IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2021, accepted August 9, 2021, date of publication August 12, 2021, date of current version August 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104515

Grayscale Image Colorization Methods: Overview and Evaluation

IVANA ŽEGER, SONJA GRGIC, (Member, IEEE), JOSIP VUKOVIĆ, (Member, IEEE), AND GORDAN ŠIŠUL, (Member, IEEE)
Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
Corresponding author: Ivana Žeger (ivana.zeger@fer.hr)

ABSTRACT Colorization is a process of converting grayscale images into visually acceptable color images. The main goal is to convince the viewer of the authenticity of the result. Grayscale images that need to be colorized are, in most cases, images with natural scenes. Over the last 20 years a wide range of colorization methods has been developed – from algorithmically simple, yet time- and energy-consuming because of unavoidable human intervention to more complicated, but simultaneously more automated methods. Automatic conversion has become a challenging area that combines machine learning and deep learning with art. This paper presents an overview and evaluation of grayscale image colorization methods and techniques applied to natural images. The paper provides a classification of existing colorization methods, explains the principles on which they are based, and highlights their advantages and disadvantages. Special attention is paid to deep learning methods. Relevant methods are compared in terms of image quality and processing time. Different metrics for color image quality assessment are used. Measuring the perceived quality of a color image is challenging due to the complexity of the human visual system. Multiple metrics used to evaluate colorization methods provide results by determining the difference between the predicted color value and the ground truth, which in several cases is not in coherence with image plausibility. The results show that user-guided neural networks are the most promising category for colorization because they successfully combine human intervention and neural network automation.

INDEX TERMS Automatic methods, black-and-white image, colorfulness, colorization, deep learning methods, example-based methods, grayscale image, image quality assessment, scribble-based methods, user-guided methods.

I. INTRODUCTION

Historical black-and-white images are regarded as irreplaceable, with exceptional artistic value. However, by looking at them it is impossible to fully imagine the actual scene, as color is a very important component of visual representation. The colorization of black-and-white images greatly alters the perspective of the viewer. The time gap between the past and the present fades away while making the scene more conceivable. However, insight into the authentic colors of early photographs is often unavailable, making satisfactory reconstruction difficult. Nevertheless, the aim of colorization is to deceive the viewer, to make him believe in the authenticity of the colorized image, and not to reconstruct the color accurately. Main applications of colorization include the revival of historical black-and-white images, movie restoration and coloring astronomy photographs.

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

Color is a subjective response of the human visual system to electromagnetic radiation of the visible spectrum with wavelengths between 380 nm and 780 nm [1]. It is a property of an object that can be described by hue, lightness, and saturation. Understanding color perception involves physics, physiology, and psychology. The perception of color depends on vision, light and individual interpretation.

Colorization is essentially a process of assuming color information where it is absent. Technically, it is a challenging process of assigning three-dimensional RGB (Red, Green, Blue) color information to each pixel with respect to intensity of a grayscale image in a visually acceptable, plausible way. To reduce the complexity of the task, a conversion to a convenient luminance-chrominance color space is used in the colorization process [2], [3]. YUV and CIELAB are color spaces derived from RGB. CIELAB is a perceptually uniform color space obtained from RGB by nonlinear transformations [4]. The uniform changes of the components in CIELAB match the uniform changes of human color experience. For

this reason, observing two distinct colors in CIELAB can be approximated by the Euclidean distance between the corresponding points in the color space. YUV is derived from RGB by linear transformations [4], [5] and is not perceptually uniform. Both YUV and CIELAB separate the luminance component from the color information, allowing the exploitation of intensity information and easier prediction of the two remaining color channels. The Y component in YUV represents the luminance, while U and V are the chrominance components. In CIELAB, L is the luminance component, while *ab* components carry the color information - *a* represents the green-red axis, while *b* represents the blue-yellow axis. Different colorization methods work with different color spaces. While some authors analyze the influence of various color spaces in the colorization process [6], [7], many choose the convenient one and develop the method with the selected color space [8]–[11].

In most cases, there is no unique color that can be associated with a particular gray object (e.g., balloons, clothing, plastic objects, etc.). Therefore, the existence of many objects in the world that appear with a great variety of colors makes colorization an ill-posed problem (no unique solution exists). This complexity provides constant interest in the research community and makes colorization a compelling problem.

Early manual colorization techniques date back to the 19th century [12]. Well-known techniques from that period include coloring a daguerreotype with a mixture of gum arabic and pigments, as well as photochrom process [13]. In the 1970s, following the impact of the digital revolution, colorization was transferred to the computer domain. The term ''colorization'' was introduced by Wilson Markle to describe the computer-assisted process of adding color to black-and-white movies or TV programs [14]. It is known that these colorization attempts resulted in low contrast, washed-out and pale colors, but the reason behind it is unknown because the details of the colorization procedures are proprietary. In addition, significant human intervention was required in the colorization process.

Technological developments have brought automated machine learning, and especially deep learning techniques into focus. These techniques have demonstrated their effectiveness in various computer vision and image processing applications [15]. In recent years, deep learning models have shown remarkable success in many different application domains (e.g., image classification, pedestrian detection and tracking, face detection, handwritten character classification, image super-resolution, photo adjustment, photo enhancement, sketch simplification, style transfer, inpainting, image blending, denoising, etc.) [9], [16] and thus promise more innovative improvements in the near future. Both machine learning and deep learning handle huge amounts of data efficiently while unfolding hidden patterns and producing satisfactory approximations of latent knowledge. While machine learning defines a set of rules in the data by extracting features regarding some form of a priori knowledge, its narrower field, deep learning, extracts regularities more independently using a hierarchical level of artificial neural networks. In this way, achieving exceptional colorization advantages has become possible. Moreover, quality assessment of the colorization results remains an active topic in the research community.

In this paper, a detailed review of the existing colorization methods was conducted along with the quantitative evaluation of the results. The algorithms of Iizuka *et al.* [6], Zhang *et al.* [8], Levin *et al.* [3], Su *et al.* [10], Vitoria *et al.* [11] and Zhang *et al.* [9] (both automatic and interactive versions) were analyzed and evaluated. Colorization was performed on five distinctive photographs made by the author. In these photographs, natural scenes with considerable differences regarding color are shown. In addition to the usual metrics for image quality evaluation, such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [17], quaternion SSIM (QSSIM) [18], a new perceptual similarity method, learned perceptual similarity metric (LPIPS) [19] and patch-based contrast quality index (PCQI) [20] were examined. Also, methods that try to measure the visual quality of the image, and not only fidelity to the original, were used. Such measures are colorfulness [21], underwater image quality measure (UIQM) [22], and underwater image quality evaluation metric (UCIQE) [23].

The remainder of this paper is organized as follows. In Section II, the classification of the methods has been performed, with the emphasis on the advantages and disadvantages of each category. In Section III, colorization results from described methods have been shown and evaluated. The paper ends with the Conclusion.

II. COLORIZATION METHODS

Many computer science research fields are intertwined and incorporated in the colorization process. Research papers dealing with colorization differ significantly because of the diversity of the proposed problem-solving approaches. The imagination and variety of the approaches make the categorization of numerous colorization methods extremely difficult. Until recently, most existing papers have classified the colorization methods considering the amount of user involvement in problem solving and the way of retrieving the required data [6], [24]–[30]. A rough division into the scribble-based and the example-based methods was introduced. The source (reference) images for the example-based methods could be obtained manually or automatically. However, this manner of classification has become obsolete. Deep learning techniques started showing remarkable advances using incomparably larger number of source images than the traditional example-based methods thereby indicating the need for separation from the example-based category. Therefore, contemporary papers introduce the use of deep learning models with a large quantity of training data as an additional criterion in the classification of colorization methods [11], [31]. Consequently, colorization methods are divided into scribble-based, example-based, and learning-based (or deep learning) methods. Deep learning methods attract the most attention. Anwar *et al.* [32] suggest dividing

the deep learning methods into seven categories based on the differences in domain type, neural network structure, auxiliary input, and final output.

Still, some of the recent research continues to divide all the existing methods into two categories [9], [33], [34], i.e., guidance (or user-guided) and no-guidance (or data-driven automatic) regarding the amount of the user interaction as the main division criterion. Controllability and interactivity are considered crucial in image editing [9]. User involvement creates the opportunity for error correction. In this case, the user-guided methods basically include the scribble-based and the example-based methods, while the automatic category can be equated with the deep learning methods. Even though colorization became involved with machine learning with the appearance of the example-based category, user intervention remained necessary in providing *a priori* knowledge about the problem or choosing suitable reference images. Some contemporary methods may be classified in both user-guided and deep learning categories [9], [10].

A. USER-GUIDED METHODS

1) SCRIBBLE-BASED METHODS

Scribble-based methods find the inspiration in early digital colorization attempts [14]. The technicians used to choose by hand the convenient colors, true or hypothetical, for each object in an image. Manual image segmentation procedures, despite being complex and time-consuming, were also applied. Each region was assigned a suitable color, like in a coloring book. Art connoisseurs had objections to colorization process because of the fear of the disruption of artistic expression which, along with financial and time costs, were the main reasons for the stagnation of colorization attempts. The increase in computer power and the development of algorithms reencouraged interest in colorization, especially in the computer vision research community.

Nowadays, scribble-based methods require annotating the grayscale image with marks of convenient colors, i.e., color strokes, scribbles. They serve as a landmark for colorization. The scribbles are user-made and placed upon certain areas of the image. The color from the scribble is propagated across the image to the borders specified by the intensity according to an optimization framework.

The basic scribble-based method is presented in Levin *et al.* [3]. Spatial continuity is exploited under the assumption that neighboring pixels in the space-time domain that have similar intensities should have similar colors. By working in YUV color space, color is assumed to be a linear function of intensity Y. The least squares optimization is used. The scribbles are formed as linear constraints of the optimization problem. Various subsequent, enhanced methods [24], [25] have taken over the optimization function of this method, or its slightly revised version. An example of the colorization result made by applying scribbles is shown in Fig 1.

Advantages of the scribble-based methods include globality – there is no need for explicit segmentation due to color propagation limited by intensity values. Furthermore, there is no need for searching an adequate, possibly unreachable reference image. The user may allocate the scribbles strategically or even add more scribbles if needed. Moreover, the user has the potential of adjusting the assigned color to a more desirable one.

On the other hand, scribble-based methods tend to be tedious while demanding significant human effort considering the necessary time, experience, and sense of aesthetics. Careful selection of palette colors is a prerequisite. Also, a substantial number of scribbles is required to yield a reliable result. In addition, color bleeding at image edges is highly possible.

2) EXAMPLE-BASED METHODS

The difficulties of scribble-based methods may be alleviated by selecting a similar reference color image for colorization. The idea behind the procedure is based on transferring color information from a color source image to the matching regions of the target grayscale image. This procedure reduces but does not exclude human activity in colorization process.

The improvement of Internet search engines and the appearance of centralized and indexable image databases enabled the transfer of the general ''atmosphere'' between images, as in Welsh *et al.* [2]. The goal is to locate pixels

(a) (b) (c)

FIGURE 1. a) Original photograph, b) scribbles applied on the grayscale version of the photograph, c) colorization results.

with corresponding luminance values in both target and reference images by means of neighboring pixels' statistical information and similar texture recognition. Although local information is crucial for properly separating the boundaries of objects within an image, the use of global information highly increases the probability of accurate color assignment.

Since 2005, mentioned methods have been reviewed, improved, combined and concatenated. Work on image processing methods and machine learning techniques has been intensified. Image segmentation is used for more efficient color assignment in Irony *et al.* [24]. As shown in Fig. 2., a suitable feature space for region differentiation is constructed using k-nearest neighbors (k-NN) algorithm. The classification of each pixel is done by voting both in feature and image spaces leading to better spatial consistency. The influence of shadows and light reflections, as well as changing lighting conditions is considered in Liu *et al.* [25]. The emphasis is switched to a global problem formulation with a statistical approach to its solving in Charpiat *et al.* [26]. This way, the framework becomes more robust to noise and local prediction errors. The probability distribution of every possible color of a pixel is calculated, thus resolving ambiguities more efficiently than previous methods.

Since 2010, more attention has been paid to the qualitative and quantitative comparison of colorization methods. Additionally, an increasing number of reference images is used while searching for the corresponding color of segmented image objects. Several local features (e.g., intensity, scale invariant feature transform (SIFT), speeded-up robust features (SURF), Gabor features) and their influence on color transfer are explored in Chia *et al.* [27] and Gupta *et al.* [28]. For feature studies, groups of pixels like patches and superpixels are used while exploiting spatial consistency. The grouping of segments assigned with similar color values is carried out with k-means algorithm, gaining reliability of color assessment. More complicated mathematical formulations of loss functions with manually adjustable parameters are used along with more advanced optimization methods in Deshpande *et al.* [34], leading to better spatial consistency and visually more appealing and convincing results.

There are several limitations of the example-based methods. The major limitation is the possible non-existence of a single suitable reference image. Also, the selection of the appropriate reference image is often done manually. The quality of the result is highly dependent on the quality of the reference image used. The target image and the reference image need to implicate visual similarity. To ensure correct color transfer, the objects in the scene are supposed to be similar in both the target and reference images. Algorithm overfitting is risked by the usage of a single reference image (or even a small number of images).

However, example-based methods are characterized by simplicity and speed.

B. DEEP LEARNING METHODS

The expectation that one or more reference images might contain sufficient color information for satisfactory colorization results is usually not realistic. The evolution of deep learning techniques has enabled training of an artificial neural network with a large number of source images. For colorization, it means automatically learning colors that naturally correspond to real objects. The methods yield better results by adding more layers to the neural network and more images to the training set. Deep neural networks were introduced to the colorization problem by Cheng *et al.* [29]. Neural networks automatically extract regularities within data by minimizing the corresponding loss function in the training phase.

The neural network model automatically learns a mapping function between the features of the pixels in a grayscale image and the color values of the source images. In recent times, the need for user intervention in colorization process has been almost entirely excluded, even though not entirely rejected. Although reducing the need for user effort is the main advantage of deep learning methods, their main drawback is that numerous parameters need to be tuned to achieve satisfactory results. Moreover, training a neural network with a large training dataset requires a significant amount of time

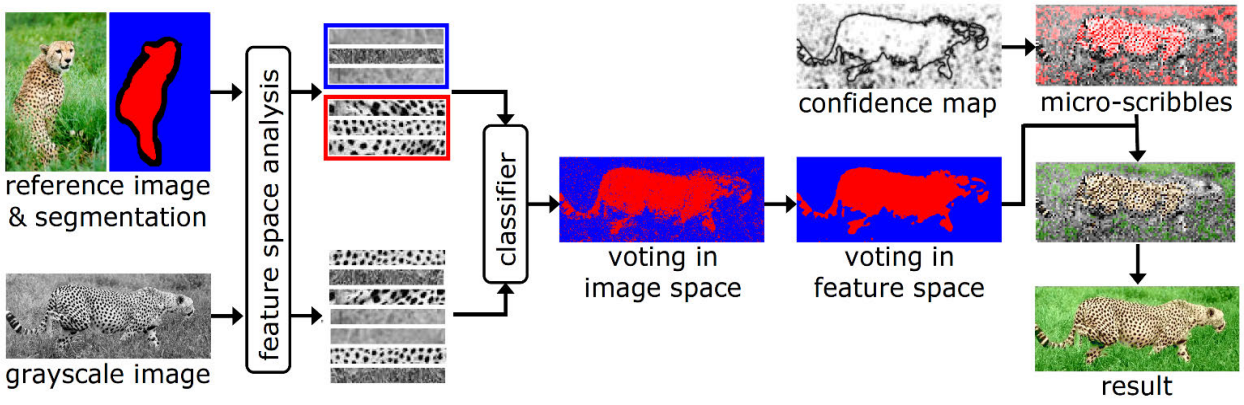


FIGURE 2. Structure of Irony *et al.* example-based method [24].

(days, even weeks). Additionally, the system uses dominant colors learned from the dataset to generate results and is therefore frequently slightly biased.

When dealing with images, basic deep neural networks remove the spatial, two-dimensional form of an image. The decomposition of an image into a one-dimensional vector of numbers occurs at the input, before forwarding to subsequent layers. Convolutional neural networks (CNNs), which are widely used in image processing, preserve the spatial information, and are therefore also used in colorization [6], [8], [16], [30], [35].

At the beginning of their development, deep learning colorization methods were considered as a modified type of the example-based methods. They were characterized as parametric methods that learn prediction functions from a large dataset and define the problem as regression onto continuous color space or classification of quantized color values [8], [29], [35]–[37]. On the other hand, traditional example-based methods were characterized as non-parametric methods transferring the color from the source data onto the analogous regions of the target image [2], [25], [26], [28], [34]. Regression methods use the Euclidean distance between the predicted and the ground truth color values as the loss function [6]. The perceptual uniformity of CIELAB is modeled by the Euclidean distance between the corresponding color points in color space. The Euclidean loss function is very appreciated because it is easy to understand and compute. However, believable results very different from the ground truth are identified as inadequate with the Euclidean loss function. Also, unfamiliar objects are assigned desaturated reddish and brownish color values. The recognition of the colorization multimodality remarkably affects the result [8], [30]. The need for modifying the loss function, in opposition to the widely accepted Euclidean distance loss is noticed by Dahl, in [35]. The cross-entropy loss is introduced for colorization by Zhang *et al.* [8], [9].

However, this simple categorization of deep learning methods became insufficient. With the emergence of deep generative models, which started producing vivid results with colors dissimilar to the ground truth, the need for a new way of method classification arose. Since 2017, deep generative models, generative adversarial networks (GANs) and variational autoencoders (VAEs), have come into focus. The main advantage of GANs is the fact that the competition between the two neural networks, the generator, and the discriminator, creates the corresponding loss function while producing vivid results [31]. Systems that avoid regression averaging of possible colors are created by modifications of ordinary architecture and loss function, while producing manifold of realistic results [36], [38].

Even more adapted loss functions are introduced in [5], [7]. The novel Color-UNet++ architecture [5] uses a linear combination of mean squared error (MSE), SSIM and PSNR as the loss function, which proved suitable for the regression problem. The method resolves the unnatural checkerboard artifacts of colorization through a careful deconvolution design. A combination of adversarial loss, cycle consistency loss and detail loss in CycleGAN [7] leads to higher authenticity of the result. Skip connections used in the method enable the improvement of feature representation.

Recently, the focus of researchers has diversified from colorization of images with natural scenes towards the application of colorization methods to other types of images, such as radar [40] and infrared images [41], comic books, cartoons, icons, fashion sketches and image synthesis from 2D and 3D models [11], [36]. New applications lead to the extension of colorization algorithms with additional techniques, e.g., language-based colorization of sketches [42], [43].

Recent extensive research [32] divides fast evolving deep learning techniques ranging from early brute-force neural networks to efficient GANs into seven categories: plain colorization neural networks, user-guided colorization neural networks, domain-specific colorization neural networks, text-based colorization neural networks, diverse colorization neural networks, multi-path colorization neural networks and exemplar-based colorization neural networks. We have found that there is overlap between some categories, so we propose modification of this classification. Our categorization consists of five categories and covers all deep learning colorization methods previously grouped into seven categories. The classification converged to five categories regarding neural network structure and user interaction as the basic criteria:

- plain colorization neural networks – simple feedforward CNNs,
- user-guided colorization neural networks – CNNs requiring user interaction,
- diverse colorization neural networks – deep generative models (GANs and VAEs) which generate results different from the ground truth,
- multi-path colorization neural networks – parallel paths of CNNs for analyzing different features,
- exemplar-based neural networks – CNNs which are given an example color image at the input along with the grayscale image.

The domain-specific colorization neural networks category from [32] is intended for colorization of specific type of images (e.g., radar images, infrared images). Regarding only the architecture used for the colorization process, they could be categorized into any of the five categories. The selected category depends on the neural network structure used for that specific problem.

The text-based colorization neural networks from [32] could be considered as user-guided colorization neural networks because user assistance in a form of providing textual input is needed in the colorization process. They were not analyzed in this paper as their successful application requires additional research on the non-image-processing-related problems such as text domain analysis, semantic ambiguity, and language bias.

Image colorization experiments were performed with methods representing each of the five categories, and the results were compared.

1) PLAIN COLORIZATION NEURAL NETWORKS

Plain colorization neural networks generally consist of a straightforward architecture: stacked layers only or naive skip connections [8], [29]. The most well-known fully automated colorization method proposed by Zhang *et al.* [8] belongs to this category. The network architecture shown in Fig. 3 consists of multiple convolutional blocks with two or three convolutional layers with rectified linear unit (ReLu) activation function. Down-sampling and up-sampling are conducted between the blocks. The method is based on the multinomial classification of pixels according to color and the class rebalancing for increasing diversity of resulting colors. The distribution of possible colors is predicted for each pixel. The classification of pixels is determined by probabilities of belonging to one of 313 segments of the discretized and quantized *ab*-plane of CIELAB color space. The major contribution of the method is the observation that in natural images the number of pixels with desaturated color values is orders of magnitude higher than the number of pixels with saturated values. Without taking this into account, the cross-entropy loss function is dominated by desaturated *ab*-values. Hence, the adjustment of the loss function is performed based on the distribution of saturated and desaturated pixels. The loss is reweighted during the training phase.

Although producing a broad range of convincing results, simple neural networks might experience difficulties with accurately capturing the color characteristics for many different scenes with distinct color styles [44]. The results often contain improper colors and noticeable artifacts [9]. An ensemble of neural networks is proposed to achieve better results than using one network alone.

Plain neural network architecture is also used for producing colorization results for images from different domains (e.g., radar) [40] and different modalities (e.g., infrared) [41].

2) USER-GUIDED COLORIZATION NEURAL NETWORKS

User-guided colorization neural networks require user involvement and are built upon deep learning foundations. Therefore, they could be categorized as both user-guided and deep learning methods simultaneously. Although the user contribution is relevant, the prevailing categorization includes them into deep learning methods because of prevailing usage of deep neural networks. The user involvement in a form of real-time or delayed distribution of sketches, strikes, points or scribbles [9], [37], or a textual phrase [42], [43] is required at the neural network input.

The first subcategory of user-guided colorization neural networks could be considered as an enhancement of the scribble-based methods because of the color hints usage. Point inputs upon the grayscale image are required, in contrast with the classic scribble-based approach which demands strokes of color. The architecture of the method from Zhang *et al.* [9] is shown in Fig 4. Two variants of the user interaction colorization neural networks are trained – the local hints and the global hints network. The local hints network processes user points and predicts the color distribution. The global hints network incorporates global statistics into the main framework. The hypercolumn approach from [30] is used by concatenating features from multiple layers of the main branch and learning a two-layer classifier on top. Necessary changes in resolution are conducted by down-sampling and up-sampling between convolutional layers. The colorization is performed in CIELAB color space. The main framework generates the result in a single feed-forward pass, enabling real-time usage. Before giving the user a chance of arbitrary marking the target image, a result of training with a million images with simulated user inputs is obtained.

The second subcategory of user-guided colorization neural networks uses textual input for colorization management. This colorization subcategory requires specific textual phrases from a limited dictionary, semantic analysis, and language processing. The association of language-based instructions and scene regions allows reusing the same instructions for consistent colorization of different images involving similar objects. The interactivity of this subcategory makes it appropriate for literacy education for children [43].

3) DIVERSE COLORIZATION NEURAL NETWORKS

Any architecture which generates more than one resulting image with colors not necessarily the same as in the original can be categorized as a diverse colorization neural

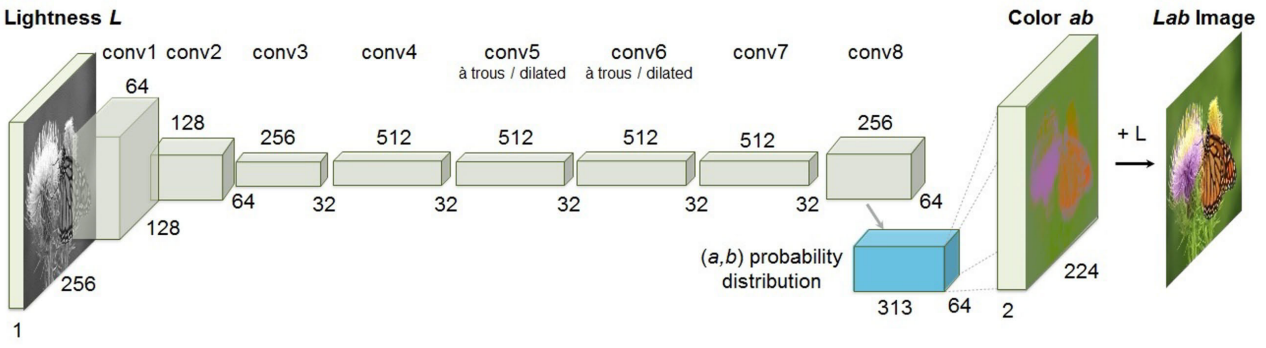


FIGURE 3. Architecture of Zhang *et al.* [8] plain colorization neural network.

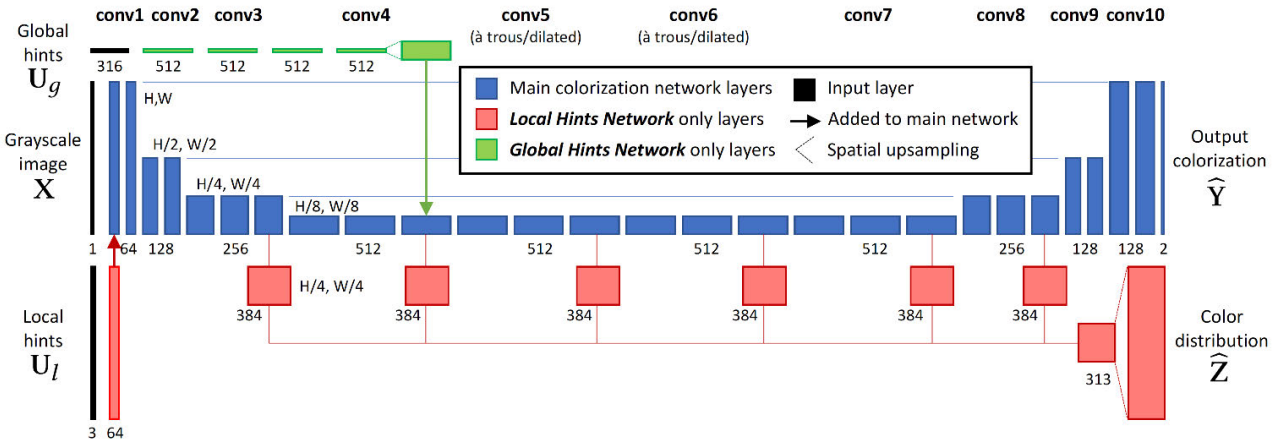


FIGURE 4. User-guided colorization neural network architecture from Zhang *et al.* [9].

network [11], [36], [38], [39]. The term ''diverse'' primarily refers to colorization results which differ from the ground truth and at the same time are characterized as realistic. Diversity of results is usually accomplished by GANs and VAEs. The GAN architecture designed by Vitoria *et al.* [11] is shown in Fig. 5. The two-part generator produces color information and classifies semantic content. The discriminator learns to differentiate between real and generated data. CIELAB color space is used for the colorization process. A combination of color, perceptual and semantic information leads to an innovative three-term loss function. Training the model with a fully self-supervised strategy (semantic clues coupled with an adversarial approach) yields to high quality, vivid results.

4) MULTI-PATH COLORIZATION NEURAL NETWORKS

Learning features from different paths is the main reason for classifying a method into the multi-path colorization neural networks category [6], [30], [33]. Colorization with capsule neural networks belongs to this category [33]. However, capsule networks have not been tested on large datasets like ImageNet [45], leading to doubt about their ability to perform better than existing approaches [46]. A frequently adverted CNN regression colorization method was developed by Iizuka *et al.* [6]. After comparing the results of using different color spaces (RGB, YUV, CIELAB) within the algorithm, the authors concluded that there was no significant difference in perceptual quality among color spaces.

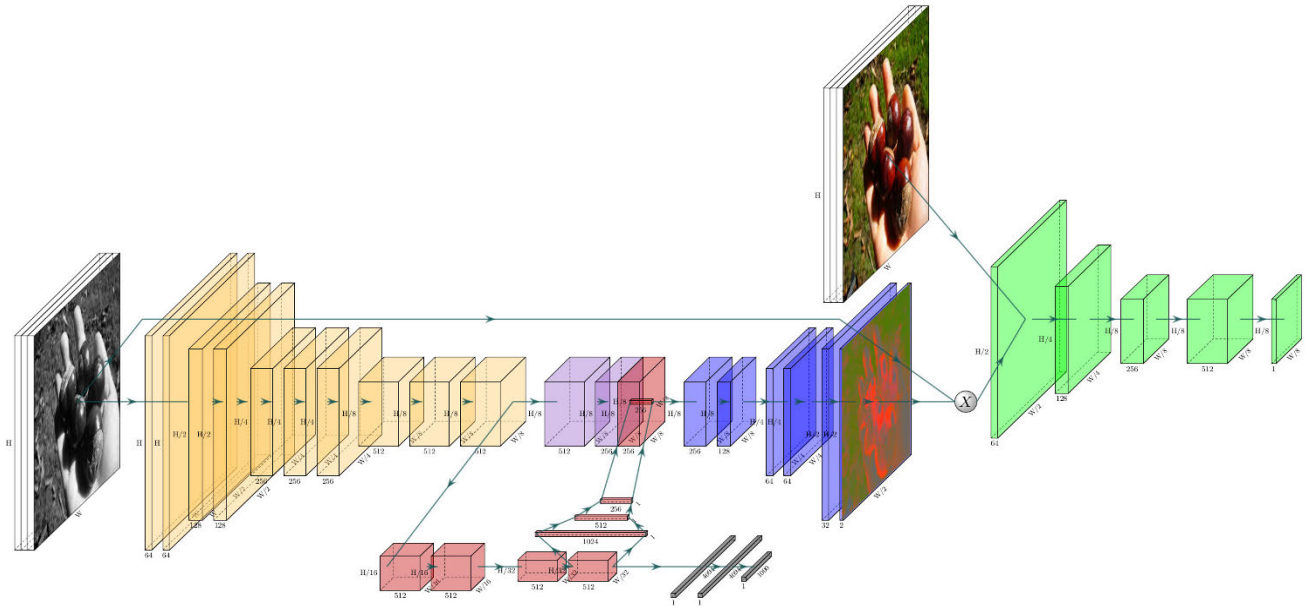


FIGURE 5. Architecture of diverse colorization neural network ChromaGAN by Vitoria *et al.* [11].

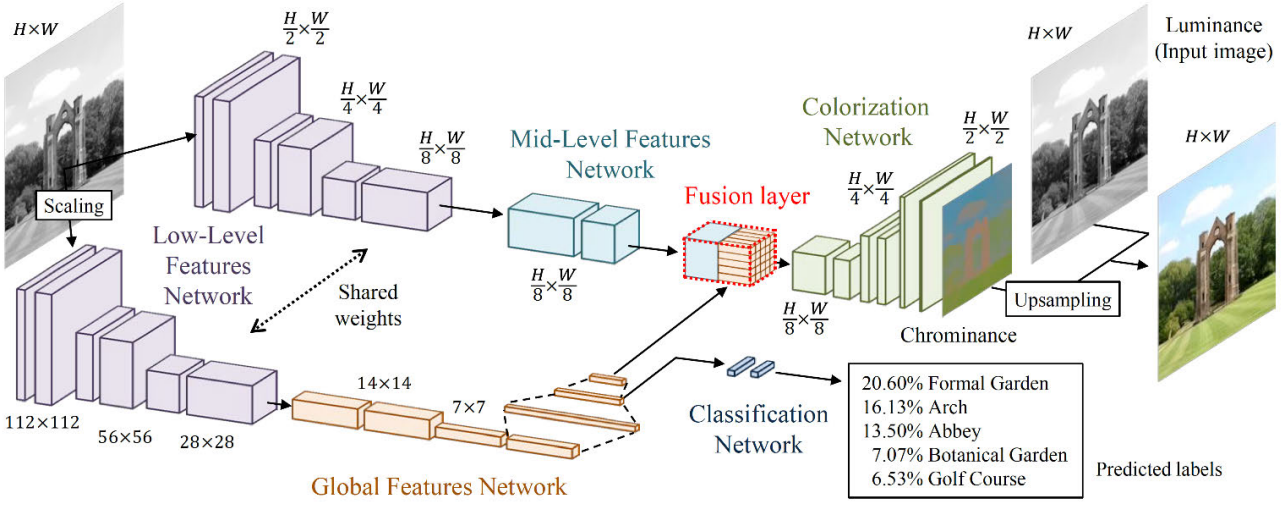


FIGURE 6. Architecture of multi-path colorization neural network from Iizuka *et al.* [6].

The importance of global features (scene semantics) is emphasized, leading to reduction of solution ambiguities. Nevertheless, local information is not diminished. The network model consists of four main components: a low-level features network, a mid-level features network, a global features network and a colorization network as shown in Fig. 6. Local and global information are fused together enabling total automaticity. The model is trained *end-to-end* on a large dataset for scene recognition with a joint colorization and classification loss. The classification of the scene into one of the predetermined categories significantly improves the result.

Previously mentioned method from Zhang *et al.* [9] can also work without user intervention. Before providing points of color, the colorization result deprived of user intervention is given at the output. Because of the division and later fusion of low-level user inputs with high-level semantic information, this automatic version of the method is classified in the multi-path category.

5) EXEMPLAR-BASED COLORIZATION NEURAL NETWORKS

Exemplar-based colorization neural networks could be considered as an extension of the example-based methods category. One or more reference images are used for transferring the color to the target image. In a new method for more successful colorization of images with multiple objects proposed by Su *et al.* [10], object detection is initially used, as shown in Fig. 7. CIELAB color space is used for colorization. In the

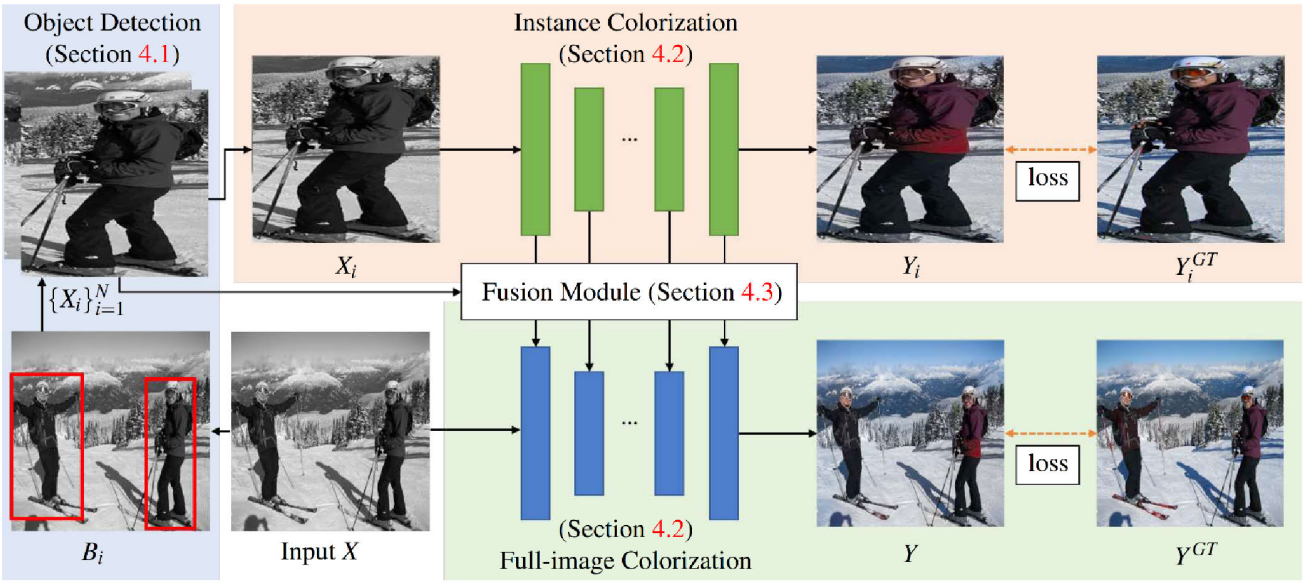


FIGURE 7. Architecture of exemplar-based colorization neural network from Su *et al.* [10].

absence of a clear figure-ground separation, when learning on an entire image, models cannot effectively locate and learn meaningful object-level semantics. Learning to colorize instances is a substantially easier task than learning on an entire image because it disregards complex background clutter. Every instance is forwarded to two different colorization neural networks. Fusion of the features is performed for the output. This way, better feature map is obtained leading to better results. The networks are initialized with the pretrained weights as in [9]. The stated fact is just one of the examples of the entanglement of the existing methods.

6) COMMONALITIES OF THE DEEP LEARNING METHODS

Although differing in many aspects regarding architecture design, construction of loss functions, variety of learning strategies, etc., all the main deep learning methods share a number of commonalities which are used for creating new, upgraded methods. Neural network topology inevitably consists of a different number of convolutional layers (blocks). A stack of convolutional layers with small kernels brings significant benefit in the accuracy compared to the singular convolutional layers with big kernels [41]. However, the number of layers and their organization are distinctive. Convolutional layers are frequently followed by batch normalization (for reducing the effect of the internal covariate shift (ICS) [33]) and a non-linear activation function. Striding or pooling layers are used (e.g., for dimensionality reduction). Dropout layers are used for preventing overfitting. These layers do not have parameters - they randomly set previous layer's activations to zero with a probability equal to dropout ratio hyperparameter. Softmax is often used as an activation function of the output of the neural network's layers. However, tanh is often used at the end of the neural network [47]. Common up-sampling techniques are used for increasing image size between the layers, e.g., bilinear [9] and nearest-neighbor [6]. Many methods use VGG-16 [11], [16], [30] or VGG-19 [33] architectures as feature extractors. Local and global features are often combined with a fusion layer [6], [10], [48]. Local features are based on small pixel areas while a global model considers the whole image for better understanding of the image content. Also, many methods find the Euclidean loss as inappropriate for the colorization task [8], [9], [11], [16], [26], [30]. Neither the Euclidean nor the cross-entropy loss are suitable to express the human subjective opinion about the acceptability or rationality of the predicted colors. The Huber loss function has been recently used because of lower sensitivity to outliers [49].

The datasets used for training the colorization methods vary substantially. They are usually intended for other image processing tasks: detection, classification, segmentation, etc. [32]. The most prominent publicly available datasets are CIFAR datasets [50], ImageNet ILSVRC2012 [51] and Places [52]. Most notable colorization methods [6], [8], [11], [33] have been trained on ImageNet ILSVRC2012. The existence of dissimilar datasets increases the amount of training material, but it also creates setbacks. For example, if using semantic labels, inconsistency in categories between different datasets leads to incompatibility [44].

Also, ResNet architecture with skip connections is frequent in colorization neural network construction [32]. Very deep CNNs are difficult to train because of vanishing and exploding gradients. Skip connections allow to take activation from one layer and feed it to another, much deeper layer. Colorization performance is superior if deeper CNNs are used [41].

Many authors agree that postprocessing is necessary to remove the incoherence and recover the lost details in an image [41]. Joint bilateral filtering is used in [29], [44].

III. EVALUATION OF COLORIZATION METHODS

A. SIGNIFICANCE OF TESTING COLORIZATION ALGORITHMS

Every algorithm has an expected behavior and a task to fulfill. Testing of image processing algorithms can be defined as a process of determining whether a particular algorithm has satisfied its specifications relating to certain criteria, such as accuracy, robustness, adaptability, sensitivity, reliability and efficiency [53]. More precisely, performance evaluation shows to what extent the behavior of an algorithm matches the required properties. The goals that should be achieved by some algorithm have to be well defined. Finding an appropriate metric which correlates with the subjective judgement is a big limitation in image processing. Visual patterns are high-dimensional and the apprehension of visual similarity (and quality) is often subjective [19]. Moreover, the performance of algorithm testing depends on various factors: the algorithm itself, the nature of images, the parameters and the metric used for evaluation. Because of the image processing algorithm diversity, the selection of a proper evaluation method is highly dependent on the task [53], [54].

Time complexity, computational complexity and the ways of hardware usage are indispensable indicators of colorization method divergence. Nevertheless, a direct comparison of certain methods is difficult to perform mainly because of the great diversity of the main ideas in problem-solving approaches. Although many methods share similarities in the approach, especially in deep learning category, some are trained with different datasets under different terms, mainly regarding training time. Also, finding the adequate manner of assessing the result of colorization remains an open issue.

The main goal of a colorization result is a convinced viewer. However, methodical subjective quality assessment demands for a group of volunteers who rate the visual quality of an image [54]. Viewers can be influenced by environmental conditions and mood oscillations, often making the results biased, far from objective. Also, the process is slow, expensive, and impractical [18], [23]. This is the main reason for turning to quantitative evaluation. Objective metrics have been designed to quantify image quality efficiently and automatically in correlation with subjective judgement [55].

However, there is no single quantitative metric which correlates completely with the image quality as perceived by the human visual system [16], [19], [53], i.e., the human perception of naturalness and detail [56].

Even though many characteristics of the human visual system are still unknown, it is acknowledged that it is much more sensitive to light intensity changes than to chrominance changes and more sensitive to contrast than to mean shift [18]. This is the main reason for the extensive use and development of the grayscale image quality metrics. Most color image quality metrics are modified grayscale image quality metrics [23]. Considering color information, visual state-of-the art quality metrics can be divided into three categories: grayscale, chrominance, and combined quality metrics [18]. In colorization, the original image and the image obtained by the process share identical luminance component. Consequently, the usage of chrominance quality metrics is necessary for the colorization evaluation.

B. METRICS USED FOR COLORIZATION EVALUATION

PSNR and SSIM are metrics which evaluate the difference between the original image and a changed one [17]. PSNR is measured in decibels. A higher PSNR value indicates a higher reconstruction quality. SSIM ranges from 0 to 1. The value of 1 indicates identity of the original and reconstructed image. Values above 0.97 are gratifying. It is important to emphasize that measuring PSNR and SSIM between color images in the process of colorization is unsuitable because they assess the luminance component only. Even though many colorization methods use CIELAB and YUV color spaces, the obtained results are transformed to RGB color space because RGB representation of images is a standard way to display colors on monitors and other devices. For this reason, PSNR and SSIM between the R, G and B components of the original and the colorized image can be used as a performance metric for colorization. Separate consideration of the R, G and B components does not provide a good insight into the overall quality of the color image. In calculation of Color PSNR (CPSNR), MSE values for R, G and B components are obtained. CPSNR is then calculated using the averaged MSE across components. Also, converting from RGB to YUV color space and measuring PSNR and SSIM of the U and V channels can be accomplished. YUV minimizes the correlation between the three coordinate axes (R, G and B) of the color space. For YUV color space the same considerations apply as for RGB color space. Consideration of the U and V channels provides a good insight into the overall quality of the color image. Overall, PSNR and SSIM are simple functions that cannot take into account fine distinctions in human perception [19].

Various metrics for the comparison of color images have been proposed over the years. Some of them are comparative and demand a reference image, while others require no reference. Most of these metrics are related to the difference between colors. However, no metric for measuring color plausibility currently exists. Nevertheless, in the examples in which color is fixed by natural laws (blue sky, green grass, etc.), numerical evaluation of color difference may have a reasonable value.

QSSIM [18] is a metric developed because it was noticed that the simple approach of expanding the grayscale quality metric into chrominance quality metric through linear combination of the separate results of each color channel (the dot-product approach) was insufficient. When performing a dot product between two vectors, only a part of the energy difference between the vectors is measured. A vector correlation is composed of two parts, the scalar correlation and the cross correlation (vector product), which form the color correlation. Quaternion image processing treats each color pixel as a single quaternion number. Quaternions are a generalization of complex numbers. A quaternion $q \in \mathbb{H}$ is composed of a real part and three imaginary parts:

$$q = s + r \cdot i + g \cdot j + b \cdot k, \tag{1}$$

where $s, r, g, b \in \mathbb{R}$ and i, j and k are its basic elements. The following rule has to be considered:

$$i^2 = j^2 = k^2 = ijk = -1. \tag{2}$$

Important definitions cover the quaternion conjugate:

$$\bar{q} = s - ri - gj - bk, \tag{3}$$

and the quaternion modulus:

$$|q| = \sqrt{s^2 + r^2 + g^2 + b^2} = \sqrt{qq^*}. \tag{4}$$

The parameter s is considered the quaternion scalar part. When s is zero, the quaternion is called a pure quaternion. The three RGB channels of the color image are encoded in the three imaginary parts of the quaternion [18], [57] thereby forming pure quaternions. QSSIM is not the only quaternion color image quality assessment metric developed. It was noticed that a quaternion matrix has a unique singular value feature vector [57]. The matrix is a representation of a color image; therefore, the singular value feature vector is unique for every color image. The similarity of two vectors can be measured by the angle between them.

S-CIELAB is a spatial extension of the CIELAB color metric useful for measuring color reproduction errors in digital images [58]. To compute the error, digital color images are spatially filtered (to simulate the spatial blur of the human visual system) and then converted to the CIELAB representation. The standard CIELAB delta E metric is suitable for use on large uniform color targets, but not on images, because color sensitivity changes as a function of spatial pattern. The sensitivity to color differences also depends on the color of the background or the adaptation state of the eye, which can be changed by ambient illumination [58], [59]. However, in colorization many convincing solutions can be created even with colors very different from the ground truth, whereas this metric is a color fidelity metric.

PCQI [20] is a metric developed for evaluating the quality of images with contrast modifications considering a reference image. The metric is a product of three independent

components: mean intensity, signal strength and signal structure. The three components uniquely describe image patches. While the distortions in contrast are observed in mean intensity and signal strength values, structure can always be determined regardless of the imperfect contrast. In [20], it is demonstrated that the PCQI metric is well correlated with subjective evaluation of image quality. Anwar *et al.* [32] show that this metric can be used for the evaluation of colorized images.

When designing a color quality metric, it is believed that two main factors need to be considered: color cast and colorfulness [21]. Colorfulness metric evaluates the perceptual impact of processing on image quality. The resulting image is considered different from the original, but not necessarily worse. The quality of the resulting image is measured, not fidelity to the original [21] since no reference image is needed. Because the main goal of colorization is a persuaded viewer and not fidelity to the original, the goal of this metric comes closest to the colorization goal. For computing colorfulness, a study of the distribution of image pixels in the CIELAB color space is conducted, indicating that colorfulness can be represented as a combination of image statistics. More precisely, it can be defined as a linear combination of the standard deviation σ and the mean value μ of the opponent blue-yellow and red-green color spaces. For the sake of computational efficiency, a simple version of the opponent color space is employed:

$$rg = R - G, \tag{5}$$

$$yb = \frac{1}{2}(R + G) - B. \tag{6}$$

Colorfulness metric can then be defined as:

$$\hat{M} = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb}, \tag{7}$$

$$\sigma_{rgyb} := \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \tag{8}$$

$$\mu_{rgyb} := \sqrt{\mu_{rg}^2 + \mu_{yb}^2}. \tag{9}$$

UIQM [22] is primarily constructed for assessing several aspects of underwater image degradation. It does not require a reference image. UIQM is a linear combination of three components: underwater image colorfulness measure (UICM), underwater image sharpness measure (UISM) and underwater image contrast measure (UIConM) that are inspired by the properties of human visual system. UICM measures the color cast. UISM specifies the clarity of edges and details. UIConM evaluates the contrast of underwater images. In [32], it is shown that UIQM can be used to evaluate colorization results of natural images.

UCIQE is a linear combination of CIELAB chroma, saturation and contrast [23]. It quantifies the non-uniform color cast, blur and low contrast of underwater images primarily. The tests conducted in the original paper verify the coherence between the quality of the results and the subjective perspective.

LPIPS is another metric used for the evaluation of the colorization result [19]. The features of the VGG neural network trained on ImageNet [45] have been revealed as an encouraging training loss function for image synthesis. Neural networks trained to solve challenging visual prediction and modeling tasks end up learning features that correlate well with perceptual judgments. The study of the metric has been conducted on the dataset containing many distortion types and real algorithm outputs, including colorization. Colorization methods generally do not show much structural variation but are prone to the effects of color bleeding and color variation.

C. RESULTS AND DISCUSSION

Visual and quantitative comparison of several state-of-the-art colorization methods with various architectures and levels of user assistance is described in the following paragraphs. The tested algorithms include user-guided scribble-based method from Levin *et al.* [3] and various deep learning methods: multi-path colorization neural network from Iizuka *et al.* [6], plain colorization neural network from Zhang *et al.* [8], exemplar-based colorization neural network from Su *et al.* [10], diverse colorization neural network from Vitoria *et al.* [11] and Zhang *et al.* [9] (both the interactive version belonging to user-guided colorization neural network and the automatic version belonging to the multi-path colorization neural network). The experiments were conducted with the code obtained from the webpages related to the papers. All deep learning methods were trained on ImageNet dataset [45]. For method comparison, several photographs from the author's collection were used. These photographs represent natural images with various scenes. The images were resized from the original resolution of 4160 × 3120 pixels to 320 × 240 pixels to reduce the processing time. Five images were carefully selected for evaluation because they have different features considering color, which is verified by the vectorscope analysis in Fig. 8. The vectorscope shows range of colors in each image. Vectorscope screens originate from the influence of the U and V components of YUV representation of the chosen images. Each dot in the vectorscope gives information about chrominance. This information contains two components: the hue (the specific color) and the saturation (the strength of the hue) of the colors in the test image. The associated vectorscopes were generated with ImageVectorscopeAnalyzer [60]. The colorization results are highly dependent on the image content.

Another reason for choosing these test images was the fact that the scenes that they represent are very distinctive – nature in daylight, differently colored typical objects, buildings, differently colored atypical objects at night and human faces. The majority of the distinctive colorization effects were described with the selected examples. A larger number of test images and the averaging of the metrics' results would not contribute to the quality of the evaluation presented in this paper and to the improvement of the comparison between

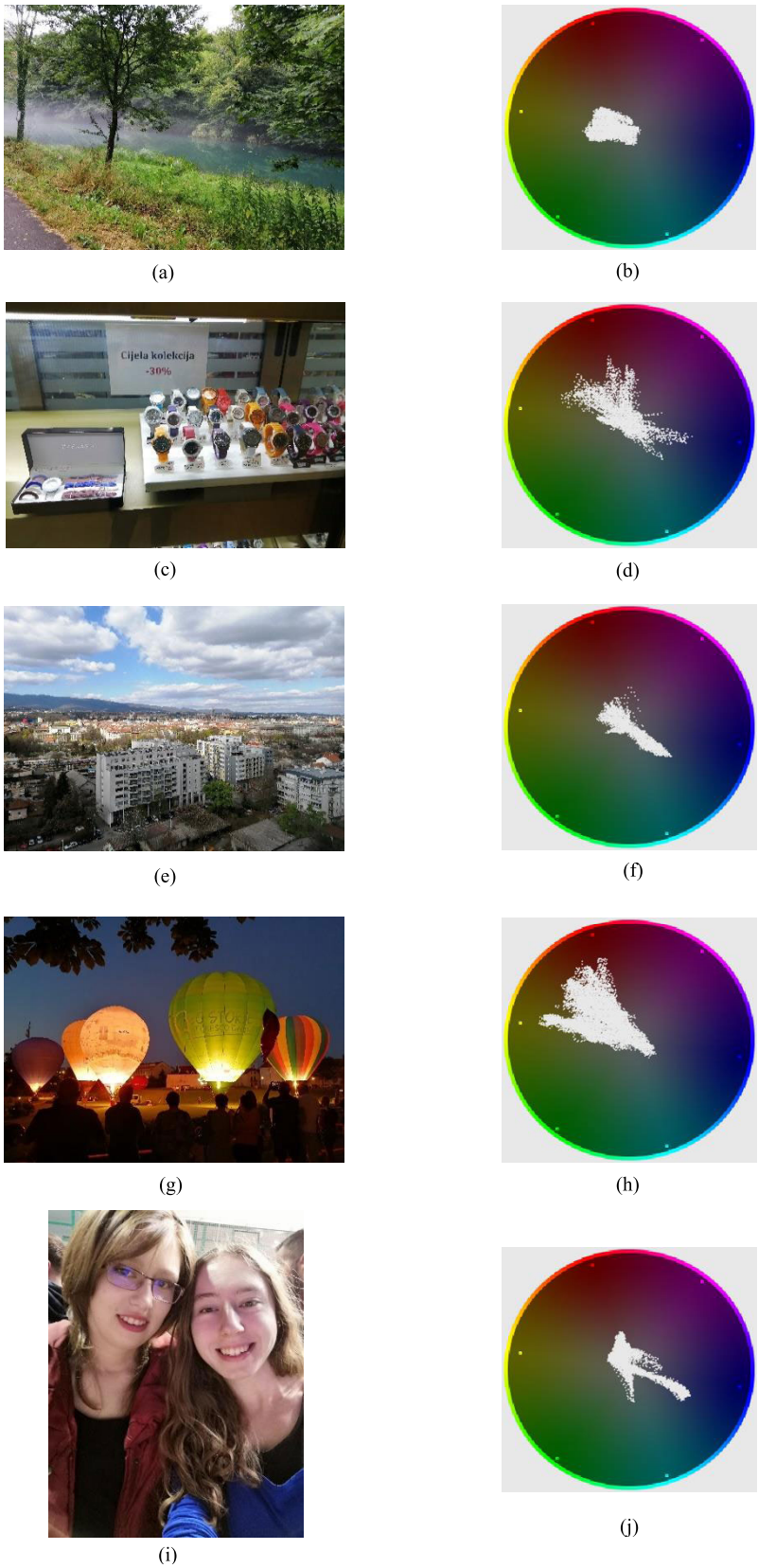


FIGURE 8. Test photograph: a) the river, b) vectorscope of river photograph, c) the watches, d) vectorscope of watches photograph, e) the buildings, f) vectorscope of buildings photograph, g) the balloons, h) vectorscope of balloons photograph, i) the people, j) vectorscope of people photograph.

the methods. Analysis on large quantity of images would not give the real insight of how the colorization methods affect the particular images with different content and color ranges. Also, it is impractical to colorize a large number of images with methods that require user intervention. Each of the selected images could be considered as a representative of a different group of natural images.

The original photographs, their grayscale versions and the colorization method comparison are shown in Figs. 9-13. In Table 1 to Table 10 the best result according to each metric is written in bold.

In Fig. 9, the river example, colorized results from the interactive version of Zhang *et al.* [9], Su *et al.* [10], Vitoria *et al.* [11] and the automatic version of Zhang *et al.* [9] can be visually estimated as highly credible, even more than the original. Also, the plain colorization neural network of Zhang *et al.* [8] gives satisfying result. Table 1 shows that mean QSSIM (MQSSIM) values do not correlate with this observation. The highest values, which imply better quality, are assigned to the automatic Zhang *et al.* [9], and Iizuka *et al.* [6]. The latter is the least convincing along with the result of the Levin *et al.* method [3], obtained from an unskillfully marked grayscale image. However, the MQSSIM results are all placed close together except for the results of Su *et al.* [10].

It is accounted to the fact that the sharpness of these images is reduced, therefore producing color bleeding. Similar observations can be noticed with mean SSIM values. A higher value of colorfulness indicates more visual appeal. Colorfulness of the result of the interactive version of Zhang *et al.* [9] is higher than of the original image. However, high value of colorfulness of the result from Levin *et al.* [3] does not indicate visual appeal. The UCIQE values mainly follow the perceptual observations. A larger UCIQE value indicates better

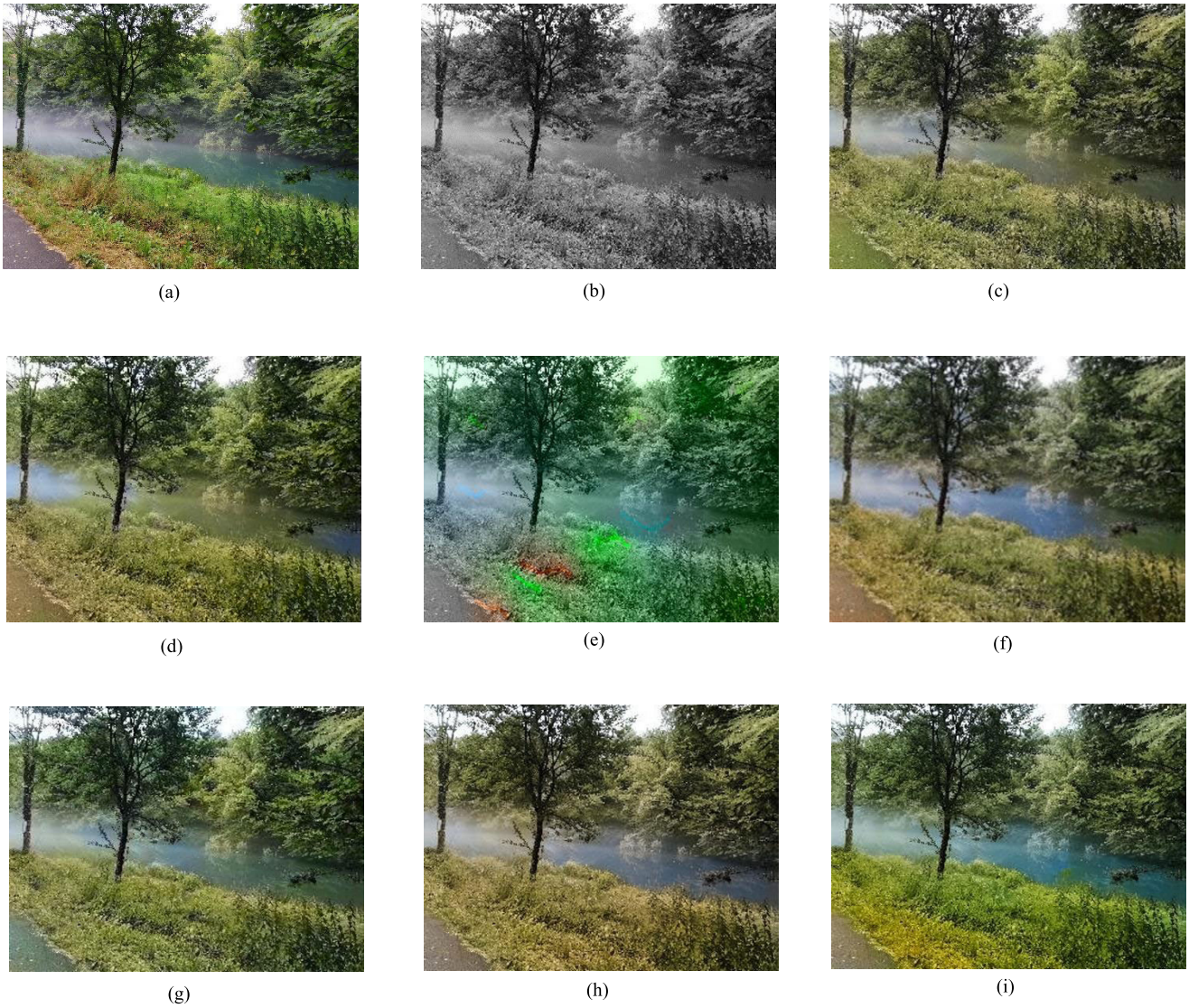


FIGURE 9. a) Original river photograph, b) grayscale version of the river photograph, c) result of Iizuka *et al.* [6], d) result of Zhang *et al.* [8], e) result of Levin *et al.* [3], f) result of Su *et al.* [10], g) result of Vitoria *et al.* [11], h) result of automatic Zhang *et al.* [9], i) result of interactive Zhang *et al.* [9].

TABLE 1. Image quality evaluation metrics for the river photograph.

Colorization method	MQSSIM	Colorfulness (orig. 36.014)	UCIQE (orig. 0.619)	LPIPS	CPSNR [dB]	SSIM (mean)	UIQM (orig. 0.824)	PCQI
Iizuka *et al.* [6]	**0.984**	27.784	0.602	0.154	25.952	0.970	0.800	**0.982**
Zhang *et al.* [8]	0.979	31.570	0.615	0.130	25.674	0.966	0.794	0.970
Levin *et al.* [3]	0.970	37.274	0.574	0.291	21.498	0.929	0.784	0.973
Su *et al.* [10]	0.712	30.096	0.610	0.421	20.723	0.703	0.723	0.466
Vitoria *et al.* [11]	0.981	30.566	0.615	**0.118**	**27.268**	0.973	0.807	0.970
Zhang *et al.* (auto) [9]	**0.984**	32.170	0.624	0.145	27.040	**0.974**	0.803	0.975
Zhang *et al.* (interactive) [9]	0.977	**46.951**	**0.644**	0.126	24.577	0.951	**0.841**	0.961

TABLE 2. PSNR and SSIM of the U and V components of the YUV river photograph.

Colorization method	PSNR_U [dB]	PSNR_V [dB]	UV_PSNR [dB]	SSIM_U	SSIM_V	UV_SSIM
Iizuka *et al.* [6]	29.167	29.065	29.116	0.880	0.932	**0.906**
Zhang *et al.* [8]	29.061	30.485	29.773	0.855	0.927	0.891
Levin *et al.* [3]	26.825	21.117	23.971	0.833	0.831	0.832
Su *et al.* [10]	29.807	30.055	29.931	0.871	0.927	0.899
Vitoria *et al.* [11]	31.593	30.111	**30.852**	0.883	0.924	0.904
Zhang *et al.* (auto) [9]	32.106	27.640	29.873	0.884	0.921	0.903
Zhang *et al.* (interactive) [9]	28.716	31.044	29.880	0.841	0.909	0.875

quality. LPIPS values also coincide with human perception (smaller values suggest better quality), except for Su *et al.* results [10]. Higher UIQM values indicate better image quality. The UIQM values give an advantage to the interactive Zhang *et al.* [9] and Levin *et al.* [3], just like colorfulness. It is because one of the components of UIQM is colorfulness measure, UICM. PCQI prioritizes the automatic Zhang *et al.* [9] method with all test images except for the river example, where the highest value is obtained by the Iizuka *et al.* [6] result. Table 2 shows the PSNR and SSIM values of the U and V components and their averaged UV_PSNR and UV_SSIM values for the river image. PSNR and SSIM of the Y component have not been evaluated because the original and the colorized image share that component. The UV_PSNR values of colorization results of all methods are located within a narrow range except for the result of Levin *et al.* [3], which is the least convincing. The highest UV_SSIM values are acquired with the results of Iizuka *et al.* [6], Vitoria *et al.* [11] and the automatic version of Zhang *et al.* [9]. The lowest value is acquired with the result of Levin *et al.* [3] and that corresponds to the human perception. The evaluated metrics have declared the last three colorized images as the most appealing: Vitoria *et al.* [11] with Zhang *et al.* [9] (both automatic and interactive) together with subjectively less appealing Iizuka *et al.* [6].

In Fig. 10, an example of colorful watches is shown. The most convincing results are generated with the interactive Zhang *et al.* [9] and Levin *et al.* [3] methods because ambiguities can be skillfully resolved by user suggestions. Other deep learning methods indicate desaturation. In this situation, Table 3 shows that MQSSIM values are not distinctive. Although colorfulness is the highest for the mentioned convincing results, many desaturated results also have high values. CPSNR and mean SSIM categorize the desaturated results of Iizuka *et al.* [6] and automatic Zhang *et al.* [9] as the best proving themselves unfitting for the task. PCQI inappropriately gives advantage to the results of automatic Zhang *et al.* [9] and Zhang *et al.* [8]. LPIPS coincides with the subjective observations. The UCIQE and UIQM metrics consider the user-guided methods better than the original, possibly indicating the fitness of the metrics for evaluating credibility of abstract scenes. In Table 4, UV_PSNR and UV_SSIM appoint the result of Iizuka *et al.* [6] as the best. Closely behind stands the automatic result of Zhang *et al.* [9].

In Fig. 11, the buildings obtained by the automatic Zhang *et al.* [9], Su *et al.* [10], Zhang *et al.* [8] and Iizuka *et al.* [6] methods are considered the most believable. Table 5 and Table 6 show that CPSNR, PCQI, UV_PSNR and UV_SSIM mostly match the observations.

In the example of balloons, Fig. 12, the result obtained by the interactive Zhang *et al.* method [9] appears more authentic than the original. UIQM and colorfulness from Table 7 confirm it. However, the result from Zhang *et al.* [8] is also labeled as more colorful than the original and that is not the case. Color stains make this image unconvincing. The results from Levin *et al.* [3] and Vitoria *et al.* [11] seem more convincing than the others. The biggest problem in the evaluation is the fundamental uncertainty of the possible balloons' colors. The results of the metrics in Table 7 and Table 8 reveal the automatic Zhang *et al.* [9] as the best. This fact does not overlap with the user perception because the colors of the balloons are not vivid.

In Fig. 13, many artifacts can be noticed. In the result of Levin *et al.* [3], the insufficient number of cautiously located scribbles makes the result utterly unbelievable. However, colorfulness, UIQM and UCIQE indicate the dominance of this method because of the overly saturated colors. The irregularities in the coloring of human faces can be seen in the result of Vitoria *et al.* [11], i.e., the skin tone is overly saturated. The result of Su *et al.* [10] wrongly assigns gray tone to the hair of a person. The results of Zhang *et al.* [8], [9] and Iizuka *et al.* [6] appear acceptable to the human eye. The averaging of all possible colors of clothes can be noticed in all

FIGURE 10. a) Original watches photograph, b) grayscale version of the watches photograph, c) result of Iizuka *et al.* [6], d) result of Zhang *et al.* [8], e) result of Levin *et al.* [3], f) result of Su *et al.* [10], g) result of Vitoria *et al.* [11], h) result of automatic Zhang *et al.* [9], i) result of interactive Zhang *et al.* [9].

TABLE 3. Image quality evaluation metrics for the watches photograph.

Colorization method	MQSSIM	Colorfulness (orig. 25.504)	UCIQE (orig. 0.617)	LPIPS	CPSNR [dB]	SSIM (mean)	UIQM (orig. 0.724)	PCQI
Iizuka *et al.* [6]	**0.974**	12.784	0.574	0.175	**26.020**	0.945	0.695	0.960
Zhang *et al.* [8]	0.952	39.258	0.632	0.234	20.819	0.903	0.746	0.961
Levin *et al.* [3]	0.956	39.307	**0.661**	0.177	22.791	0.913	**0.808**	0.954
Su *et al.* [10]	0.905	31.750	0.620	0.257	21.308	0.877	0.623	0.795
Vitoria *et al.* [11]	0.961	29.190	0.607	0.211	22.474	0.931	0.704	0.959
Zhang *et al.* (auto) [9]	**0.974**	17.115	0.583	0.171	25.523	**0.953**	0.690	**0.962**
Zhang *et al.* (interactive) [9]	0.960	**43.719**	0.641	**0.169**	22.361	0.920	0.757	0.958

the examples that do not require the user intervention. Despite that, images with appropriately colored human characteristics (skin and hair) are considered satisfactory. Table 9 shows that LPIPS and CPSNR consider the user-manipulated result from [9] as the best, which matches the visual impression. Moreover, UV_PSNR from Table 10 confirms that inference.

The time complexity of the core of the colorization process of every automatic method was measured. The configuration used for colorization process was a PC based on AMD Ryzen 7 4800H processor and NVIDIA GeForce GTX 1650 GPU. The numbers in Table 11 represent seconds. More precisely, the beginning of the measurement is set when the preprocessed image is forwarded to the model and the end is set just after the colorization is done. The postprocessing step is not included in the measurement. The time spent on colorizing images with user-guided methods,

TABLE 4. PSNR and SSIM of the U and V components of the YUV watches photograph.

Colorization method	PSNR_U [dB]	PSNR_V [dB]	UV_PSNR [dB]	SSIM_U	SSIM_V	UV_SSIM
Iizuka *et al.* [6]	29.131	28.477	**28.804**	0.898	0.883	**0.891**
Zhang *et al.* [8]	24.597	21.851	23.224	0.820	0.759	0.790
Levin *et al.* [3]	27.549	22.649	25.099	0.875	0.827	0.851
Su *et al.* [10]	25.846	23.548	24.697	0.863	0.825	0.844
Vitoria *et al.* [11]	26.804	22.754	24.779	0.863	0.780	0.822
Zhang *et al.* (auto) [9]	28.867	27.289	28.078	0.888	0.870	0.879
Zhang *et al.* (interactive) [9]	25.378	25.280	25.329	0.838	0.844	0.841

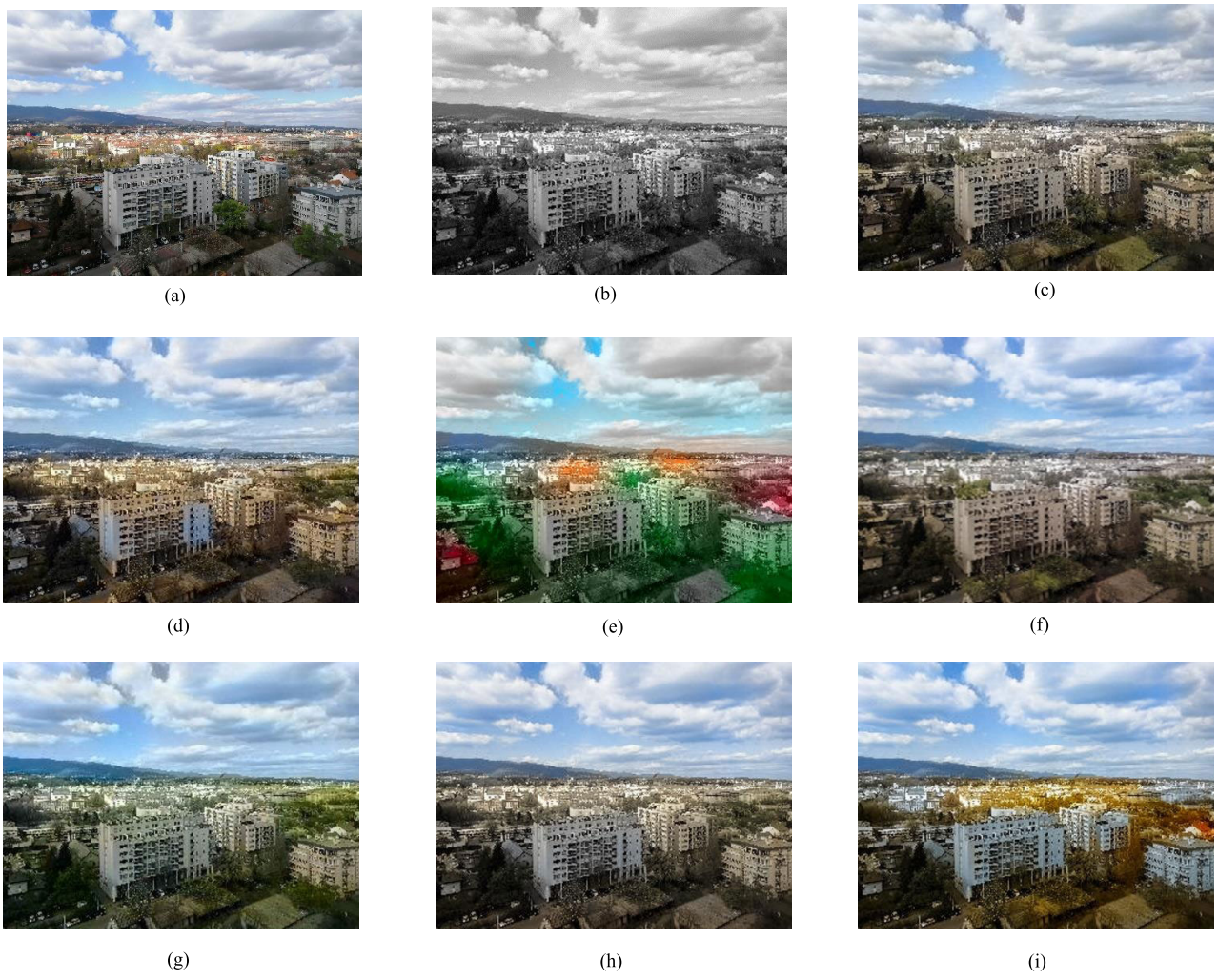


FIGURE 11. a) Original buildings photograph, b) grayscale version of the buildings photograph, c) result of Iizuka *et al.* [6], d) result of Zhang *et al.* [8], e) result of Levin *et al.* [3], f) result of Su *et al.* [10], g) result of Vitoria *et al.* [11], h) result of automatic Zhang *et al.* [9], i) result of interactive Zhang *et al.* [9].

TABLE 5. Image quality evaluation metrics for the buildings photograph.

Colorization method	MQSSIM	Colorfulness (orig. 26.155)	UCIQE (orig. 0.590)	LPIPS	CPSNR [dB]	SSIM (mean)	UIQM (orig. 0.959)	PCQI
Iizuka *et al.* [6]	0.978	20.591	0.594	0.115	29.466	0.970	0.964	0.974
Zhang *et al.* [8]	0.977	29.301	0.625	0.118	27.462	0.964	1.000	0.975
Levin *et al.* [3]	0.959	**40.952**	0.596	0.229	23.047	0.909	0.986	0.961
Su *et al.* [10]	0.885	30.721	0.613	0.242	24.03	0.873	0.791	0.756
Vitoria *et al.* [11]	0.970	33.266	0.615	0.130	26.44	0.954	1.010	0.968
Zhang *et al.* (auto) [9]	**0.979**	25.939	0.605	**0.102**	**29.977**	**0.971**	0.983	**0.978**
Zhang *et al.* (interactive) [9]	0.960	38.100	**0.636**	0.148	24.845	0.942	**1.021**	0.976

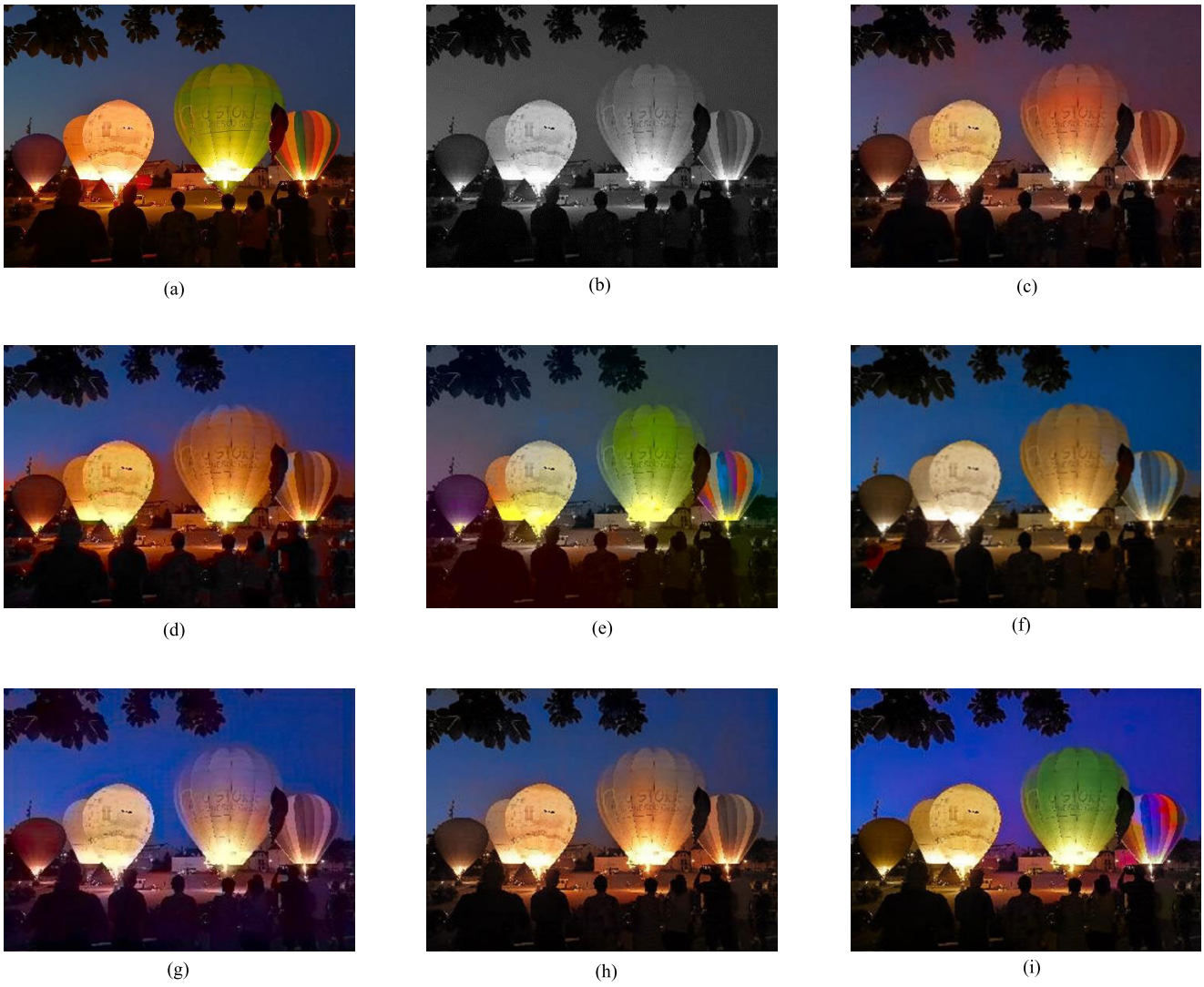


FIGURE 12. a) Original balloons photograph, b) grayscale version of the balloons photograph, c) result of Iizuka *et al.* [6], d) result of Zhang *et al.* [8], e) result of Levin *et al.* [3], f) result of Su *et al.* [10], g) result of Vitoria *et al.* [11], h) result of automatic Zhang *et al.* [9], i) result of interactive Zhang *et al.* [9].

TABLE 6. PSNR and SSIM of the U and V components of the YUV buildings photograph.

Colorization method	PSNR_U [dB]	PSNR_V [dB]	UV_PSNR [dB]	SSIM_U	SSIM_V	UV_SSIM
Iizuka *et al.* [6]	32.886	31.915	30.401	0.921	0.904	0.913
Zhang *et al.* [8]	31.016	29.226	30.121	0.903	0.879	0.891
Levin *et al.* [3]	28.909	22.478	25.694	0.901	0.795	0.848
Su *et al.* [10]	31.200	29.509	30.355	0.914	0.887	0.901
Vitoria *et al.* [11]	30.611	28.070	29.341	0.909	0.881	0.895
Zhang *et al.* (auto) [9]	33.719	32.106	**32.913**	0.933	0.907	**0.920**
Zhang *et al.* (interactive) [9]	27.827	27.977	27.902	0.867	0.856	0.862

TABLE 7. Image quality evaluation metrics for the balloons photograph.

Colorization method	MQSSIM	Colorfulness (orig. 66.572)	UCIQE (orig. 0.723)	LPIPS	CPSNR [dB]	SSIM (mean)	UIQM (orig. 0.957)	PCQI
Iizuka *et al.* [6]	0.924	48.206	0.679	0.212	21.456	0.829	0.866	0.945
Zhang *et al.* [8]	0.896	67.245	**0.739**	0.160	22.129	0.799	1.027	0.932
Levin *et al.* [3]	0.862	46.971	0.681	0.227	20.636	0.781	0.899	0.924
Su *et al.* [10]	0.887	55.073	0.702	0.204	19.752	0.754	0.830	0.797
Vitoria *et al.* [11]	0.834	60.792	0.684	0.271	17.257	0.701	1.023	0.883
Zhang *et al.* (auto) [9]	**0.925**	54.138	0.706	**0.150**	**22.588**	**0.843**	0.944	**0.947**
Zhang *et al.* (interactive) [9]	0.896	**77.363**	0.692	0.175	18.045	0.799	**1.045**	0.930

FIGURE 13. a) Original people photograph, b) grayscale version of the people photograph, c) result of Iizuka *et al.* [6], d) result of Zhang *et al.* [8], e) result of Levin *et al.* [3], f) result of Su *et al.* [10], g) result of Vitoria *et al.* [11], h) result of automatic Zhang *et al.* [9], i) result of interactive Zhang *et al.* [9].

TABLE 8. PSNR and SSIM of the U and V components of the YUV balloons photograph.

Colorization method	PSNR_U [dB]	PSNR_V [dB]	UV_PSNR [dB]	SSIM_U	SSIM_V	UV_SSIM
Iizuka *et al.* [6][10]	25.844	21.768	23.806	0.878	0.816	0.847
Zhang *et al.* [8]	25.649	23.627	24.638	0.864	0.785	0.825
Levin *et al.* [3]	24.510	21.543	23.027	0.816	0.801	0.809
Su *et al.* [10]	24.100	20.103	22.102	0.890	0.806	0.848
Vitoria *et al.* [11]	19.318	22.707	21.013	0.818	0.768	0.793
Zhang *et al.* (auto) [9]	25.482	24.956	**25.219**	0.915	0.854	**0.885**
Zhang *et al.* (interactive) [9]	19.808	23.014	21.411	0.878	0.833	0.856

TABLE 9. Image quality evaluation metrics for the people photograph.

Colorization method	MQSSIM	Colorfulness (orig. 36.046)	UCIQE (orig. 0.632)	LPIPS	CPSNR [dB]	SSIM (mean)	UIQM (orig. 0.677)	PCQI
Iizuka *et al.* [6]	**0.981**	36.112	0.628	0.100	24.653	0.959	0.689	0.986
Zhang *et al.* [8]	0.970	41.733	0.656	0.115	23.856	0.933	0.734	0.975
Levin *et al.* [3]	0.917	**64.067**	**0.681**	0.190	21.437	0.834	**0.752**	0.958
Su *et al.* [10]	0.965	35.065	0.634	0.141	23.440	0.924	0.664	0.919
Vitoria *et al.* [11]	0.969	48.109	0.658	0.115	22.426	0.925	0.726	0.972
Zhang *et al.* (auto) [9]	**0.981**	36.687	0.636	0.107	24.843	**0.958**	0.703	**0.995**
Zhang *et al.* (interactive) [9]	0.976	50.135	0.661	**0.091**	**25.934**	0.887	0.718	0.962

TABLE 10. PSNR and SSIM of the U and V components of the YUV people photograph.

Colorization method	PSNR_U [dB]	PSNR_V [dB]	UV_PSNR [dB]	SSIM_U	SSIM_V	UV_SSIM
Iizuka *et al.* [6]	27.581	26.813	27.197	0.941	0.898	**0.920**
Zhang *et al.* [8]	27.044	25.389	26.217	0.923	0.870	0.897
Levin *et al.* [3]	25.493	21.996	23.745	0.865	0.805	0.835
Su *et al.* [10]	27.052	24.676	25.864	0.932	0.884	0.908
Vitoria *et al.* [11]	25.558	23.981	24.770	0.922	0.872	0.897
Zhang *et al.* (auto) [9]	27.790	26.849	27.320	0.937	0.903	**0.920**
Zhang *et al.* (interactive) [9]	29.976	26.830	**28.403**	0.930	0.886	0.908

TABLE 11. Time needed for the colorization step of each method measured in seconds.

Colorization method	River	Watches	Buildings	Balloons	People
Iizuka *et al.* [6]	77.95	77.32	77.30	77.74	77.62
Zhang *et al.* [8]	6.71	6.69	6.63	6.47	6.65
Su *et al.* [10]	0.01	0.16	0.01	0.51	0.14
Vitoria *et al.* [11]	1.35	1.94	1.23	1.78	0.89
Zhang *et al.* (auto) [9]	0.28	0.24	0.29	0.25	0.52

Levin *et al.* [3] and Zhang *et al.* [9], was not measured because it cannot be done objectively. It is strictly dependent on the user's will – how much time someone is ready to dedicate to the task. Still, Zhang *et al.* [9] recommend the interaction of a few seconds to couple minutes.

In conclusion, the results in Table 11 indicate that hardware accelerators and architectural shortcuts of modern deep learning methods [9]–[11] have notably reduced the time complexity of the colorization process.

IV. CONCLUSION

The colorization of natural images is a challenging image processing and computer vision task. It is an ill-posed process with multimodal uncertainty. The colorization is successful if the viewer is persuaded in the naturalness of the colorized image. In this paper, the algorithms with different architectures and level of user guidance have been analyzed taking into account objective image quality metrics and time needed for colorization.

The user-guided colorization neural network from Zhang *et al.* [9] provides the most visually convincing results because of the successful combination of human effort and technology advances regarding neural networks. This fact was also confirmed by colorfulness, UIQM and UCIQE objective metrics. However, this colorization method requires extensive human intervention and vast amount of time. If user intervention cannot be applied or the processing time must be lower, the automatic version of Zhang *et al.* [9] shows superior results because its neural network architecture includes separate parts trained for the local features (color distribution) and the global features (semantic information).

More complex architectures, such as GANs in the diverse category [11], provide plausible results because of better adaptation to the colorization problem. The architecture uses more convenient loss function that takes into account color, perceptual and semantic information. The multi-path colorization neural network from Iizuka *et al.* [6] also indicates convincing results mainly because of extracting different levels of features. The plain colorization neural network from Zhang *et al.* [8] and the exemplar-based colorization neural network [10] show good results in colorization quality, but not as good as already discussed methods. The plain colorization neural networks do not capture the color characteristics of different scenes successfully because of their simplicity. The exemplar-based colorization neural network [10] requires the lowest time for colorization because providing reference images helps in color transfer to the target image.

Among the methods requiring the user intervention, the user-guided colorization neural network achieves better colorization than the scribble-based method. The use of neural networks improves the visual impression.

Multiple objective image quality metrics were applied, but none of them has demonstrated universality in the qualitative evaluation of colorization results. The objective image quality evaluation results frequently did not match the subjective impression. Colorfulness, UIQM and UCIQE metric often assign highest values to images with excessively saturated colors which appear attractive to the human eye. The values of LPIPS metric almost always coincided with human observations. UV_PSNR also assigned the highest values to visually acceptable results.

It is shown that there is a need for improvement of image quality metrics that would be able to better assess the colorization results according to the characteristics of human perception.

Existing deep learning colorization methods can be enhanced by careful selection of training parameters such as number of layers, number of epochs, and learning rate to achieve a balance between training time and complexity of the neural network structure. Further improvement of neural networks could lead to better understanding of features and context of images. Separate extraction of local and global features contributes to the improvement of the results. Choosing an appropriate loss function is a big challenge in colorization problem, because the common loss functions used in existing neural networks result in unsaturated colors. More adapted loss functions should take care of a combination of several different aspects that may affect the final result (e.g., color, detail, perceptual and semantic information). Future development of colorization systems will focus on creating even more imaginative architectures adapted to the colorization problem.

REFERENCES

[1] K. Nassau. *Colour*. Accessed: 2020. [Online]. Available: https://www.britannica.com/science/color

[2] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

[3] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.

[4] F. Bianconi, R. Bello, A. Fernández, and E. González, "On comparing colour spaces from a performance perspective: Application to automated classification of polished natural stones," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 71–78.

[5] Y. Di, X. Zhu, X. Jin, Q. Dou, W. Zhou, and Q. Duan, "Color-UNet++: A resolution for colorization of grayscale images using improved UNet++," *Multimedia Tools Appl.*, pp. 1–20, Mar. 2021, doi: 10.1007/s11042-021-10830-2.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[7] S. Huang, X. Jin, Q. Jiang, J. Li, S.-J. Lee, P. Wang, and S. Yao, "A fully-automatic image colorization scheme using improved CycleGAN with skip connections," *Multimedia Tools Appl.*, vol. 80, pp. 26465–26492, May 2021.

[8] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[9] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[10] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7965–7974.

[11] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2445–2454.

[12] A. Weltz. *The Basics of Hand-Coloring Black-and-White Prints*. Accessed: 2020. [Online]. Available: https://www.bhphotovideo.com/explora/photography/tips-and-solutions/the-basics-of-hand-coloring-black-and-white-prints

[13] G. Gene. (2009). *Scanning Around With Gene: The Miracle of Photochrom*. Accessed: 2020. [Online]. Available: https://creativepro.com/scanning-around-gene-miracle-photochrom/

[14] E. Trex. (2011). *How (and Why) Are Black and White Films Colorized?* Accessed: 2020. [Online]. Available: https://www.mentalfloss.com/article/26956/how-and-why-are-black-and-white-films-colorized

[15] S. Moraes. (2019). *Wonders in Image Processing With Machine Learning*. Accessed: 2020. [Online]. Available: https://medium.com/ODSC/wonders-in-image-processing-with-machine-learning-9c6f2e070e99

[16] D. Varga and T. Szirányi. (2017). *Convolutional Neural Networks for Automatic Image Colorization*. Accessed: 2020. [Online]. Available: http://eprints.sztaki.hu/9292/1/Varga_1_3306455_ny.pdf

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[18] A. Kolaman and O. Pecht, "Quaternion structural similarity: A new quality index for color images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1526–1536, Apr. 2012.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 568–595.

[20] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2387–2390, Dec. 2015.

[21] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Proc. SPIE*, Jun. 2003, pp. 87–95.

[22] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, Jul. 2016.

[23] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, Dec. 2015.

[24] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Proc. Eurographics Symp. Rendering*, Konstanz, Germany, 2005, pp. 201–210.

[25] X. Liu, L. Wan, Y. Qu, T. T. Wong, S. Lin, S. C. Leung, and P. A. Heng, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 152-1–152-9, 2008.

[26] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2008, pp. 126–139.

[27] 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 Trans. Graph.*, vol. 30, no. 6, pp. 1–8, Dec. 2011.

[28] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, Nara, Japan, 2012, pp. 369–378.

[29] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 415–423.

[30] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.

[31] S. Koo, "Automatic colorization with deep convolutional generative adversarial networks," Stanford Univ., Stanford, CA, USA, Tech. Rep. CS231n, 2016.

[32] S. Anwar, M. Tahir, C. Li, A. Mian, F. Shahbaz Khan, and A. W. Muzaffar, "Image colorization: A survey and dataset," 2020, *arXiv:2008.10774*. [Online]. Available: http://arxiv.org/abs/2008.10774

[33] G. Ozbulak, "Image colorization by capsule networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2150–2158.

[34] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 567–575.

[35] R. Dahl. (2016). *Automatic Colorization*. Accessed: 2020. [Online]. Available: https://tinyclouds.org/colorize/

[36] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2017, pp. 151–166.

[37] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8456–8465.

[38] A. Deshpande, J. Lu, M.-C. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2877–2885.

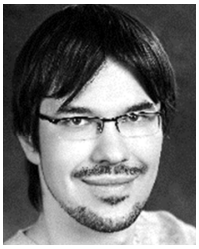
[39] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo, "Coloring with limited data: Few-shot colorization via memory augmented networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11275–11284.

[40] Q. Song, F. Xu, and Y.-Q. Jin, "Radar image colorization: Converting single-polarization to fully polarimetric using deep neural networks," *IEEE Access*, vol. 6, pp. 1647–1661, 2017.
[41] M. Limmer and H. P. A. Lensch, "Infrared colorization using deep convolutional neural networks," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, Dec. 2016, pp. 61–68.
[42] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis, "Learning to color from language," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2018, pp. 764–769.
[43] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 233:1–233:16, 2019.
[44] Z. Cheng, Q. Yang, and B. Sheng, "Colorization using neural network ensemble," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5491–5505, Nov. 2017.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[46] A. Misra. (2019). *Capsule Networks: The New Deep Learning Network*. Accessed: 2021. [Online]. Available: https://towardsdatascience.com/capsule-networks-the-new-deep-learning-network-bd917e6818e8
[47] E. Wallner. (2017). *Colorizing B&W Photos With Neural Networks*. Accessed: 2020. [Online]. Available: https://blog.floydhub.com/colorizing-b-w-photos-with-neural-networks/
[48] M. R. Joshi, L. Nkenyereye, G. P. Joshi, S. M. R. Islam, M. Abdullah-Al-Wadud, and S. Shrestha, "Auto-colorization of historical images using deep convolutional neural networks," *Mathematics*, vol. 8, no. 12, p. 2258, Dec. 2020.
[49] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 53, no. 1, pp. 73–101, 1964.
[50] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[52] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using Places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[53] M. Wirth, M. Fraschini, M. Masek, and M. Bruynooghe, "Performance evaluation in image processing," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–3, Apr. 2006, Art. no. 45742, doi: 10.1155/ASP/2006/45742.
[54] K. K. Gupta and R. P. Pareek, "A survey of image quality assessment techniques for medical imaging," *Indian J. Med. Informat.*, vol. 8, no. 2, pp. 72–74, 2014.
[55] Y. Shi, Y. Ding, R. Zhang, and J. Li, "An attempt to combine structure and color for assessing image quality," in *Proc. Int. Conf. Digit. Image Process.*, Mar. 2009, pp. 314–318.
[56] R. Maharaj and B. Naidoo, "An analysis of objective and human assessments in contrast enhancement," *Int. J. Appl. Eng. Res.*, vol. 13, no. 22, pp. 15843–15859, 2018.
[57] Y. Wang, W. Liu, and Y. Wang, "Color image quality assessment based on quaternion singular value decomposition," in *Proc. Congr. Image Signal Process.*, 2008, pp. 433–439.
[58] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color image reproduction," *J. Soc. Inf. Display*, vol. 5, no. 1, pp. 61–63, 1996.
[59] X. Zhang, D. A. Silverstein, J. E. Farrell, and B. A. Wandell, "Color image quality metric S-CIELAB and its application on halftone texture visibility," in *Proc. IEEE COMPCON Dig. Papers*, Feb. 1997, pp. 44–48.
[60] S. Laguerre. (2016). *Mathworks—Vectorscope Image Analyzer*. Accessed: 2021. [Online]. Available: https://ch.mathworks.com/matlabcentral/fileexchange/56546-vectorscope-image-analyzer
[61] I. Zeger and S. Grgic, "An overview of grayscale image colorization methods," in *Proc. Int. Symp. ELMAR*, Zadar, Croatia, Sep. 2020, pp. 109–112.
[62] T.-H. Sun, C.-H. Lai, S.-K. Wong, and Y.-S. Wang, "Adversarial colorization of icons based on contour and color conditions," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 683–691.
[63] Y. Shih, S. Paris, F. Durand, and W. T. Freeman, "Data-driven hallucination of different times of day from a single outdoor photo," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 200-1–200-11, Nov. 2013.
[64] S. Raj. (2014). *RGB to YUV Format*. Accessed: 2020. [Online]. Available: https://ch.mathworks.com/matlabcentral/fileexchange/47786-rgb-to-yuv-format
[65] V. Laparra, J. Muñoz-Marí, and J. Malo, "Divisive normalization image quality metric revisited," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 27, no. 4, pp. 852–864, Apr. 2010.
[66] A. Kolaman. *Quaternion Structural Similarity*. Accessed: 2021. [Online]. Available: http://www.ee.bgu.ac.il/~kolaman/QSSIM
[67] C. Yi Li, R. Mazzon, and A. Cavallaro, "Underwater image filtering: Methods, datasets and evaluation," 2020, *arXiv:2012.12258*. [Online]. Available: http://arxiv.org/abs/2012.12258

IVANA ŽEGER received the B.S. degree in electrical engineering and information technology and the M.S. degree in information and communication technology from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2018 and 2020, respectively, where she is currently pursuing the Ph.D. degree in electrical engineering.

In 2020, she was a Student Intern engaged in machine learning. Since 2020, she has been a Research and Teaching Assistant with the Department of Communication and Space Technologies, Faculty of Electrical Engineering and Computing, University of Zagreb. Her research interests include machine and deep learning technologies and their application in image processing and communications.

SONJA GRGIC (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 1989, 1992, and 1996, respectively. She is currently a Professor in multimedia technologies and communication systems with the Faculty of Electrical Engineering and Computing, University of Zagreb. Her research interests include image processing and machine learning, picture quality evaluation, video communication technologies, and image forensics.



JOSIP VUKOVIĆ (Member, IEEE) was born in Sisak, Croatia, in 1986. He received the B.S. degree in computing, the M.S. degree in information and communication technology, and the Ph.D. degree in electrical engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2009, 2011, and 2017, respectively.

From 2011 to 2017, he worked as a Research Assistant with the University of Zagreb. In 2017, he worked as a Researcher with Innovation Centre Nikola Tesla, Zagreb. Since 2017, he has been a Postdoctoral Researcher with the University of Zagreb. His research interests include GNSS, ionosphere, space weather, CubeSat technologies, remote sensing, and image processing.

GORDAN ŠIŠUL (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 1996, 2000, and 2004, respectively. He is currently employed as a Full Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb. His academic interests include wireless communications, signal processing applications in communications, modulation techniques, and radio propagation.