SCGAN: Saliency Map-Guided Colorization With Generative Adversarial Network

Yuzhi Zhao, *Graduate Student Member, IEEE*, Lai-Man Po, *Senior Member, IEEE*, Kwok-Wai Cheung, *Member, IEEE*, Wing-Yin Yu, and Yasar Abbas Ur Rehman, *Member, IEEE*

***Abstract*— Given a grayscale photograph, the colorization system estimates a visually plausible colorful image. Conventional methods often use semantics to colorize grayscale images. However, in these methods, only classification semantic information is embedded, resulting in semantic confusion and color bleeding in the final colorized image. To address these issues, we propose a fully automatic Saliency Map-guided Colorization with Generative Adversarial Network (SCGAN) framework. It jointly predicts the colorization and saliency map to minimize semantic confusion and color bleeding in the colorized image. Since the global features from pre-trained VGG-16-Gray network are embedded to the colorization encoder, the proposed SCGAN can be trained with much less data than state-of-the-art methods to achieve perceptually reasonable colorization. In addition, we propose a novel saliency map-based guidance method. Branches of the colorization decoder are used to predict the saliency map as a proxy target. Moreover, two hierarchical discriminators are utilized for the generated colorization and saliency map, respectively, in order to strengthen visual perception performance. The proposed system is evaluated on ImageNet validation set. Experimental results show that SCGAN can generate more reasonable colorized images than state-of-the-art techniques.**

***Index Terms*— Colorization, generative adversarial network, saliency map.**

I. Introduction

IMAGE colorization is the process of assigning plausible and perceptual colors to each pixel in the input image. It has found a wide array of applications in computer vision, such as multispectral image colorization [1], [2], image compression [3], cartoon colorization [4], [5], restoration of old photographs and films [6], fake colorization detection [7] and even assisting other tasks like classification and segmentation [8]. However, without prior information on the colors of the objects in the input intensity image, the colorization results may vary largely from system to system. Notably, the semantic confusion (which color should be assigned to each object in the image), color bleeding (spreading of colors beyond the object boundary), edge distortion, and object intervention are some key problems in the current automatic image colorization tasks.

There are multiple possible colors for an object in the image. Assigning a proper color to the object in an image is still an open research problem in multiple domains. In recent decades, a multitude of algorithms have been proposed to solve this problem. These algorithms can be divided into three possible categories: (1) Scribble-based methods [9]–[17], (2) example-based methods [18]–[30], and (3) fully-automatic methods [31]–[45]. The first two categories of algorithms require human interactions for assigning reasonable colors to various objects in the input-intensity image. As a result, these algorithms are highly correlated with the rationality of the human hints, which makes them labor-intensive and less robust to errors. For example, the scribble-based methods utilize the color hints, provided by the user, to assign different colors to the objects in the image. Similarly, the example-based methods require an additional color image to infer the chrominance intensity of different objects in the input image.

On the other hand, fully automatic approaches utilize end-to-end learning to directly learn the relationship between an input grayscale image and the corresponding color embeddings, without any human intervention. Most of these approaches utilize the deep Convolutional Neural Networks (CNN). Normally they are trained on large-scale datasets such as ImageNet [46] (1.3M images) and Places [47] (1.8M images) to encode the semantic information for image colorization. For instance, Larsson *et al.* [33] utilized hyper-column from a VGG-Net [48] pre-trained on ImageNet for semantic feature extraction. However, it requires high computational footprints which makes the inference slower during test time. Iizuka *et al.* [38] on the other hand jointly trained a classification sub-network and auto-encoder stream. It not only obtains semantic features but also establishes a reasonable scene context for colorization. Based on a VGG-Net backbone, Zhang *et al.* [37] introduced cross-channel encoding and class rebalancing techniques to generate unimodal distribution of color embeddings.

The automatic image colorization systems achieve better results. However, the problems of color bleeding and

Manuscript received April 16, 2020; revised August 11, 2020, October 3, 2020, and November 5, 2020; accepted November 8, 2020. Date of publication November 16, 2020; date of current version August 4, 2021. This work was supported by an Internal Funds Scheme from the City University of Hong Kong under Project 9678141. This article was recommended by Associate Editor H. Xiong. *(Corresponding author: Yuzhi Zhao.)*

Yuzhi Zhao, Lai-Man Po, and Wing-Yin Yu are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: yzzhao2-c@my.cityu.edu.hk; eelmpo@cityu.edu.hk; wingyinyu8-c@my.cityu.edu.hk).

Kwok-Wai Cheung is with the School of Communication, The Hang Seng University of Hong Kong, Hong Kong (e-mail: keithcheung@hsmc.edu.hk).

Yasar Abbas Ur Rehman is with TCL Corporate Research Hong Kong, Hong Kong (e-mail: yasar.abbas@my.cityu.edu.hk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2020.3037688.

Digital Object Identifier 10.1109/TCSVT.2020.3037688

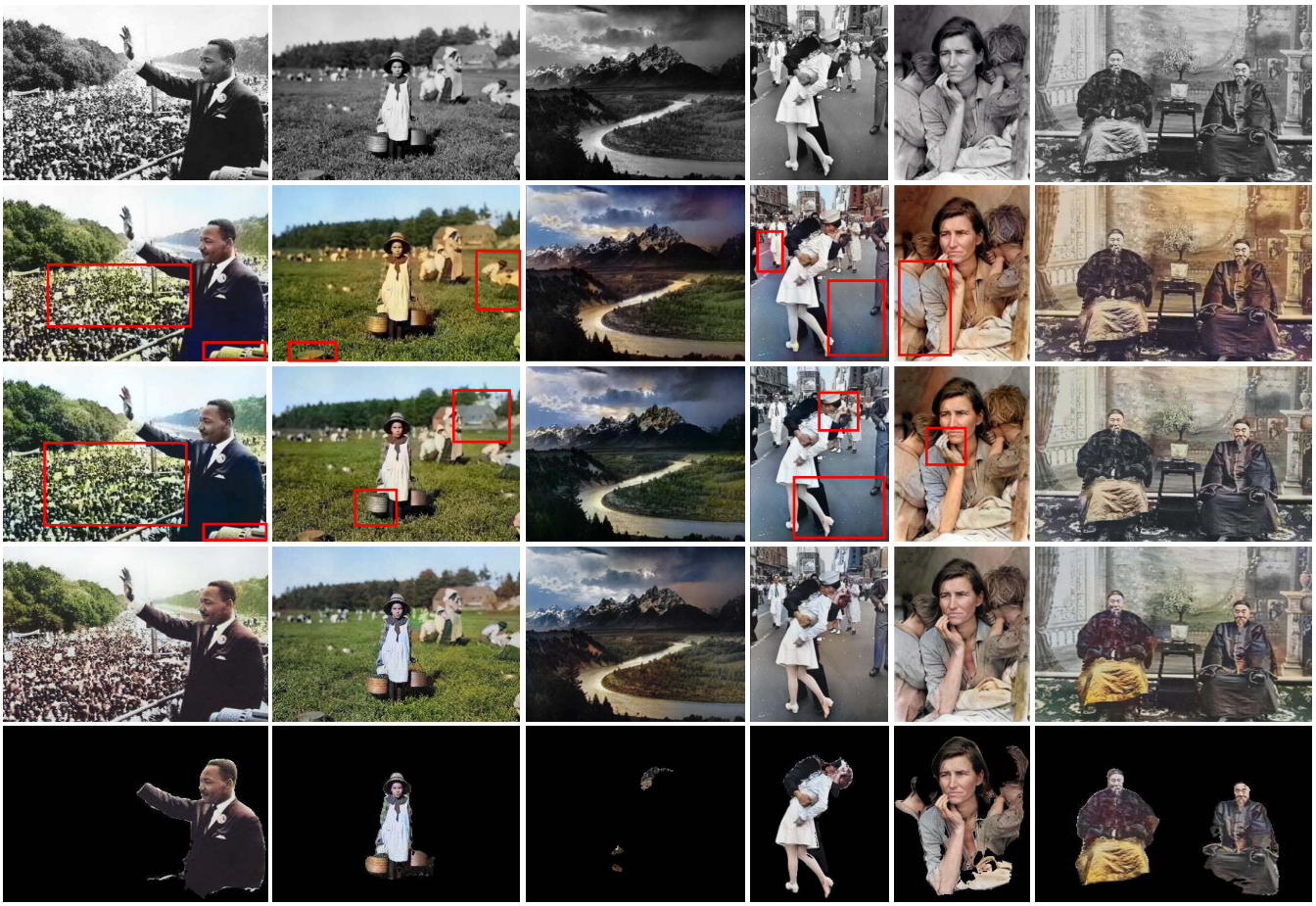

Fig. 1. Illustration of coloirzation results by [37], [44] and proposed method on old black and white photographs. The rows from top to bottom represent grayscale input, colorization results of [37], [44] and proposed method and saliency map generated by proposed method, respectively. The red rectangles highlight specific regions suffer from color bleeding or semantic confusion. Our model learns the different colorization representations in multiple scenes: speech, countryside, landscapes, city street, and human portraits. Photos were taken from the US National Archives (public domain). Please visit https://github.com/zhaoyuzhi/Semantic-Colorization-GAN (supplementary material) to see more colorization results.

unreasonable assignment of colors still exist. Figure 1 shows some common examples of the failure cases of [37] and [44] on some legacy photos. For instance, there is color bleeding in the first column by methods [37], [44] since color of trees spreads to crowd. Also, the roads and human faces are colorized in blue (in column 4 and 5, respectively) by methods [37], [44]. It leads to semantic confusion effect in output images. To address the problems, some regularization terms such as image gradients [31] and segmentations [34] have been added to the optimization process. However, these constrains are not useful for some situations. Since image gradients cannot represent semantics, it is hard for the colorization system to judge the colors for objects with similar boundaries, e.g. trees and crowd. In addition, only a few datasets include segmentation labels with limited categories.

Considering these limitations, we propose to use saliency map to improve the image colorization quality for following three aspects. Firstly, it identifies perceptually significant regions in the image. The colorization system can then be guided to focus more on the key objects while less influenced by the backgrounds. The key objects are richer in color while the backgrounds often contain green and blue colors, e.g. trees and sky. Moreover, it reduces the bias of the system to the colors that make up the majority of images. Secondly, it assists the network to localize objects at pixel level. It represents semantically salient areas with relatively clear boundaries. Thus, it is beneficial for colorization network to alleviate color bleeding artifact. Finally, since saliency map is adaptive to different objects in an image, it is convenient to be applied to multiple datasets in colorization area.

Specifically, we perform colorization and predict saliency map simultaneously by utilizing a Saliency Map-guided Colorization with Generative Adversarial Network (SCGAN) architecture. The proposed SCGAN has the following advantages. Firstly, it adopts dual encoders, one of which is a well-trained VGG-Net [48] for extracting semantic information. Since semantic information is implied in this VGG-based encoder, the proposed system can distinguish plausible colors for objects with similar edges. Secondly, the decoder of proposed system has two branches for producing colorization and saliency map, respectively. To augment visual salient area, we compute multiplication of the two outputs to obtain a weighted image representing salient areas, as shown in last row of Figure 1. Then, we leverage an attention loss to emphasize

the salient areas at training. Finally, it includes two discriminators for entire image and weighted image, respectively. The adversarial training strategy [49] enhances sharpness and color vividness of images. Moreover, the saliency map branch better assists the mainstream to generate plausible colorization.

In addition, conventional fully automatic colorization approaches often require large training datasets such as ImageNet [46] and Places [47]. The proposed SCGAN can be trained on a relatively small dataset (e.g. subset of ImageNet, 0.13M images). It utilizes a saliency map-based guidance method to produce visually plausible colorization in the salient regions in the image. We notice that acquisition of large dataset in some low-level vision applications is much harder compared to natural image colorization. For example, the multispectral image colorization [1], [2], [50] requires complex imaging system and precise alignment technique. The proposed saliency map-based guidance method is beneficial to such applications.

Compared with the existing methods, the main contributions of this paper are as follows:

1) We employ the saliency map as an additional proxy task in the proposed SCGAN that improves the performance;

2) We propose a saliency map-based guidance method that helps our system effectively predict a fine colorization with a relatively small training dataset;

3) We firstly use an effective evaluation criterion CCI (Color Colorfulness Index) to evaluate colorization quality and show its high correlation with human observers;

4) We apply SCGAN architecture with saliency map-based guidance method to multispectral image colorization and obtain state-of-the-art results.

II. Related Work

A. Scribble-Based Colorization

Scribble-based colorization method is the most straightforward way to achieve colorization of grayscale image, but it is extremely labour-intensive. It is based on prior color scribbles and then propagates them to the rest of the image. Levin *et al.* [9] proposed an optimization-based system and assumed that the adjacent pixels with same illuminance could have similar colors. This technique was enhanced using an additional adaptive edge detection algorithm by Huang *et al.* [10]. Yatziv and Sapiro [11] proposed a fast colorization algorithm based on the concepts of luminance-weighted chrominance blending. To enhance long-range color propagation, Xu *et al.* [13] performed an affinity-based edit scheme and Chen *et al.* [15] utilized the locally linear embedding to model the linear combination for adjacent pixels in a feature space. However, the main weakness is that they only concentrate on one aspect of propagating local or global color hints. The results are highly related to the number and location of given color scribbles. To address the ambiguity brought by sparse scribbles. Xu *et al.* [12] proposed a novel approximation scheme requiring much less time and memory and Paul *et al.* [51] proposed a 3D steerable pyramids approach for occlusion handling. Since the aforementioned methods require accurate scribbles for colorization, Zhang *et al.* introduced an additional deep prior from a CNN to ensure plausible colorization when no given scribbles. Those methods are still easy to overfit to scribbles. Moreover, scribbles with similar pixel locations often lead to color bleeding in colorized images.

B. Example-Based Colorization

In contrast, the example-based colorization approaches exploit color information from a reference image to guide colorization. It reduces the difficulty of choosing many color scribbles. They mainly match spatial features between reference image and input grayscale image by statistical analysis [20], [26]. This idea was enhanced by characterizing the image patches using GMM [27], discriminating different regions by segmentation maps [25], predicting probability for each pixel by global optimization [22], and modelling color selection by energy-minimization method [18]. Moreover, superpixels [19], [21], [24] were utilized to model the correspondences between grayscale input and reference. To alleviate effort of selecting proper reference images, Chia *et al.* [29] developed an image retrieval method to download appropriate reference images from Internet. However, those methods are highly based on references which are remarkably close to grayscale input. The colors of output images often appear unnatural when given images not similar to input. In order to generalize to more reference images, He *et al.* [23] and Zhang *et al.* [52] applied deep image analogy technique and neural network to match the semantics of the target image and reference accurately. In addition, researchers used more types of references as guidance for colorization such as words [53], [54] and complete sentence [55]. However, the combination of examples and input grayscale image is difficult in terms of transferring examples to useful color information.

C. Fully Automatic Colorization

Recently, fully automatic colorization methods have outperformed traditional methods due to their robustness and generalization. They are based on CNN to learn mapping from grayscale to color embeddings as a self-supervised task chiefly. Cheng *et al.* [40] first adopted a deep neural network to colorize the images based on the extracted features from different patches. However, their training dataset is too small and network structure is simple. Without using handcrafted features, Larsson *et al.* [33] proposed an end-to-end CNN architecture. The hyper-column of a pre-trained VGG-Net is utilized to augment original grayscale input; whereas its memory consumption is too high. Iizuka *et al.* [38] developed a two-stream architecture to jointly predict the color embedding and category of the scene. The semantics from classification sub-network are merged into mainstream by a fusion layer. Zhang *et al.* [37] adopted a VGG-styled network with added depth and dilated convolutions. They introduced cross-channel encoding and class rebalancing techniques to resolve the inherent ambiguity and multimodal nature of the colorization problems. However, those methods retain common artifacts in colorization area such as color bleeding and semantic confusion. To address these problems, Zhao *et al.* [34], [35] added

segmentation information and Lei and Chen [43] proposed a bilateral loss for self-regularization.

Moreover, some generative models have been leveraged for multimodal colorization. Isola *et al.* [36] proposed a general image-to-image translation framework based on conditional GAN [49]. The experimental results demonstrated that the vividness of colorized images was enhanced due to adversarial training. Deshpande *et al.* [31] utilized a mixture density network (MDN) to map the grayscale images to GMM. There are numerous possible vectors sampled from GMM and each corresponds to a colorization type. It was enhanced by Messaoud *et al.* [56] by introducing structural consistency. Based on capturing dependencies of neighbouring pixels to ensure color consistency, Royer *et al.* [32] and Guadarrama *et al.* [42] developed a PixelCNN network to produce multiple plausible and vivid colorizations for a given grayscale image.

D. Salient Object Detection

The early works of salient object detection (or saliency detection) were based on hand-crafted features such as color variation [57], boundaries [58], and center prior [59]. They preserve the edges of images well but ignore the integral structural features. To predict robust saliency maps, Li and Yu [60] proposed a multiscale feature extraction for superpixel saliency detection. Liu *et al.* [61] combined image-level and superpixel-level features into saliency detection. However, hand-crafted features are hard to generalize to different scenes. Thus, the CNN is adopted to improve generalization ability of saliency detection algorithms. Researchers developed diverse architectures such as recurrent network [62], encoder-decoder [63]–[66], feature pyramid network [67]–[71] to fuse low-level edge details and high-level semantics. Some methods [64], [65], [69] used attention mechanism, which further improved the accuracy due to use of dense connections for each pixel. Recently, some extensions focus on improvement of network architecture to effectively use features. For instance, Liu *et al.* [66] designed a pooling-based pyramid architecture to accurately locate salient areas. Pang *et al.* [72] effectively used multi-level and multi-scale information and proposed a feature aggregation module. Zhao and Wu [70] proposed a pyramid attention network that integrates different levels of information from VGG-Net. In conclusion, edge guidance, attention mechanism and semantics greatly improve the performance of saliency detection. In this paper, we choose the approach [70] to generate robust and accurate saliency maps.

E. Generative Adversarial Network

The GAN was proposed by Goodfellow *et al.* [49] to generate data in an unsupervised manner. It contains a generator that learns to produce realistic data and a discriminator that judges whether the input is generated by generator or sampled from ground truth. The system is trained to minimize the JS-divergence between generated samples and target dataset. To stabilize its convergence, some advanced divergences for estimating feature disparity were proposed, such as f-divergence [73], Pearson χ^2 divergence [74], and Earth-Mover distance [75], which was further improved by adding a gradient punishment [76]. Compared to traditional pixel-level loss, the adversarial loss minimizes the various divergences between the generated images and the real images in the target domain, leading to a substantial boost of the results. The proposed SCGAN framework aims at producing perceptually high-quality colorizations.

F. Comparative Analysis of Colorization Methods

Early colorization methods often require human hints such as scribbles and reference images as guidance. They [10], [16], [24], [25], [30] mainly utilized hand-crafted features including low-level SIFT or edges and high-level scene or location categories. The limitation of these works is not general to images in different scenes. Recently, deep neural networks have been utilized to address this problem. They mainly adopted pre-trained networks to enhance colorization quality but individual optimization skills. Thus, their colorization effects are different, e.g. classifying color for pixels [37] promotes very colorful results; training with scene classification [38], [44] ensures overall color correctness; contextual loss [52], [77] facilitates color similarity with ground truth. Moreover, to alleviate color bleeding and semantic confusion, additional constrains such as gradient loss [31], segmentation loss [34], [35], and bilateral loss [43] were proposed. They worked well in some circumstances; whereas saliency map is more general to all images compared with them. In this paper, we propose to extract semantics and use a novel joint training with saliency detection.

III. Methodology

A. SCGAN Architecture

An overview of the SCGAN framework is shown in Figure 2 and Figure 3. Our method is based on a hierarchical GAN architecture that produces colorizations and saliency maps from grayscale images jointly. It consists of four parts: global feature encoder, main colorization network, saliency prediction network, and patch-based discriminators. The first three components constitute generator. The global feature encoder adopts VGG-16-Gray [48] architecture, all max-pooling layers of which are replaced by convolutional layers with stride of 2. While the main colorization network is based on U-Net structure [78] with skip connection between each encoder layer i and decoder layer $n-i$ with same resolution, where n is the total number of layers, as green lines shown in Figure 2. It effectively prevents gradient vanishment and accelerates convergence. In order to fuse global information and local low-level information, the resultant layer of global feature encoder is concatenated with the middle layer of main colorization network. Moreover, three layers of the decoder are used to predict the saliency map with same spatial resolution as colorization.

Two discriminators share the same PatchGAN [36] architecture, as shown in Figure 3, which effectively models the image as Markov random field and strengthens high-frequency correctness in local image patches. The first discriminator judges

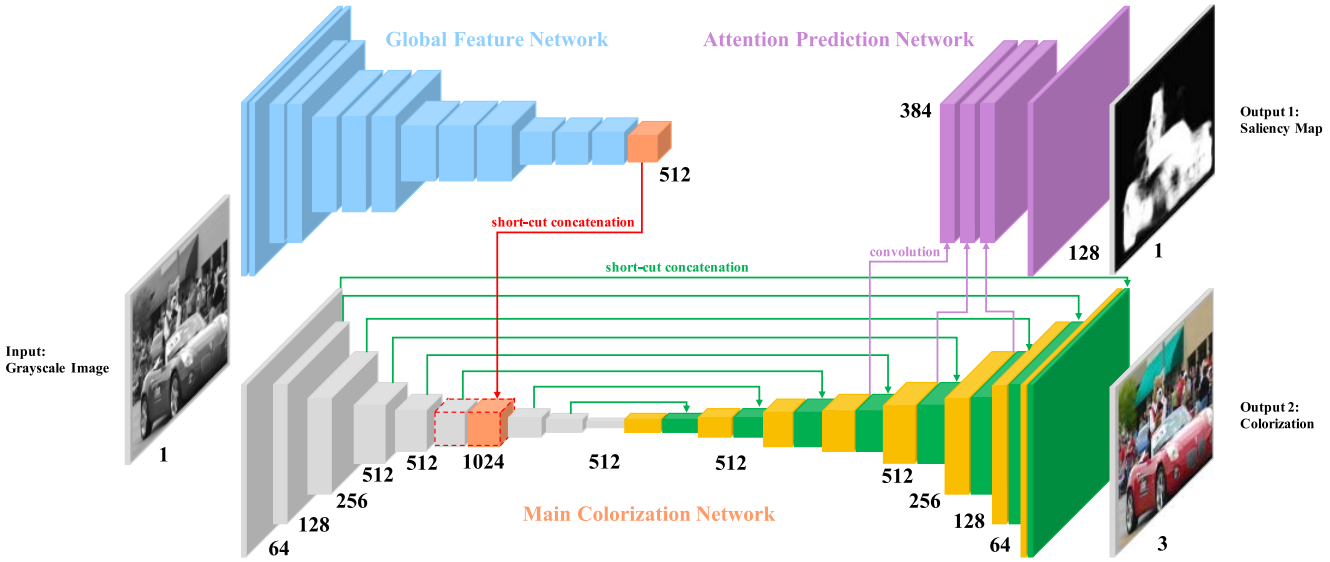

Fig. 2. Architecture of the proposed SCGAN generator. It receives a grayscale image as input and predicts a colorized image with a corresponding saliency map. The scalar denotes number of channels for relevant block. Different colors represent the distinct parts of generator architecture.

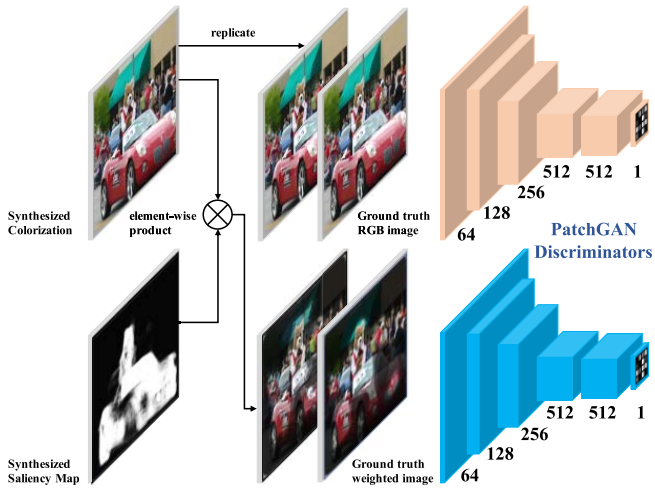

Fig. 3. Architecture of the proposed SCGAN discriminators. The inputs of the two discriminators (color image discriminator and attention-based discriminator) are pairs of colorized images and the images with attention region, respectively.

the colorized image and ground truth color image. In addition, we perform element-wise product on colorized image with generated saliency map to obtain a weighted image. Similarly, we can obtain a ground truth weighted image by same computation scheme. Then, we feed the pair to the second attention-based discriminator, which judges whether the input is real weighted image or not. Based on the work in [36], we choose 70×70 PatchGAN architecture including reasonable parameters for better visual quality.

B. Attention Mechanism and Training Schedule

Saliency maps are commonly used to explicitly represent the visual attention areas. Based on this observation, we assume these salient areas have more colorful patterns or higher variance, which are essential for enhancing rare colors when developing a colorization algorithm. To emphasize the areas, we propose to perform element-wise product between the colorful image and its saliency map. The output weighted image includes regions rich in color while filtering out flatten regions, as shown in Figure 3. By enforcing an additional attention loss, as represented in Equation (2), on weighted image, the saliency prediction network assists the main colorization network in revising its bottom layers. Therefore, this attention mechanism serves as a kind of guidance for colorization.

Since GAN architecture is highly nonlinear, random initialization tends to converge to local minima. To facilitate and stabilize its convergence, we defined a two-phase training procedure. Firstly, SCGAN generator is only trained with L1 loss, which can remove outliers so that the generator achieves better generalization than L2 loss. Therefore, a stable adversarial training process can be maintained by striding a balance between generator and discriminators. At second stage, we construct the whole SCGAN by alternately training the generator and discriminators. Note that, the saliency map-based guidance method exists in both stages.

C. Objectives

At first stage, the L1 losses for colorized image and weighted image are jointly considered. Thus, they emphasize both pixel fidelity and perceptually significant regions of the generated images. The losses are defined as:

$$L_1 = \mathbb{E}[||G_c(x) - c||_1], \tag{1}$$

$$L_A = \mathbb{E}[||G_c(x) \odot G_s(x) - c \odot s||_1], \tag{2}$$

where x, c and s represent input grayscale image, ground truth colorful image and saliency map, respectively. The $G_c(x)$ and $G_s(x)$ are the colorized image and corresponding saliency map. The operator $\odot$ means element-wise product.

At second stage, we add two additional discriminators $D_c(*)$ and $D_A(*)$, respectively. The WGAN-GP loss [75] items are defined as:

$$L_G = -\mathbb{E}[D_c(G_c(x))] - \mathbb{E}[D_A(G_c(x) \odot G_s(x))], \quad (3)$$

$$\begin{aligned} L_D = {} & \mathbb{E}[D_c(G_c(x))] + \mathbb{E}[D_A(G_c(x) \odot G_s(x))] \\ & - \mathbb{E}[D_c(c)] - \mathbb{E}[D_A(c \odot s)] \\ & + \lambda \mathbb{E}[(||\nabla_{\ddot{c}} D_c(\ddot{c})||_2 - 1)^2] \\ & + \lambda \mathbb{E}[(||\nabla_{\ddot{s}} D_s(\ddot{s})||_2 - 1)^2] \end{aligned} \quad (4)$$

where Equation (3) and the first four terms of Equation (4) constitute original WGAN loss, the remaining of Equation (4) is a gradient penalty. According to [76], we define $\ddot{c}$ and $\ddot{s}$ sampling uniformly along straight lines between pairs of points sampled from the synthesized images $G_c(x)$, $G_s(x)$ and ground truth images c, s, respectively. The gradient penalty coefficient λ is set to 10.

In order to improve perceptual quality, we measure the image semantic similarity in high-level feature space by perceptual loss [79]. It is defined as:

$$L_p = \mathbb{E}[||\phi_l(G_c(x)\ ) - \phi_l(c)||_1], \quad (5)$$

where $\phi_l(*)$ represents the features of the l-th layer of the pre-trained network. In our experiment, we use the ReLU [80] activated $conv_{3\_3}$ layer of VGG-16 network pre-trained on ImageNet dataset.

The total loss function of the generator for the second stage includes Equation 1, 2, 3, and 5, which is given by:

$$Loss = L_1 + \lambda_G L_G + \lambda_A L_A + \lambda_p L_p. \quad (6)$$

IV. Experiment

A. Implementation Details

For training set, a subset of ImageNet [46] (0.13M images) is utilized, which is only one tenth of the size of training dataset comparing with the state of the art [33], [37], [38], [43]–[45]. We randomly sample images from 1000 categories, corresponding to the proportion of the entire dataset. It provides enough modes for SCGAN to learn the mapping robustly. The images are rescaled to 256 × 256. They are normalized within [-1, 1] range and an additive Gaussian noise with standard deviation of 0.005 is added. In addition, the corresponding saliency maps are generated by [70], which are set as ground truth. They are normalized in [0, 1] range, which represent different levels of significance for salient regions.

For network architecture, the global feature network is trained from scratch until its Top-1 accuracy is verified to be sufficiently high and stable. It adopts VGG-16-Gray architecture, where each max-pooling layer is replaced with strided convolutional layer to maintain more spatial information. Batch normalization [81] and LeakyReLU activation function [82] are attached to each convolutional layer of SCGAN except the input and output layers. The reflection padding scheme is utilized to avoid border effects. Moreover, with spectral normalization [83] attached to each discriminator layer, the SCGAN meets 1-Lipschitz continuity.

For optimization details, the parameters of SCGAN are initialized using zero mean Gaussian distribution with standard deviation of 0.02 except global feature network. We train SCGAN generator with L1 loss and attention loss for 10 epochs at first stage and the learning rate is fixed to 2×10^{-4}. At the second stage, we train the generator and discriminators collaboratively for 30 epochs. The initial learning rate for both generator and discriminator are 1×10^{-4} but halved every 10 epochs. We use Adam optimizer [84] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size of 8. The discriminators and generator are trained alternately until the SCGAN converges. The trade-off parameters λ_G, λ_A, and λ_p are empirically set to 0.05, 0.5, and 5, respectively. We implement our system with PyTorch framework and train it on a NVIDIA Titan Xp GPU. It takes approximately 7 days to complete the whole training process.

B. Experimental Settings

1) Dataset: To assess colorization quality, we set up 10000 images from ImageNet validation set [46], same as in [33], [37] for evaluation. They are randomly selected and have a balanced representation for ImageNet categories. All the validation images encoded as grayscale are excluded and rescaled to 256×256. To further demonstrate the effectiveness of several network components, we use both quantitative and qualitative analysis.

2) Quantitative Metrics: On the one hand, we apply pixel-level PSNR (peak signal to noise ratio) and SSIM (structural similarity index) [85] metrics to evaluate the pixel fidelity of an image. On the other hand, since there might be many reasonable color guesses given the grayscale input, we use high-level Top-1 accuracy (computed by a well-trained VGG-16 [48]) to measure semantic interpretability of synthesized images.

3) Color Colorfulness Index: In addition, we firstly use a new non-reference evaluator called CCI (color colorfulness index) for colorization evaluation. Basically, CCI is a professional index to measure the color vividness and naturalness [86]–[88]. Compared with traditional PSNR, CCI focuses more on color shift and saturation level. It can be viewed as a significant index for evaluating color reasonability of generated images and is defined as:

$$CCI_k = S_k + \sigma_k, \quad (7)$$

where S_k is the average saturation of image k, and σ_k is the standard deviation. Note that CCI varies from 0 (achromatic image) to ∞ (most colorful image). However, the optimum range of CCI for a generated color image is between 16 and 20 based on large amount of experiments [88]. The correlation of optimum range of CCI with human perception equals to 95.3%. Since the human visual system (HVS) captures color information in opponent color space [87], [88], the RGB image is first transformed into opponent color space to compute CCI value. The transformation functions are defined as:

$$rg = R - G, \quad (8)$$

$$yb = (R + G)/2 - B. \quad (9)$$

TABLE I
COMPARISON RESULTS OF SCGAN AND STATE-OF-THE-ART FULLY AUTOMATIC COLORIZATION ALGORITHMS

Method	PSNR	SSIM	Top-1 Accuracy	CCI Ratio	Color Naturalness	Color Bleeding Removal	Color Colorfulness
Ground Truth	/	1	63.44%	/	/	/	/
Grayscale	23.23	0.9394	49.78%	/	/	/	/
Larsson *et al.*	**24.42**	0.9229	**55.16%**	14.93%	9.14	8.58	8.22
Isola *et al.*	23.25	0.9386	52.29%	11.26%	8.56	7.93	8.96
Zhang *et al.*	22.49	0.9153	53.97%	3.300%	9.05	7.10	**9.50**
Iizuka *et al.*	24.32	0.9464	53.05%	19.60%	9.17	8.34	8.76
DeOldify	23.14	0.9194	53.45%	14.73%	9.20	8.57	9.01
Lei *et al.*	22.96	0.9146	51.46%	11.40%	8.02	7.14	7.45
Vitoria *et al.*	24.32	0.9273	53.65%	11.24%	9.03	8.24	9.17
SCGAN	23.80	**0.9473**	53.47%	**21.41%**	**9.32**	**8.68**	9.04

Hasler and Suesstrunk [86] proposed a more accurate method for computing CCI, which is used in our experiment and defined as:

$$CCI_k = \sigma_{rgyb} + 0.3\mu_{rgyb}, \tag{10}$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \tag{11}$$

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \tag{12}$$

where σ_* and μ_* represent standard deviation and mean value, respectively. We calculate the ratio of the number of generated images in optimum range to the whole 10000 validation images, which represents the reasonability degree for the colorization system.

4) Human Perceptual Evaluation: Since the evaluation of color naturalness, colorfulness, and color bleeding removal are highly subjective, we perform a qualitative perceptual evaluation. The color naturalness denotes whether colorized images are similar to real-world images. It emphasizes color reasonability rather than high brightness or vividness. Conversely, color colorfulness score is high as long as generated images are very colorful, even though the color is not authentic. Moreover, color bleeding artifact exists in an image when color of one object permeates through other objects. The color bleeding removal judges the ability of colorization systems to prevent or reduce such artifact.

There are overall 20 observers participating in the test. Each observer was given 30 groups of grayscale images, ground truth colorful images, and images colorized by different algorithms. For each result, the observer was required to decide its color naturalness and severity of color bleeding by scoring 0-10. For instance, 0 represents the most achromatic or severely color bleeding images and 10 indicates the reverse. Finally, we calculate the average score across all 30 colorized images and from every observer.

C. Comparison With State of the Art

We utilize 7 state-of-the-art fully automatic algorithms [33], [36]–[38], [43]–[45] for comparisons. Some colorized results of proposed SCGAN and other methods are shown in Figure 4 for qualitative measurement. The results from [33], [38], [43] look more unsaturated than others in the second and fifth columns. In the third and fourth columns, there is semantic confusion effect. For instance, the grass of fourth row from [36], [37] is polluted since the methods fail to classify grass and wave with similar jagged edges. Moreover, the color of sea permeates through fish, as shown in fifth row from [37], [44], [45]. In sixth row from [33], [36], [38], [44], the colors of some fruits bleed into others. However, the results from proposed SCGAN are more reasonable and natural. For human perceptual evaluation, the scoring results are summarized in Table I. The SCGAN has better performance than other methods since it produces more natural colors. The saliency map could provide attention segmentation for SCGAN, which is beneficial for removing color bleeding effect.

In addition, the results of quantitative metrics are shown in Table I. SCGAN ranks first in the SSIM metric. It means that proposed system could accurately model the perceptual structure of reconstructed images. As PSNR is not highly related to human visual system (HVS) [85], SSIM is proposed to grasp the structural characteristics (luminance, contrast, and structure) of the image. We think SSIM is better to estimate whether the colorization is distorted or not. The proposed SCGAN with high SSIM can generate structural consistent results compared with original color images, which demonstrates the colorization is more reasonable. The SCGAN also has sound results across other quantitative evaluators.

The CCI distributions of all validation images for different algorithms are shown in Figure 5. On the one hand, methods [36], [37] have larger CCI means and variances than others, indicating that the colorization is very saturated and color shifts very much in many generated images, respectively. Moreover, the method [37] has the best performance of color colorfulness but the worst color bleeding removal. It demonstrates CCI metric focuses more on color reasonability and contrast. On the other hand, methods [33], [38], [44], [45] obtain rational CCI distribution and good PSNR and SSIM values since they generate more natural colorization. But they have less scores in perceptual evaluation and SSIM than SCGAN. Finally, the proposed SCGAN occupies the most compact distribution over CCI near the optimal range [16, 20] than other methods obviously. It demonstrates that our system produces plausible colorization and has less probability to encounter semantic confusion and color bleeding.

The human perceptual evaluation indicates that SCGAN achieves the best performance over color naturalness and color bleeding removal. Since saliency map assists the system to retain a clear separation of foreground and background, the color bleeding effect of SCGAN is less than other methods.

Fig. 4. Comparison of colorization results between the proposed SCGAN and the state-of-the-art approaches [33], [36]–[38], [43]–[45] by 10 samples. The first column shows the grayscale input images. Column 2-9 show automatically generated results from the state-of-the-art approaches and the proposed method. The final column shows the ground truth images.

Moreover, we use attention loss with adversarial training in SCGAN. They promote the system to strengthen colorization on key objects. Thus, SCGAN produces more natural colorizations. Zhang *et al.* [37] obtains the highest color colorfulness due to its classification-based training scheme. From these experiment results, we notice that CCI ratio metric is highly related to color naturalness. When the average of CCI is very high, the system tends to produce colorful samples, although they may be not natural. However, it cannot represent the ability of removing color bleeding since it focuses on the global characteristic of images.

D. Ablation Study

In order to further demonstrate the effectiveness of several network components, we analyze different components of our system on validation dataset quantitatively. Basically, there are 7 settings to exclude some parts from original structure:

1) SCGAN w/o attention loss. Drop the saliency prediction network and its corresponding discriminator in order to analyze the effect of attention mechanism of system. We utilize twice amount of data (one fifth of ImageNet training set) for training to demonstrate the effectiveness of saliency map-based guidance method;

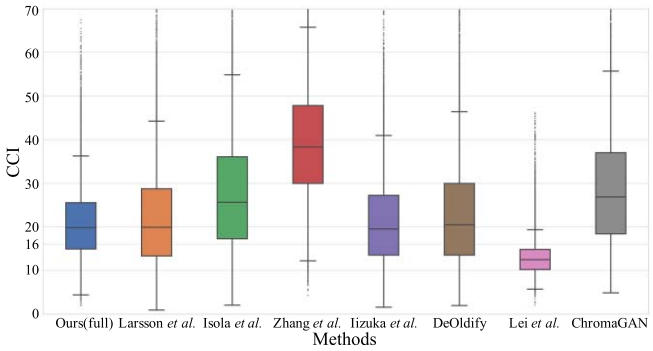

Fig. 5. Box plot of CCI distributions for the proposed SCGAN and state-of-the-art methods.

TABLE II
QUANTITATIVE RESULTS OF ABLATION STUDY ON 10000 IMAGENET VALIDATION SET

Method	PSNR	SSIM	Top-1 Acc	CCI Ratio
w/o attention loss	**23.81**	0.9368	52.34%	18.56%
w/o GAN loss	23.28	0.9305	52.89%	20.45%
w/o perceptual loss	23.80	0.9443	52.11%	21.31%
with LSGAN loss	23.46	0.9390	53.42%	20.86%
w/o pre-weights	23.15	0.9280	52.59%	18.20%
w/o global feature	23.61	0.9356	52.16%	17.55%
with L2 loss	23.67	0.9436	53.26%	19.58%
SCGAN (full)	23.80	**0.9473**	**53.47%**	**21.41%**

2) SCGAN w/o GAN loss. Drop the two discriminators of colorized images and weighted images, with the adversarial training to analyze the adversarial loss in SCGAN. This setting will not affect the architecture of generator;

3) SCGAN w/o perceptual loss. Drop the perceptual loss at second stage. This setting only changes the optimization method, while the network architecture is remained;

4) SCGAN with LSGAN loss. Replace original WGAN-GP training strategy with LSGAN [74] at the second stage. It minimizes the Pearson χ^2 divergence between the generated samples and ground truth;

5) SCGAN w/o pre-weights. Initialize the parameters of global feature network using Gaussian distribution. It evaluates the utility of pre-trained weights for global feature network since SCGAN architecture is unchanged;

6) SCGAN w/o global feature. Delete the global feature network to analyze the effect of this module. Although it will reduce complexity of the system, the main idea of this setting is to evaluate the effectiveness of semantic context information;

7) SCGAN with L2 loss. Use L2 loss instead of L1 loss at both training stages. The optimization method remains unchanged.

As shown in Figure 6, the full SCGAN has the best perceptual performance compared with the six ablation study settings. If global feature network or its pre-trained weights are removed, the color of generated images is unimaginative. The system without global semantics predicts wrong colorizations and causes semantic confusion. The attention loss emphasizes the significant parts, thus the main objects in colorizations are clear separated from backgrounds. For instance, the color of chicken in Figure 6 first row is obvious for full SCGAN; whereas the edges between chicken and background are blurry for system without attention loss. In addition, the perceptual loss and GAN loss enhance the sharpness of colorization. In Figure 6 column 3-5, the samples are less natural and colorful than full SCGAN.

The quantitative analysis is summarized in Table II. Firstly, if visual saliency information and attention branch are removed (setting 1), the system tends to generate samples with shifted distribution over CCI. Although we utilize double amount of data to train the system, it lacks visual saliency information so that the optimization is altered. Moreover, we also train the system without attention loss using same data (one tenth of ImageNet training set) compared with normal process. The performance is still inferior to full losses. Secondly, GAN loss (setting 2) promotes the SCGAN to produce more colorful images. The perceptual loss (setting 3) facilitates the semantic interpretability of system and generates sharper images. L1 loss performs slightly better than L2 loss (setting 7) according to color abundance. Thirdly, the SCGAN without global feature network or its pre-trained weights (setting 5 and 6) have much worse ability to represent the semantics, leading to bad results over the metrics, especially classification accuracy. Finally, SCGAN with LSGAN loss (setting 4) produces worse results than the WGAN-GP loss. In conclusion, each component of the proposed SCGAN is indispensable.

E. Saliency Map-Based Guidance Method

The SCGAN produces colorization with corresponding saliency map for grayscale image, which enhances the attention interpretation ability. As aforementioned, saliency map provides attention intensity and segmentation information [89] in an unsupervised manner. We illustrate the attention region by the multiplication between the colorization and saliency map and comparison with SCGAN without saliency map, as shown in Figure 7. Firstly, the foreground objects are obviously highlighted in all generated colorizations (last column of Figure 7). The saliency maps may contribute to less color bleeding effect since the foreground and background are well separated. Secondly, the colorizations generated by full SCGAN are more colorful than it without saliency map, especially the key objects. For instance, in 1st and 4th rows, the cheetah and bird colorized by full SCGAN are more realistic. As a proxy task, the generated saliency map assists the system to pay more attention to visually important regions. This mechanism can be viewed as a guidance, making SCGAN more efficient at training stage.

In order to further demonstrate the saliency map is more efficient, we propose to compare the variance of color for salient regions and the opposite. To measure color characteristics, the H channel in HSV color space is utilized in our experiment. Since the saliency map is irregular, we alternatively choose small patches (64×64 in experiment) of each training image to include high response area. The definition of salient region in RGB image is that there should be at least 80% pixels

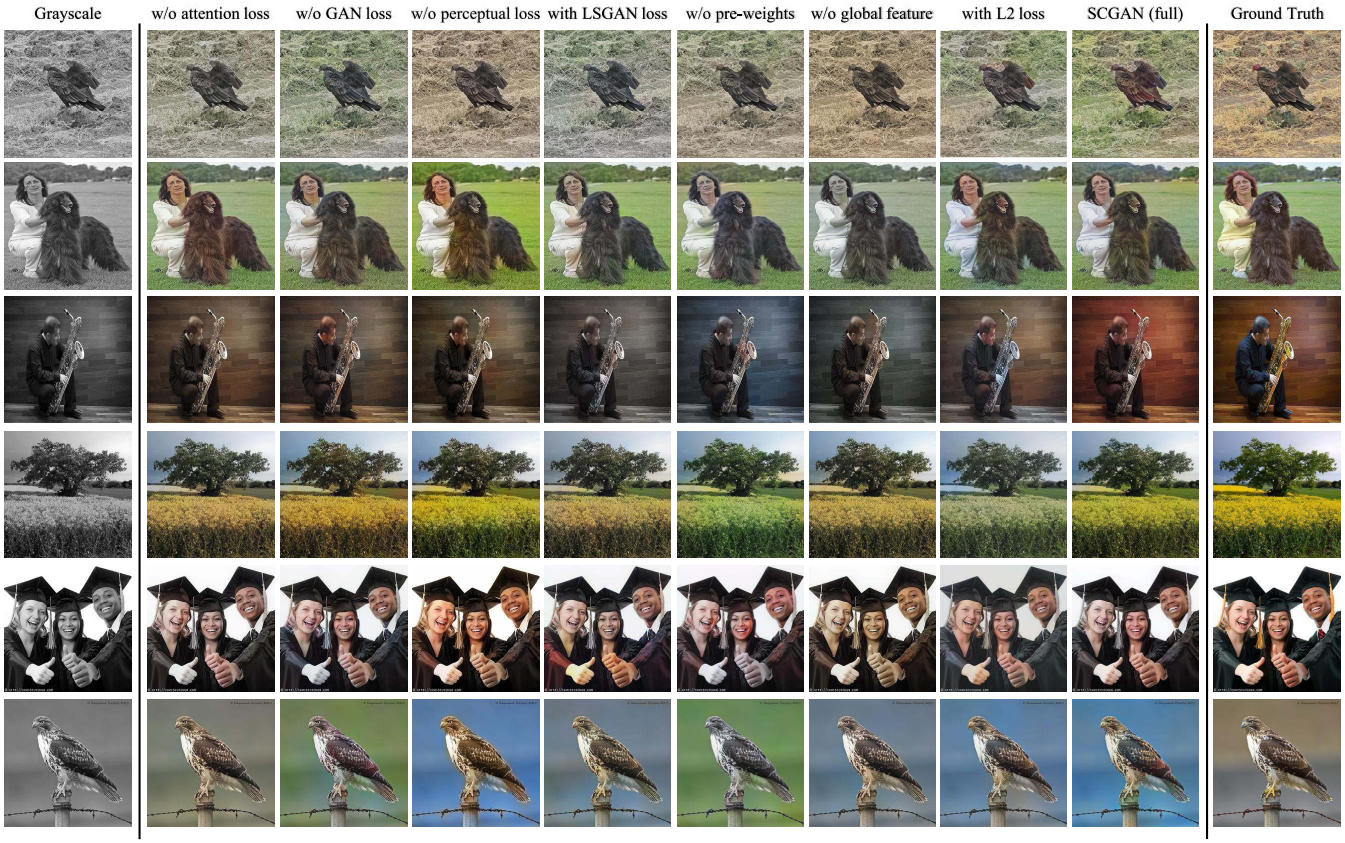

Fig. 6. Comparison of colorization results of different ablation study settings and full SCGAN. The first column shows the grayscale input images. Column 2-9 represent the colorization results of different settings. The final column is the ground truth colorful images.

having high value in same spatial location of corresponding saliency map. Conversely, unsalient region indicates the no response area. For fair comparison, we also count random regions as reference, as shown in Figure 8. The H value represents the color category, expressed by angle in HSV color space. The H variance of salient regions is much larger than unsalient regions that demonstrates they contain more colors. At training, the saliency map with attention loss stresses such regions, which contributes to the regression of diverse colors.

Moreover, the colorization system tends to learn green and blue colors first (please see supplementary for examples) since they are common in natural images, e.g. lawn and sky. The salient regions have less percent (26.56%) of green-blue range than randomly selected regions (30.52%) and unsalient regions (31.29%), as shown in Figure 8. Thus, other colors are more probable to be utilized for SCGAN optimization. It can be regarded as a color augmentation. We believe this mechanism enhances colorization system to produce more plausible results.

F. Colorizing Multispectral Images

In order to further verify the advance of SCGAN architecture and saliency map-based guidance method, we perform a multispectral image colorization experiment on KAIST multispectral pedestrian detection dataset [50]. There are four network architectures used for comparison: Pix2Pix [36] and it with proposed saliency map-based guidance method (Pix2Pix+Sal), UResNet [2] and it with proposed saliency map-based guidance method (UResNet+Sal), and the proposed SCGAN. The L1 loss is adopted for optimization while attention loss with same trade-off parameter λ_A is utilized for Pix2Pix+Sal, UResNet+Sal and SCGAN. Each network is trained for 20 epochs from scratch. Some colorized results are shown in Figure 9. If the network is trained with attention loss, the output is richer in color and clearer, e.g. the cars in Figure 9. The results from SCGAN are sharper than other methods. Moreover, the quantitative analysis on KAIST validation set is summarized in Table III. With saliency map-based guidance method, UResNet can obtain higher PSNR and SSIM values. Since the KAIST dataset only contains approximately 90000 training pairs, which are much less than ImageNet, the function of the proposed saliency map-based guidance

TABLE III
QUANTITATIVE RESULTS OF MULTISPECTRAL IMAGE COLORIZATION ON KAIST VALIDATION SET

Method	PSNR	SSIM	Saliency Map Guidance
Pix2Pix	23.55	0.8165	-
Pix2Pix+Sal	23.53	0.8164	✓
UResNet	23.66	0.8219	-
UResNet+Sal	23.72	0.8244	✓
SCGAN	**24.59**	**0.8396**	✓

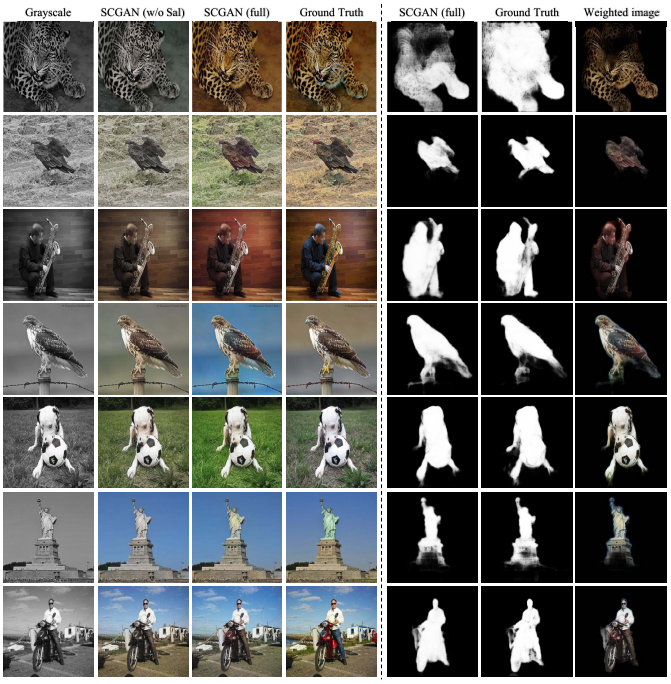

Fig. 7. Attention representation of the proposed SCGAN. The columns from left to right indicate that input grayscale images (1st), colorizations generated by SCGAN without saliency map (2nd), colorizations generated by full SCGAN (3rd), ground truth colorful images (4th), saliency maps generated by full SCGAN (5th), ground truth saliency maps (6th) and weighted images (7th), respectively. The weighted images are obtained by the multiplication of two outputs from SCGAN full system.

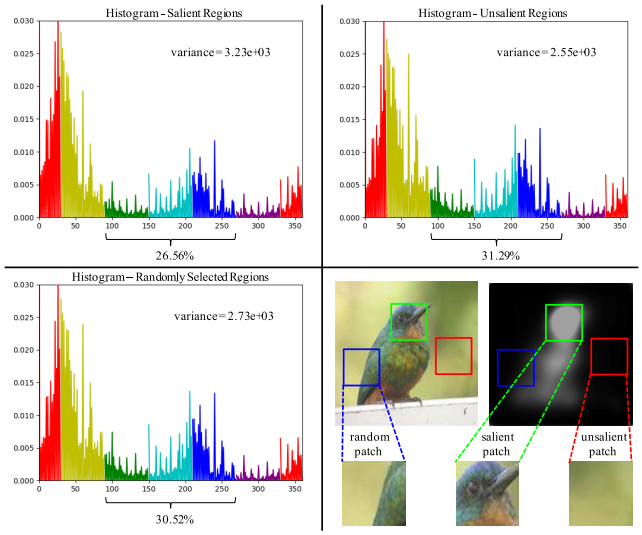

Fig. 8. Illustration of histogram of H channel for salient, unsalient and randomly selected regions. The percent represents the ratio of green, cyan, and blue color to all colors. The figure at lower right corner represents the patch selection scheme for the experiment. The rectangles with different colors imply 3 kinds of patches. The bird image is a training sample from ImageNet dataset.

method is evident. The SCGAN framework achieves the best performance across all the methods, since the convolutional layers of global feature network are general to multispectral images that boost the performance. It demonstrates the SCGAN network architecture is also more advance.

Fig. 9. Comparison of multispectral image colorization results between the proposed SCGAN and other approaches [2], [36]. The first row is the multispectral input while last row is ground truth visible RGB images. Row 2-6 represent colorization results of proposed SCGAN and other methods. We highlight two patches in each pair and the location is indicated by green and yellow rectangles.

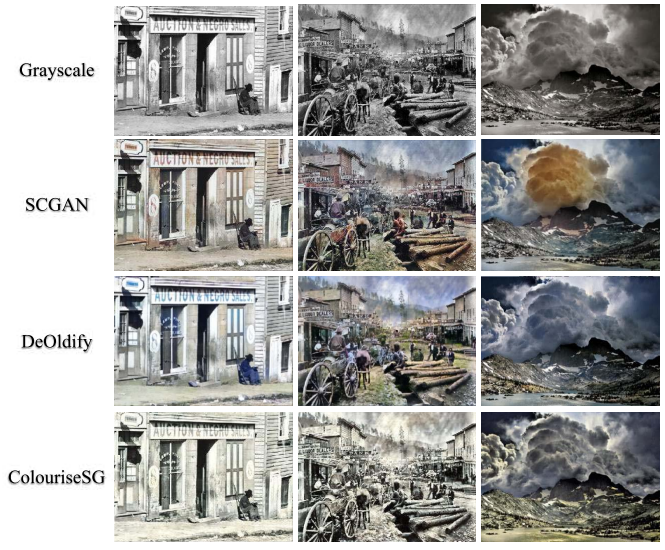

Fig. 10. Comparison of other legacy image-specialization colorization algorithms. The first row is the grayscale input. Row 2-4 are colorization results of the proposed SCGAN, DeOldify [45], and ColouriseSG [90], respectively.

G. Colorizing Legacy Photographs

We also test SCGAN on legacy black and white photographs and illustrate the colorization results along with the results of

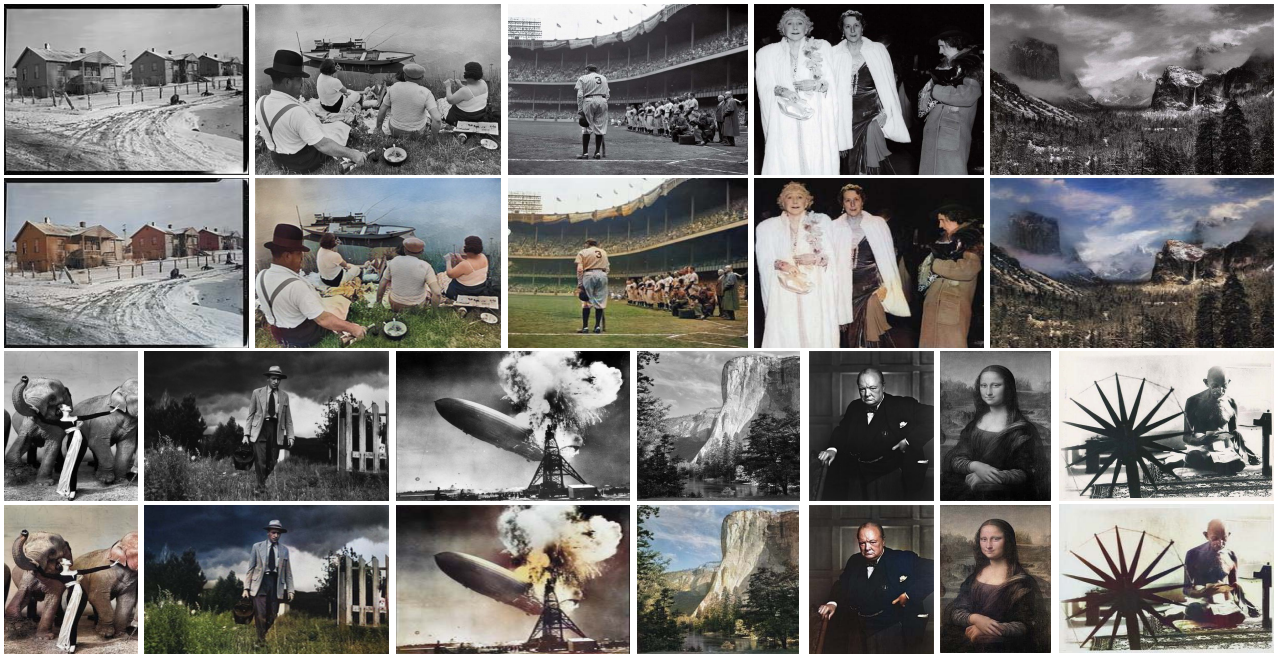

Fig. 11. Colorization results on historical photographs. Our colorization system still predicts visually high-quality reasonable images. The photos were taken from the US National Archives (Public Domain). For more colorized legacy photographs, please see the supplementary material.

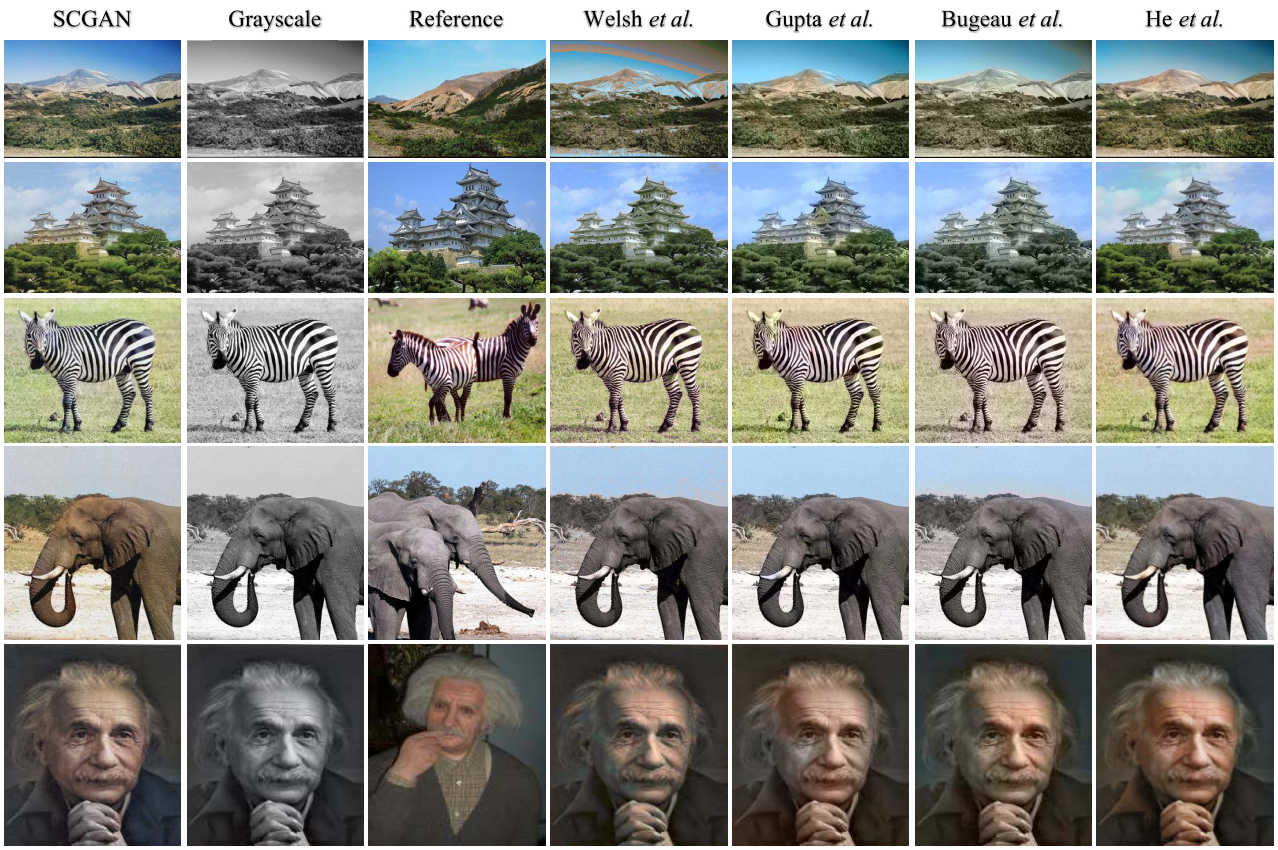

Fig. 12. Comparison of colorization results between the proposed SCGAN and the state-of-the-art exemplar-based algorithms [18], [23], [24], [26]. The first column shows the automatic colorization results of proposed method. The second column shows the grayscale input images. The third column is the references for remaining four algorithms, which are shown in column 4-7.

two recent open-use automatic colorization systems [45], [90], as shown in Figure 10 and 11. Due to the age and type of past photos and films, the statistic details are quite different from our training set and their edges are quite blur. However, SCGAN could produce plausible colorizations, which demonstrates its strong fitting ability. Since the training samples

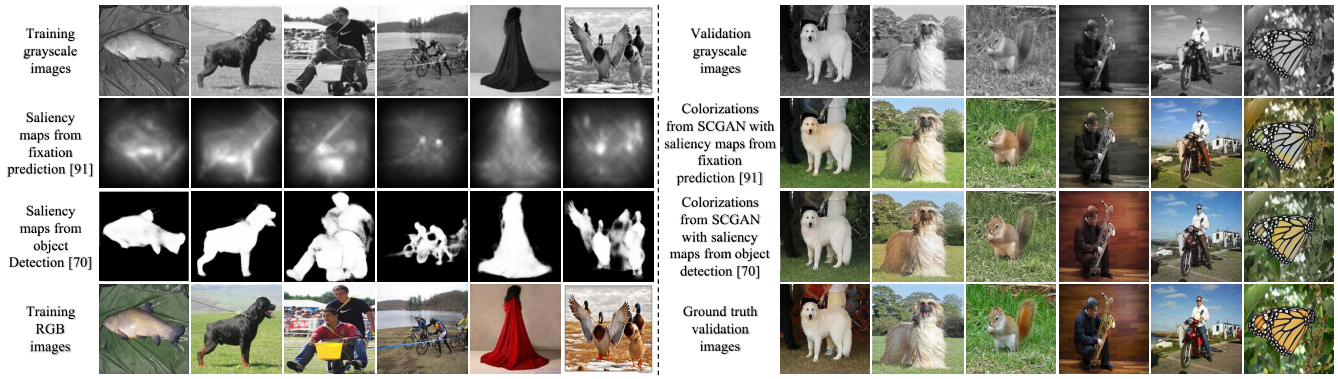

Fig. 13. Comparison of the two types of saliency maps. The left part of the figure includes the samples from ImageNet training set; whereas the right part represents the colorization results by SCGAN trained with different saliency maps. The first row and last row represent the grayscale input and ground truth RGB images. The saliency maps from fixation prediction and salient object detection are computed by [96] and [70], respectively.

Fig. 14. Examples of the most common failure cases. The top, middle and bottom rows include grayscale ground truth, the generated images, and colorful ground truth, respectively. The SCGAN may be sensitive to small objects like complicated scene, special patterns, and details respectively as shown in left 3 samples. The images generated by SCGAN may be not very colorful, as illustrated in the samples.

and legacy photographs are both real-world images, we further assume that SCGAN tends to learn general information primarily. During the optimization process, the system first reconstructs the profile of the objects and background. Then, it adds simple colors, like green and blue. Finally, it fixes the details and attaches special colors (please see supplementary material for illustration). It indicates that CNN-based models have strong ability to capture low-level image statistics [91] while natural images have similar statistical features. It demonstrates that SCGAN has great generalization ability on legacy photographs.

H. Comparison With Example-Based State of the Art

We also compare our system with state-of-the-art example-based algorithms [18], [23], [24], [26]. We report the comparison results in Figure 12. All the test grayscale images are accompanied with corresponding references. Compared with the example-based methods, the proposed SCGAN can still generate realistic and reasonable colorizations even though there are no references. The color styles of the images generated by the proposed method are implied in the training strategy and network architecture.

I. Discussion on the Usage of Saliency Maps

Basically, there are two methods [92], [93] to label the "ground truth" saliency maps: fixation prediction [94]–[100] and salient object detection [57]–[72], [101], [102]. The saliency maps from fixation prediction record the eye fixations of a user; whereas the saliency maps from salient object detection focus more on entire key objects. We show some saliency map samples generated by fixation prediction [96] and salient object detection [70] in Figure 13, respectively. The saliency maps from salient object detection have clearer edges of objects than from fixation prediction, which are beneficial for removing color bleeding artifact. Also, the key objects have more vivid colors than other areas. To compare the effects of two types of saliency maps, we additionally train the SCGAN using the saliency maps generated from fixation prediction [96] and salient object detection [70], respectively. The training strategies for them are the same. Some generated saliency maps and colorization samples are shown in Figure 13. In first three columns of right part of Figure 13, there are less color bleeding artifacts for SCGAN with saliency maps from salient object detection. While in last three columns, the colorizations from row 3 are more natural than row 2. In conclusion, the SCGAN trained with saliency maps from salient object detection achieves better perceptual quality. The saliency maps in this paper denote the ones from salient object detection.

J. Failure Cases

The proposed SCGAN can predict relatively reasonable colorizations in many samples; however, there are some common

failure cases, shown in Figure 14. It produces colorization and saliency map jointly so that core objects in images are well highlighted. There is less color bleeding effect in most of generated images. However, there is no specific loss item or network design for enhancing colors of details or small objects. Thus, SCGAN is difficult to identify plausible colors for such objects. Some failure cases are illustrated in Figure 14 first row. As we only use 0.13M training images, the system cannot include all the situations of input. Some generated images are not very colorful, as shown in Figure 14 second row. In the future, we will develop new methods generalized to small objects while generating more realistic colors.

V. Conclusion

In this paper, we presented a hierarchical GAN architecture called SCGAN. It generates perceptually reasonable and photorealistic colorful images and their corresponding saliency maps from grayscale input images automatically. This is achieved through a pre-trained VGG-16-Gray global feature network embedded to mainstream so that low-level and high-level semantic information are combined. In addition, we proposed a novel saliency map-based guidance method to perform the joint colorization and saliency map prediction. These designs help the system minimize semantic confusion and color bleeding in the colorized images. The proposed SCGAN framework can be trained with only one-tenth of ImageNet training data to achieve state-of-the-art colorization performance. Furthermore, we found that our system has potential to colorize multispectral images and legacy photographs with sundry scenes. Finally, we validated our system on ImageNet dataset against several state-of-the-art methods. Experiment results demonstrated that SCGAN can generate high-quality reasonable colorizations.

Acknowledgment

The authors would like to thank Mengyang Liu, Yujia Zhang, Weifeng Ou, Tiantian Zhang, Chang Zhou, and Kin-Wai Lau for many helpful comments. The authors would also like to thank Tingyu Lin for drawing the Figure 7 and 8. The authors would also like to thank Wei Liu for the support in the revision. The authors would also like to thank the anonymous reviewers and the editors for their kind suggestions.

References

[1] A. Nyberg, A. Eldesokey, D. Bergstrom, and D. Gustafsson, "Unpaired thermal to visible spectrum transfer using adversarial training," in *Proc. ECCVW*, 2018, pp. 657–669. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-11024-6_49

[2] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1143–1152.

[3] M. H. Baig and L. Torresani, "Multiple hypothesis colorization and its application to image compression," *Comput. Vis. Image Understand.*, vol. 164, pp. 111–123, Nov. 2017.

[4] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Jan. 2019.

[5] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, Jul. 2006.

[6] Y. Chen, Y. Luo, Y. Ding, and B. Yu, "Automatic colorization of images from Chinese black and white films based on CNN," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2018, pp. 97–102.

[7] Y. Guo, X. Cao, W. Zhang, and R. Wang, "Fake colorized image detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 1932–1944, Aug. 2018.

[8] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6874–6883.

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

[10] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 351–354.

[11] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1120–1129, May 2006.

[12] L. Xu, Q. Yan, and J. Jia, "A sparse control model for image and video editing," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–10, Nov. 2013.

[13] K. Xu, Y. Li, T. Ju, S.-M. Hu, and T.-Q. Liu, "Efficient affinity-based edit propagation using K-D tree," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–6, Dec. 2009.

[14] R. Zhang *et al.*, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[15] X. Chen, D. Zou, Q. Zhao, and P. Tan, "Manifold preserving edit propagation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–7, Nov. 2012.

[16] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proc. EGSR*, 2007, pp. 309–320.

[17] B. Sheng, H. Sun, M. Magnor, and P. Li, "Video colorization using parallel optimization in feature space," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 407–417, Mar. 2014.

[18] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 298–307, Jan. 2014.

[19] B. Li, F. Zhao, Z. Su, X. Liang, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5188–5202, Nov. 2017.

[20] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 4, pp. 34–41, Jul./Aug. 2001.

[21] F. Fang, T. Wang, T. Zeng, and G. Zhang, "A superpixel-based variational model for image colorization," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 10, pp. 2931–2943, Oct. 2020.

[22] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *Proc. ECCV*, 2008, pp. 126–139.

[23] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Trans. Graphics*, vol. 37, no. 4, p. 47, 2018.

[24] 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)*, 2012, pp. 369–378.

[25] R. Ironi, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Proc. EGSR*, 2005, pp. 201–210.

[26] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2002, pp. 277–280.

[27] Y.-W. Tai, J. Jia, and C.-K. Tang, "Local color transfer via probabilistic segmentation by expectation-maximization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 747–754.

[28] S. Iizuka and E. Simo-Serra, "DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, Nov. 2019.

[29] A. Y.-S. Chia *et al.*, "Semantic colorization with Internet images," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–8, Dec. 2011.

[30] X. Liu *et al.*, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–9, Dec. 2008.

[31] 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. 6837–6845.

[32] A. Royer, A. Kolesnikov, and C. Lampert, "Probabilistic image colorization," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–15.

[33] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. ECCV*, 2016, pp. 577–593.

[34] J. Zhao, L. Liu, C. G. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," in *Proc. BMVC*, 2018, pp. 1–12.

[35] J. Zhao, J. Han, L. Shao, and C. G. Snoek, "Pixelated semantic colorization," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 1–17, 2019.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[37] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. ECCV*, 2016, pp. 649–666.

[38] 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, p. 110, 2016.

[39] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Proc. ECMLPKDD*, 2017, pp. 151–166.

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

[41] 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.

[42] S. Guadarrama, R. Dahl, D. Bieber, J. Shlens, M. Norouzi, and K. Murphy, "PixColor: Pixel recursive colorization," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–17.

[43] C. Lei and Q. Chen, "Fully automatic video colorization with self-regularization and diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3753–3761.

[44] 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.

[45] *Deoldify*. Accessed: Jan. 29, 2020. [Online]. Available: https://github.com/jantic/DeOldify

[46] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[47] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–14.

[49] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[50] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[51] S. Paul, S. Bhattacharya, and S. Gupta, "Spatiotemporal colorization of video using 3D steerable pyramids," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1605–1619, Aug. 2017.

[52] B. Zhang *et al.*, "Deep exemplar-based video colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8052–8061.

[53] H. Bahng *et al.*, "Coloring with words: Guiding image colorization through text-based palette generation," in *Proc. ECCV*, 2018, pp. 431–447.

[54] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2Pix: Line art colorization using text tag with SECat and changing loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9056–9065.

[55] 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. 1–16, Nov. 2019.

[56] S. Messaoud, D. Forsyth, and A. G. Schwing, "Structural consistency and controllability for diverse colorization," in *Proc. ECCV*, 2018, pp. 596–612.

[57] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[58] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[59] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2043–2050.

[60] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[61] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[62] J. Wei and B. Zhong, "Saliency detection using fully convolutional network," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 825–841.

[63] J. Pan *et al.*, "SalGAN: Visual saliency prediction with adversarial networks," in *Proc. CVPRW*, 2017, pp. 1-2.

[64] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[65] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[66] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[67] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

[68] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[69] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[70] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[71] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3534–3545, 2020.

[72] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[73] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. NeurIPS*, 2016, pp. 271–279.

[74] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[75] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.

[76] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NeurIPS*, 2017, pp. 5767–5777.

[77] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. ECCV*, 2018, pp. 768–783.

[78] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[79] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[80] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[81] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[82] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, p. 3.

[83] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, 2018, pp. 1–26.

[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014, pp. 1–15.

[85] 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.

[86] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[87] G. Yue, C. Hou, and T. Zhou, "Blind quality assessment of tone-mapped images considering colorfulness, naturalness, and structure," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3784–3793, May 2019.
[88] K.-Q. Huang, Q. Wang, and Z.-Y. Wu, "Natural color image enhancement and evaluation algorithm based on human visual system," *Comput. Vis. Image Understand.*, vol. 103, no. 1, pp. 52–63, Jul. 2006.
[89] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.
[90] *Colourisesg*. Accessed: Jan. 29, 2020. [Online]. Available: https://colourise.sg
[91] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.
[92] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," *Int. J. Comput. Vis.*, vol. 126, no. 1, pp. 86–110, Jan. 2018.
[93] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.
[94] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[95] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. NeurIPS*, 2006, pp. 155–162.
[96] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NeurIPS*, 2007, pp. 545–552.
[97] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[98] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. CVPR*, Jun. 2011, pp. 433–440.
[99] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.
[100] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, Mar. 2013.
[101] T. V. Nguyen and J. Sepulveda, "Salient object detection via augmented hypotheses," in *Proc. IJCAI*, 2015, pp. 1–7.
[102] T. V. Nguyen, K. Nguyen, and T.-T. Do, "Semantic prior analysis for salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3130–3141, Jun. 2019.

Yuzhi Zhao (Graduate Student Member, IEEE) received the B.Eng. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, City University of Hong Kong. His research interests include image processing, computational photography, and deep learning.

Lai-Man Po (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively.

He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor with the Department of Electrical Engineering. He has authored more than 150 technical journals and conference papers. His research interests include image and video coding with an emphasis deep learning-based computer vision algorithms. He is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairman of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of *HKIE Transactions* in 2011 to 2013. He also served on the Organizing Committee, of the IEEE International Conference on Acoustics, Speech, and Signal Processing in 2003, and the IEEE International Conference on Image Processing in 2010.

Kwok-Wai Cheung (Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees from the City University of Hong Kong, in 1990, 1994, and 2001, respectively, all in electronic engineering. He worked with Hong Kong Telecom as an Engineer, from 1990 to 1995. He was a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, from 1996 to 2002. He was an Assistant Professor with the Chu Hai College of Higher Education, Hong Kong, from 2002 to 2016. He has been with the School of Communication, The Hang Seng University of Hong Kong, as an Associate Professor, since 2016. His research interests include image processing, machine learning, and social computing.

Wing-Yin Yu received the B.Eng. degree in information engineering from the City University of Hong Kong, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. His research interests include deep learning and computer vision.

Yasar Abbas Ur Rehman (Member, IEEE) received the B.Sc. degree in electrical engineering (telecommunication) from the City University of Science and Information Technology, Peshawar, Pakistan, in 2012, the M.Sc. degree in electrical engineering from the National University of Computer and Emerging Sciences, Pakistan, in 2015, and the Ph.D. degree in electrical engineering from the City University of Hong Kong, Hong Kong, in 2019. He is currently working with TCL Corporate Research (HK) Co., Ltd., as a Post-Doctoral Researcher. His research interests include the computer vision, machine learning, deep learning and its applications in facial recognition, biometric anti-spoofing, and video understanding.