn't exist. In this case, color information from the database can be introduced as additional information to perform colorization [10]. Thus, utilizing color information from large scale data and the reference image simultaneously attracts increasing research interests for exemplar based colorization with three types of colors, which are 1) semantic colors associated with the objects in reference image, 2) the global color distribution such as tones of reference image, and 3) the color information from the database.

Normally, it is hard to take these three types of colors into account for colorization at the same time under a unified framework. The challenge relies on the fact that when colors are from different color sources for a particular object, the problem of multi-modal is aroused which causes the colorization a hard problem. Furthermore, because multiple colors introduced from the large scale database could be assigned to the same object, to remove this ambiguity makes the colorization a even harder problem.

To address this challenge, He *et al.* [10] proposed a colorization method that utilized both reference image and database. When the semantic correspondence between reference image and the grayscale image was not established, they utilized the correspondence between semantics and colors contained in database and attempted to produce final color image naturally. However, this approach ignores the global tone of the reference image provided by the user, which causes the colorized results may look different from the reference image. On the other hand, in [21], Lu *et al.* used the color statistical information of reference image to predict the colors of the pixels in grayscale image when they were not semantic related to the reference image. In this situation, this approach does not consider the inherent relationship between semantics and colors, which tends to produce unnatural colorized image.

To solve aforementioned problems, we propose a new colorization method which fully employs the correspondence between semantics and colors for both reference image and images in the database to ensure the colorized image looks like the reference image and also natural. To achieve this goal, several contributions are made in this paper:

- We modeled the colorization as a query-assignment problem for different color sources, where the candidate colors from different color sources are determined by using the same attention mechanism. Under this unified framework, the colors, no matter they are from reference images or from the database images, they are selected and assigned to the grayscale image based on the same criterion where the semantic features are applied as the key.
- To solve the problem of ambiguity when colors are transferred from the database, the color distribution of reference image is utilized as the prior in the attention operation to constrain the searching process in the database. By using this idea, the colorized image can be like the reference image and natural at the same time, even reference image and the grayscale image are semantic independent.
- To encourage the proposed colorization system to select the color from one source and assign it to the particular pixel in grayscale image, a gating mechanism is designed in our work, where a simple but effective training loss (sparse loss) is proposed. This gating mechanism along with the sparse loss training guarantee the color selection operation is differentiable.
- Based on the proposed unified attention mechanism based colorization framework, superior colorization performance is achieved compared with other state-of-the-art approaches.

2 PREVIOUS WORK

Unlike the pioneer colorization works mostly relied on heuristic rules, recent researches tended to use learning based approach to discover the underlying patterns of colors according to the statistic properties of training samples. In [8], by using a random forest, an object function conditioned on image features was proposed, where correlations between colors in a long range within the image can be learned and used for colorization. Based on a cascade feature matching scheme, the correspondence between superpixels of the reference and input grayscale images was found initially for the color seeds assignment on grayscale image, which were smoothed by a voting approach afterwards [9].

Benefitting from the extensive success of deep learning, most recent colorization approaches relied on the deep neural network to learn the knowledge of colors from large scale data or transferred the colors from reference image to grayscale image.

In [18], the relationship between semantic and colors were learned from large amounts of data, the grayscale image was colorized according to its semantic information. When predict the color of grayscale image in ab channel of quantified CIELAB color space, to

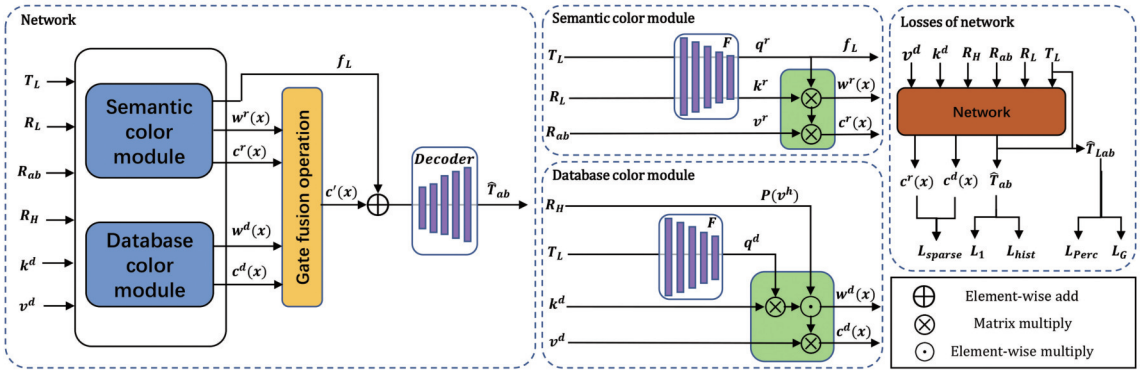


Figure 2: The architecture of the proposed networks and losses, where the left part of Fig. 2 is the overall framework containing two modules, which are semantic color module and database color module. As shown in the middle part of Fig. 2, T_L is the input grayscale image, R_{Lab} is the reference image, R_H is the color histogram of reference image, k^d and v^d represent the semantic and color information of the database respectively, and $\hat{T}_{ab}$ is the predicted ab channel of grayscale image.

avoid the unbalanced distribution of ab values in natural images, the class rebalance loss was introduced in [35]. In order to use the color 'hints' from users to boost colorization performance, the method proposed by [36] designed a convolutional neural network (CNN) to combine the high level color information from dataset and low level color cues from reference image to transfer colors to grayscale image. To overcome the problem of insufficient training pairs of images with meaningful semantics, [19] was proposed to apply the identical images with geometric distortion as the references to enlarge the corpus for sketch image colorization's training, where the data augmentation was implemented by a self-attention based network. By applying the existing object detectors, [28] effectively located and learned object level semantics and the colors assigned to the input image were predicted by the proposed similarity measure network based on semantics. In [31], the colorization was fulfilled in a two-stage manner, where a coarse color assignment was applied using a faster transfer network and the refined colorization was obtained by a robust network.

To thoroughly employ the colors from reference images, most modern colorization approaches applied pre-calculated semantic similarities between reference image and input grayscale image using empirical rules [6] or neural networks [10]. However, if each pixel's color in the grayscale image can be represented by the colors of certain areas in the reference image based on their similarities, the colorization can be modeled by attention mechanism, where each pixel of the grayscale image attends to different sources of colors to express it. Recently, attention based neural networks were widely used for NLP and CV and achieved the state-of-the-art performance in these areas [11, 29].

3 PROPOSED APPROACH

3.1 Problem formulation

Given a grayscale image $T_L \in \mathbb{R}^{H\times W\times 1}$ and a reference image $R_{Lab} \in \mathbb{R}^{H\times W\times 3}$, the exemplar based colorization aims to predict image T_L's missing ab channel $T_{ab} \in \mathbb{R}^{H\times W\times 2}$ in CIELAB color space, where H, W are the height and width of the image. Particularly, for a given pixel p_x in the grayscale image T_L, the optimal color is obtained by searching against reference image R_{Lab} or database and assigned to the pixel p_x based on its semantic property.

When the semantic correspondence exists between the pixel of grayscale image and reference image, we transfer colors from reference image R_{Lab} to grayscale image T_L by attention mechanism directly. Formally, the colors provided by the reference image can be represented as color values $v^r(i) \in \mathbb{R}^2, 1 \le i \le \mathcal{N}^r$ in ab channel of CIELAB color space, where $\mathcal{N}^r$ is the number of the available colors. For each color value $v^r(i)$, a semantic key $k^r(i)$, which is the corresponding semantic feature, is extracted from the L channel of image R_{Lab}. Thus, for a pixel p_x in the grayscale image T_L such that $1 \le x \le H \times W$, it can be considered as a query by using its multi-resolution semantic feature $q^r(x)$ to find the proper color values $v^r(i)$ through $k^r(i)$, which can be modeled as a query-assignment problem:

$$c^r(x) = \frac{1}{\mathcal{N}^r} \sum_{i\in\mathcal{N}^r} h\big(q^r(x), k^r(i)\big) v^r(i), \tag{1}$$

where the $h(\cdot)$ is a function measuring the similarity between query $q^r(x)$ and key $k^r(i)$, and the superscript r indicates the color is from reference image.

Based on the same idea, we can model the colorization process as a query-assignment problem by introducing additional color information from database when semantic correspondence between the pixel of grayscale image and reference image is not established, where the semantic color from reference image can hardly be transferred to the grayscale image. Thus, given the colors provided by the database, which can be expressed as color value $v^d(i) \in \mathbb{R}^2, 1 \le i \le \mathcal{N}^d$, the corresponding semantic features can also be extracted from the database and applied as the semantic key $k^d(i)$. And the high-level semantic feature for each pixel p_x in the image T_L is still considered as the query $q^d(x)$. Therefore, the process of finding the proper color value $v^d(i)$ for pixel p_x based on the semantic correspondence between query $q^d(x)$ and key $k^d(i)$ can also be modeled employing Eq. 1 instead of using colors from database, which is indicated by superscript d.

However, given a pixel p_x in grayscale image T_L, there might have multiple images/objects in the database that share the same

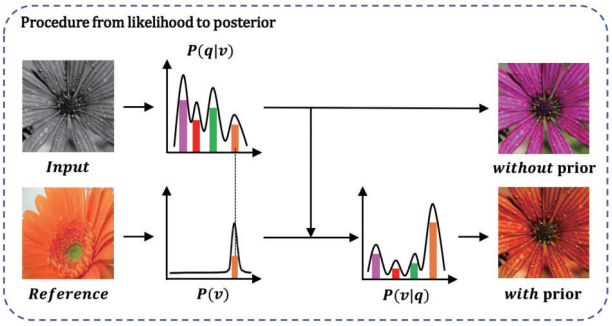


Figure 3: The procedure of converting the likelihood to posterior. $p(q|v)$ is the likelihood of colors from database, the prior $p(v)$ is the color histogram of reference image, $p(v|q)$ is the posterior of colors.

semantic as the grayscale image but have different colors. To tackle this ambiguity problem and inspired by the work from [21], where the color distribution of reference image is applied to guide the colorization when semantic correspondence between the reference image and grayscale image does not exist, we propose to use the color distribution $p(v)$ from the reference image as the prior to constrain the color searching process in database. And under the probabilistic framework, given a color v of a pixel, the likelihood of the corresponding semantic $p(q|v)$ can be learned from the database. Hence, based on the Bayesian rule, the probability of the color v of a pixel in the image given its semantic, which is the posterior of colors can be obtained:

$$p(v|q) \propto p(v) \cdot p(q|v), \tag{2}$$

In the scenario of transfer colors from database to the grayscale image, the likelihood $p(q|v)$ can be approximated by $h(q^d(x), k^d(i))$ because each key $k^d(i)$ in the database is associated with a color v, and the similarity between $q^d(x)$ and $k^d(i)$ can be used to measure the probability of semantic $q^d(x)$ of a pixel given this color. Thus, the posterior of colors from the database is proportional to:

$$p(v^h(i)) \cdot p(q^d(x)|v^h(i)) \propto p(v^h(i)) \cdot h(q^d(x), k^d(i)), \tag{3}$$

where $h(\cdot)$ is a function measuring the similarity between query $q^d(x)$ and key $k^d(i)$ in the database.

Based on the posterior obtained by Eq. 3, the process of assigning colors from database to the grayscale image can be modeled as:

$$c^d(x) = \frac{1}{\mathcal{N}^d} \sum_{i \in \mathcal{N}^d} p\big(v^h(i)\big) \cdot h\big(q^d(x), k^d(i)\big) \cdot v^d(i). \tag{4}$$

The proposed procedure of transfer colors from database to the grayscale image with the prior is demonstrated in Fig. 3. As can be seen, when the color distribution of the reference image is ignored, the colorization system tends to find the optimal color (e.g., purple) from the database based on the likelihood of this database and the colorized image and the reference image may have irrelevant appearance due to the color ambiguity issue. However, once we introduce the color distribution of the reference image as the prior into the system and convert the likelihood to the posterior, the searching space in the database is constrained by the prior provided by the reference image, which produces a colorized image with the same tone as the reference image (e.g., orange), which is preferred by the user.

In the proposed framework, there are more than one color source are available, which are colors from reference image and database respectively. Hence, each pixel in the grayscale image can obtain multiple assigned colors $c(x)$ from different sources, which can be combined to generate the final color according to:

$$c'(x) = \sum_{j \in M} \alpha^j c^j(x), \tag{5}$$

where superscript j means the type of color source, the set of which can be denoted by $M = \{r, d, \cdots\}$, and α^j represents the fusion weight for different color sources which satisfies $\alpha^j \geq 0$ and $\sum_{j \in M} \alpha^j = 1$.

To encourage each pixel of the grayscale image to be colorized by an existing color from different information sources rather than a mixed color by multiple color sources, the sparsity of fusion weights α^j for different color sources should be guaranteed, so the constraint $-\sum_{j=1}^{M} \alpha^j \cdot \log \alpha^j = 0$ is applied.

In summary, the process of colorization from multiple color sources is shown in the Algorithm 1.

Algorithm 1 Attention based colorization from multiple color sources

Input: a grayscale image $T_L \in R^{H \times W \times 1}$ and a reference image $R_{Lab} \in R^{H \times W \times 3}$, where H and W are height and width of the image

Output: the color $c'(x)$ for each pixel in T_L, $1 \leq x \leq H \times W$

1: **for** $j \in M = \{r, d, \cdots\}$ **do**
2: Build query $q^j(x)$ for each pixel of T_L
3: Build keys $k^j(i)$ for color values $v^j(i)$ of jth source, where $1 \leq i \leq \mathcal{N}^j$, $\mathcal{N}^j$ is the number of usable color for jth source
4: Get the color $c^j(x)$ for each pixel in T_L by:

$$c^j(x) = \frac{1}{\mathcal{N}^j} \sum_{i \in \mathcal{N}^j} p\big(v(i)\big) \cdot h\big(q^j(x), k^j(i)\big) v^j(i), \tag{6}$$

where $h(\cdot)$ is a similarity function, the $p\big(v(i)\big)$ is a prior information in different sources
5: **end for**
6: Use Eq. 5 to integrate different color $c^j(x)$
7: Return $c'(x)$

3.2 Colorization network

From the previous discussion, we can see that the challenge existing for the color assignment to the input grayscale image T_L lies in the establishment of correspondence between query $q(x)$ and color value $v(i)$ from different information sources. Also, the fusion of colors $c(x)$ from different color sources to each pixel p_x of T_L is another great challenge. To tackle these difficulties, a fully attention based colorization framework is proposed, which is shown in Fig. 2.

In our implementation, three modules are designed where the semantic color module is applied to extract optimal colors from the reference image by the semantic correspondences between

reference and grayscale images. And database color module can find semantic relationship between the database and grayscale image, then transfers appropriate colors from database to each pixel of the grayscale image. These two color information can be combined by gate fusion operation, whose output is passed into the decoder to produce the final colorful image $\hat{T}_{ab}$.

3.2.1 Semantic color module (SCM). When the pixels in the grayscale image are related to the reference image semantically, the prior knowledge of colors $p(v)$ expressed in algorithm 1 is a uniform distribution. In this case, we can transfer the semantic colors associated with the objects in reference image R_{Lab} to the grayscale image T_L directly based on their semantic similarities.

To obtain the query $q^r(x)$ of grayscale image's pixel and the key $k^r(i)$ of colors from the reference image, we adopt the feature extractor F, which is implemented by CNN networks [36], to extract the semantic features from grayscale image and the L channel of reference image respectively. Specifically, features from six layers of the encoder F and the L channel of the image are extracted, where each channel of these features are down-sampled or up-sampled to the same spatial size and concatenated to obtained query $q^r(x)$ and key $k^r(i)$. Consequently, the semantic correspondence between grayscale and reference image can be computed by performing attention calculation between query $q^r(x)$ and key $k^r(i)$. Based on the obtained attention weight between $q^r(x)$ and $k^r(i)$, it can be applied on the colors value v^r, which is from the ab channel of the reference image, to get the most appropriate color $c^r(x)$ for the grayscale image.

Moreover, in order to determine the importance of the color feature $c^r(x)$ for colorization, a confidence map $w^r(x)$ is adopted, which is obtained by passing the semantic correspondence between grayscale and reference images into a fully connected layer.

3.2.2 Database color module (DCM). When the available color information is from the database, the prior information $p(v)$ in Algorithm 1 is the color distribution of the reference image. Thus, we can transfer the color information of the database to the grayscale image T_L according to Eq. 4.

Similar to the semantic color module, we adopt the same feature extractor F to obtain the semantic feature as query $q^d(x)$ from grayscale image, but only use the features of last layer of F instead of intermediate features. Unlike the semantic color module that trains the query and the key simultaneously, due to the color ambiguity nature of the database, where a single color in the database might have multiple semantic representations, we only train the feature extractor F for query $q^d(x)$ but force the semantic key $k^d(i)$ of database color to a fixed value. In our implementation, an identify matrix is applied where each element represents the corresponding $k^d(i)$ for the database color $v^d(i)$. Therefore, the feature extractor F is trained to map multiple query $q^d(x)$ to $k^d(i)$. To establish the one-to-one mapping between $q^d(x)$ to $k^d(i)$, the color histogram R_H of the reference image, which is implemented with the same method as described in [35], is implemented as the prior to constrain the training process, as introduced in Eq. 4.

Same as the semantic color module, a confidence map $w^d(x)$ is extracted by passing the semantic correspondence between the database and grayscale image into a fully connected layer.

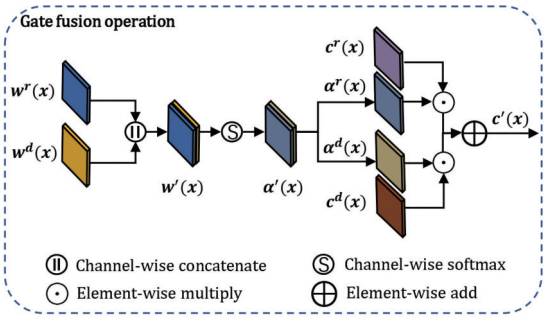


Figure 4: Structure of gate fusion operation.

More detailed structures of the proposed semantic color module and database color module can be seen in our supplemental material.

3.2.3 Gate fusion operation. According to Algorithm 1, the color $c^r(x)$ from semantic color module and $c^d(x)$ from database color module are combined based on Eq. 5. Without gate fusion operation, each pixel of input grayscale image T_L will be assigned with a mixed color from multiple color sources rather than an existing color selected from a single color source, which is not ideally w.r.t. the user's preference .

In order to obtain the fusion weights of different color sources, we concatenate the confidence map $w^r(x)$ and $w^d(x)$ learned from the semantic color module and database color module as described in previous subsection to form weight matrix $w'(x)$. Afterwards, the weight matrix $w'(x)$ is fed into a softmax layer channel-wisely to calculate normalized weight $\alpha'(x)$, which is exported separately as the fusion weight $\alpha^r(x)$ and $\alpha^d(x)$. These fusion weights along with the obtained color $c^r(x)$ from semantic color module and $c^d(x)$ from database color module are used to compute the fused color $c'(x)$ according to Eq. 5. The detailed structure of this gate fusion module is illustrated in Fig.4.

Finally, we utilize the decoder with the same structure as the network proposed in [34] to upsample the fused color $c'(x)$ and the semantic features f_L of grayscale image T_L to the original size and produce the final colorful image $\hat{T}_{ab}$ by using the de-convolutional layers.

3.3 Losses

3.3.1 L1 loss. In order to constraint the relationship between T_{ab} and $\hat{T}_{ab}$ without using the average scheme to solve the multi-modal ambiguity problem of colorization [10, 33, 36], the $\mathcal{L}_1$ loss is applied in our work:

$$\mathcal{L}_1 = ||T_{ab} - \hat{T}_{ab}||_1, \tag{7}$$

3.3.2 Perceptual loss. To obtain a more perceptually plausible output, we use perceptual loss [16] to measure the semantic difference between the output $\hat{x}$ and the ground truth image x:

$$\mathcal{L}_{Perc} = ||\Phi^L_{\hat{x}} - \Phi^L_x||^2_2, \tag{8}$$

where Φ^L represents the features extracted from the $5th$ ReLU layer in the pre-trained VGG19 [27] network to obtain more meaningful semantic features.

3.3.3 GAN loss. In order to encourage the proposed image colorization network to produce realistic images, large amounts of real color images X are used with the LS-GAN loss [23] in the training phase. The losses of the generator L_G and discriminator L_D are defined as:

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{2}E_{\hat{T}_{Lab}\sim P_T}[(D(\hat{T}_{Lab}))^2 - 1], \tag{9}$$

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}E_{\hat{T}_{Lab}\sim P_T}[(D(\hat{T}_{Lab}))^2] + \frac{1}{2}E_{X\sim P_X}[(D(X)-1)^2], \tag{10}$$

where P_T and P_X are the distributions of output colorful images and reference colorful images X respectively.

The patch based discriminator [15] is also applied in our work to fully capture the high-frequency structures from the images.

3.3.4 Sparse loss. Aiming at assigning a single color from different sources to each pixel of the grayscale image T_L efficiently, the sparse loss is proposed in our gate fusion operation to guarantee the sparsity of fusion weight $\alpha^j(x)$. Therefore, for each pixel in the grayscale image, a color from an existing information source rather than a new mixed color by multiple color sources is assigned. When the loss converges, the fusion weight $\alpha^j(x)$ which is close to 0 or 1 infinitely can indicate which color source is selected. The loss function is expressed as:

$$\mathcal{L}_{sparse} = \sum_{j\in M} -\alpha^j(x)\cdot log(\alpha^j(x)), \tag{11}$$

where j denotes the color source, which is selected from set $M = \{r, d\}$ in our case.

3.3.5 Color histogram loss. To transfer the global tone of the reference image to the input grayscale image, the color histogram loss $\mathcal{L}_{hist}$ [21] is employed, which forces the input grayscale image to have the same global color distribution as the reference image.

3.3.6 Total loss. The losses described above are added with different weights to compose the total loss, which can be expressed as:

$$\mathcal{L}_{total} = \lambda_{Perc}\mathcal{L}_{Perc} + \lambda_{\mathcal{G}}\mathcal{L}_{\mathcal{G}} + \lambda_{sparse}\mathcal{L}_{sparse} + \lambda_1\mathcal{L}_1 + \lambda_{hist}\mathcal{L}_{hist}, \tag{12}$$

where λ_{Perc}, λ_{adv}, λ_{sparse}, λ_1, λ_{hist} are the weights of different losses.

4 EXPERIMENTAL RESULTS

4.1 Datasets

To train and test the proposed colorization framework, we use the same training dataset as [21], which contains more than 1.2 millions images from 1000 categories. And the training images are all resized to ensure the short side of each image is 288 pixels.

To evaluate the performance of our model, two subsets containing 5000 image pairs whose short side is cropped into 256 pixels are constructed from the ImageNet ILSVRC 2012 Test Dataset [7] and the palette dataset [1]. To build the semantic related subset (subset-1) with 2500 image pairs, given an input grayscale image T_L randomly selected from ImageNet Test Dataset, an image from the same dataset with closest distance to T_{Lab} in VGG19 feature space is selected as the corresponding reference image R_{Lab}. The semantic independent subset (subset-2) contains 2500 image pairs. Among these images, half of them are constructed by selecting a corresponding reference image R_{Lab} with the farthest distance to T_{Lab} in VGG19 feature space from ImageNet Test Dataset for each input grayscale image T_L. And the corresponding reference images R_{Lab} for the other 1250 grayscale images are randomly chosen from palette dataset.

4.2 Training setup

The training procedure is carried out in two stages. In the first stage, we train the feature extractor F and decoder network together for 10 epochs, the input and output of this part are T_L and $\hat{T}_{ab}$ respectively. The loss functions in use are perceptual loss, GAN loss and $\mathcal{L}_1$ loss. In the second stage, the total network, which takes T_L, R_{Lab}, R_H, k^d and v^d as the inputs and has the same output as the first stage, is trained for 10 epochs end-to-end. The color histogram loss and sparse loss are used in this stage along with other loss functions mentioned in the first stage.

In this paper, We set λ_{Perc} = 1000.0, $\lambda_{\mathcal{G}}$ = 0.1, λ_{sparse} = 1.0, λ_1 = 1000.0, λ_{hist} = 1.0. Adam solver is adopted for optimization with parameters β_1 = 0.5 and β_2 = 0.99, where the learning rate is set to be 2e-5 without any decay schedule. The total network is trained for 20 epochs with batch size of 6.

4.3 Comparison with the state-of-the-art

To evaluate the performance of the proposed colorization framework, the results of our approach are compared with other state-of-the-art image colorization methods quantitatively and qualitatively, including one automatic colorization method [18], two histogram colorization methods [30, 36] and two exemplar based colorization methods [10, 21].

4.3.1 Qualitative comparison. A qualitative comparison of selected representative cases is shown in Fig. 1.

When semantic related reference images are provided, one can observe that our method along with [10] can transfer the color of reference image to gray image effectively, such as the bird and plant shown in the first row and the second row in Fig. 1. By contrast, automatic colorization method [18] assigns colors to grayscale image from the database but ignores colors from the reference image. And due to the lower resolutions of the training samples, the latest work [21] fails to capture the details of semantic colors associated with the objects in the reference image, which can be seen from the body of the bird in the first row. Moreover, the histogram colorization based methods [30, 36] are defective which only transfer the global tone of the reference image.

When the input and reference images are semantic independent, and the reference image itself is semantically meaningful as shown in the third row of Fig. 1, our approach along with histogram colorization based methods [30, 36] can transfer the global tone of reference image to the grayscale image successfully. However, the latter provides a colorized result with lower aesthetic quality compared with ours. As can be seen in the third row of Fig. 1, He *et al.*'s approach [10] only transfers colors from database and fails to provide an acceptable result because of the ignorance of the reference image. Also from the third row of Fig. 1, one can see that the inconsistent color of the sky obtained by [21] is unnatural

Figure 5: The colorization results of different approaches, which indicates excessive color transfer phenomenon when a higher HIS is accompanied with a lower AS.

compared to our approach, which shows the positive effects of the sparse loss in constraining the fusion of multiple color sources.

Under the circumstance that the reference image does not any meaningful semantic information, as shown in the forth row of Fig. 1, the colorized result of our method is more saturated compared with [21, 30, 36]. This phenomenon shows the database color module successfully utilizes the global tone of reference image as prior to constrain the colorization to avoid color ambiguity. Compared to our result, the colorized image of [10, 18] is less satisfactory due to ignoring the global tone of reference image.

From the analysis of Fig. 1, it can be observed that the proposed method provides remarkable performance for different types of reference image, which fully demonstrates the high efficiency of our colorization framework. More comparisons of our work and other state-of-the-art approaches can be found in the supplement.

4.3.2 Quantitatively comparison. To evaluate our approach quantitatively, the scores of histogram intersection similarity (HIS) [15] is used to measure the color transfer quality from reference image to colorized result. Furthermore, the aesthetics scores (AS) in [12] is also applied in our work to evaluate the aesthetic sense degree of colorized image, which represents the level of users' preference, where a higher AS represents a more visually pleasure image.

The comparisons of HIS and AS for different colorization approaches are shown in Table 1. As can be noticed from this table, benefitting from the proposed attention based colorization framework, our method achieves the highest AS score in both semantic related/independent situations. However, several baseline approaches over-transfer the colors to the grayscale image, whose colorized results have low aesthetic quality but with high HIS. More examples of this phenomenon can be found in Fig. 5 where the reference images are semantic related to the grayscale images. Compared to the over-transferred colorized results in the first row by [10, 21, 36], only our result preserves the uniform color for corn and background. And from the second row in Fig. 5, it can be seen that only the proposed method successfully transfer the color of insect from the reference image to the grayscale image appropriately without over transferring the color to the background. Based on this phenomenon, we conclude that a colorization method with high HIS is not necessarily to have high AS, and the HIS is more appropriate in our colorized results because they deliver more satisfactory images.

Table 1: HIS and AS scores of colorization results for different methods on our evaluation datasets.

	Semantic Related		Semantic Independent	
Approach	HIS ↑	AS ↑	HIS ↑	AS ↑
Larsson [18]	0.44	3.63	0.29	3.83
Zhang [36]	0.67	3.71	0.57	3.9
Xiao [30]	0.66	3.56	0.56	3.76
He [10]	0.70	3.64	0.53	3.87
Lu [21]	0.74	3.71	0.62	3.89
Ours	0.67	3.75	0.51	3.98

4.4 User study

To evaluate the attractiveness of our results by human observers, the perceptual assessment is performed. In this experiment, our method is compared with five state-of-the-art approaches by evaluators based on the quality of colorized results. 50 pairs of evaluation samples from the test dataset are randomly selected, where the semantic related pairs and semantic independent pairs are evenly distributed. 300 colorized images are generated and demonstrated to 20 evaluators. We follow the same scoring procedure as described in [21].

The average quality scores for each colorization method are listed in Table 2. It can be observed that the automatic colorization method [18] achieves the lowest perceptual quality score due to the ignorance of color information from reference image. And one can see that the perceptual quality score of the proposed method is the highest in two subsets, which indicates our approach tends to obtain colorized results with more visual pleasure compared with other exemplar based approaches.

4.5 Ablation

4.5.1 Database module with attention. To investigate the effectiveness of database color module, we build an ablated model without it. From Table 3, one can observe the decrease of HIS and AS scores of the ablated model compared with the full model, when there is few

Table 2: Quality scores of colorization results provided by evaluators for different methods.

Eval set	Larsson [18]	Zhang [36]	Xiao [30]	He [10]	Lu [21]	Ours
Subset-1	2.69	3.33	3.11	3.76	3.52	4.08
Subset-2	2.51	3.27	2.94	2.81	3.42	3.95
All set	2.6	3.3	3.03	3.29	3.47	4.02

Input Reference Without DCM With DCM

Figure 6: The colorization results with or without database color module (DCM).

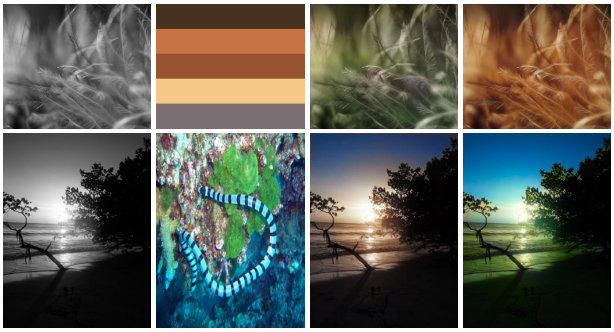
semantic correspondence between grayscale and reference image. As shown in Fig. 6, the sky in the first line and the apple in the second line are colorized unsuccessfully by ablated model, while full model can utilize colors from database efficiently by the attention mechanism which establishes the correspondence between colors and semantics in the database.

Table 3: HIS and AS scores of colorization results for ablated model and our full model.

	Semantic Related		Semantic Independent	
Approach	HIS ↑	AS ↑	HIS ↑	AS ↑
w/o attention in database	0.65	3.73	0.44	3.90
w/o prior in database	0.66	3.74	0.47	3.93
w/o sparse loss	0.64	3.74	0.48	3.93
Full	0.67	3.75	0.51	3.98

4.5.2 Color histogram as prior from reference image. To demonstrate the importance of the prior, which is the color histogram of reference image, we construct a database color module without prior. When there is few semantic correspondence between grayscale and reference images, the ablated model assigns colors from database to the grayscale image directly based on the semantic similarity between grayscale image and images in the database. An obvious decrease of both HIS and AS scores in ablated model can be observed in Table 3, which indicates a weakened color transfer ability. Also, as shown in Fig. 7, the grass in the first line and the nature scenery in the second line are colorized by the information from database in our ablated model which are unlike the reference image. However, by introducing the color histogram of reference image as the prior, the full model outputs colorized results with user's preference.

4.5.3 Sparse loss. To demonstrate the importance of sparse loss constraining multiple color sources fusion, all colors provided by



Input Reference Without prior With prior

Figure 7: The colorization results with or without the prior from reference image in our database color module.

Input Reference Without sparse loss With sparse loss

Figure 8: The colorization results with or without sparse loss.

different sources are directly added before being processed by the decoder in the ablated model. It can be seen that there are obvious decreases for both HIS and AS scores in Table 3. And the color discontinuity also appears in the results of ablated model, such as the background in the fist line of Fig. 8 and the sky in the second line. Based on the aforementioned analysis, our sparse loss is able to guide the process of selecting colors with higher aesthetic quality for grayscale images.

For the space limit, our attentions are primarily paid to verify the effect of sparse loss. The validity of other losses are proved in [21].

5 CONCLUSIONS

In this paper, a general colorization framework is proposed to perform colorization from multiple sources, where attention is all you need. It's ability to model valuable color source has largely increased performance of colorization. To eliminate the ambiguity introduced by correspondence between colors and semantic information from the database, we adopt color histogram of reference image as the prior to improve the accuracy of colors utilization from database. Moreover, to ensure the sparsity of fusion weights for different color information sources, we propose a sparse loss which allows our framework to effectively combine different color sources. Our method achieves plausible results compared with the state-of-the-art colorization approaches.

REFERENCES

[1] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Coloring with Words: Guiding Image Colorization Through Text-based Palette Generation. In *Computer Vision – ECCV 2018*. 431–447.

[2] A. Bugeau, V. Ta, and N. Papadakis. 2014. Variational Exemplar-Based Image Colorization. *IEEE Transactions on Image Processing* 23, 1 (2014), 298–307.

[3] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. 2008. Automatic Image Colorization Via Multimodal Predictions. In *Computer Vision – ECCV 2008*. 126–139.

[4] Z. Cheng, Q. Yang, and B. Sheng. 2015. Deep Colorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 415–423.

[5] Z. Cheng, Q. Yang, and B. Sheng. 2017. Colorization Using Neural Network Ensemble. *IEEE Transactions on Image Processing* 26, 11 (2017), 5491–5505.

[6] Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. 2011. Semantic Colorization with Internet Images. *ACM Trans. Graph.* 30, 6 (2011), 1–8.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[8] A. Deshpande, J. Rock, and D. Forsyth. 2015. Learning Large-Scale Automatic Image Colorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 567–575.

[9] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. 2012. Image Colorization Using Similar Images. In *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*. 369–378.

[10] Mingming He, Dongdong Chen, Jing Liao, Pedro V. Sander, and Lu Yuan. 2018. Deep Exemplar-Based Colorization. *ACM Trans. Graph.* 37, 4, Article 47 (2018).

[11] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Advances in Neural Information Processing Systems 32*. 11137–11147.

[12] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective Aesthetics Prediction With Multi-Level Spatially Pooled Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9375–9383.

[13] Yi-Chin Huang, Yi-Shin Tung, Jun-Cheng Chen, Sung-Wen Wang, and Ja-Ling Wu. 2005. An Adaptive Edge Detection Based Colorization Algorithm and Its Applications. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*. 351–354.

[14] Revital Irony, Daniel Cohen-Or, and Dani Lischinski. 2005. Colorization by Example. In *Eurographics Symposium on Rendering*.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-To-Image Translation With Conditional Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1125–1134.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision – ECCV 2016*. 694–711.

[17] S. H. Kang and R. March. 2007. Variational Models for Image Colorization via Chromaticity and Brightness Decomposition. *IEEE Transactions on Image Processing* 16, 9 (2007), 2251–2261.

[18] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Learning Representations for Automatic Colorization. In *Computer Vision – ECCV 2016*. 577–593.

[19] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. 2020. Reference-Based Sketch Image Colorization Using Augmented-Self Reference and Dense Semantic Correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5801–5810.

[20] Anat Levin, Dani Lischinski, and Yair Weiss. 2004. Colorization Using Optimization. *ACM Trans. Graph.* 23, 3 (2004), 689–694.

[21] Peng Lu, Jinbei Yu, Xujun Peng, Zhaoran Zhao, and Xiaojie Wang. 2020. Gray2ColorNet: Transfer More Colors from Reference Image. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3210–3218.

[22] Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and HeungYeung Shum. 2007. Natural Image Colorization. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*. 309–320.

[23] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. 2017. Least Squares Generative Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2813–2821.

[24] Yuji Morimoto, Yuichi Taguchi, and Takeshi Naemura. 2009. Automatic Colorization of Grayscale Images Using Multiple Images on the Web. In *SIGGRAPH '09: Posters*.

[25] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. 2006. Manga Colorization. In *ACM SIGGRAPH 2006 Papers*. 1214–1220.

[26] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2017. Scribbler: Controlling Deep Image Synthesis With Sketch and Color. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5400–5409.

[27] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR*.

[28] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-Aware Image Colorization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7968–7977.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. 5998–6008.

[30] Chufeng Xiao, Chu Han, Zhuming Zhang, Jing Qin, Tien-Tsin Wong, Guoqiang Han, and Shengfeng He. 2020. Example-Based Colourization Via Dense Encoding Pyramids. *Computer Graphics Forum* 39, 1 (2020), 20–33.

[31] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. 2020. Stylization-Based Architecture for Fast Deep Exemplar Colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9363–9372.

[32] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. 2005. Local color transfer via probabilistic segmentation by expectation-maximization. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 747–754 vol. 1.

[33] Bo Zhang, Mingming He, Jing Liao, Pedro V. Sander, Lu Yuan, Amine Bermak, and Dong Chen. 2019. Deep Exemplar-Based Video Colorization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8052–8061.

[34] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. 2019. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2019), 1947–1962.

[35] Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful Image Colorization. In *Computer Vision – ECCV 2016*. 649–666.

[36] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Trans. Graph.* 36, 4, Article 119 (2017).