IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 1, 2020, accepted May 7, 2020, date of publication May 12, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994148

Invertible Grayscale via Dual Features Ensemble

TAIZHONG YE[1], YONG DU[2], JUNJIE DENG[1], AND SHENGFENG HE[1], (Member, IEEE)
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China
[2]Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China

Corresponding author: Shengfeng He (hesfe@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972162 and Grant 61702194, in part by the Guangzhou Key Industrial Technology Research Fund under Grant 201802010036, and in part by the CCF-Tencent Open Research Fund (CCF-Tencent) under Grant RAGR20190112.

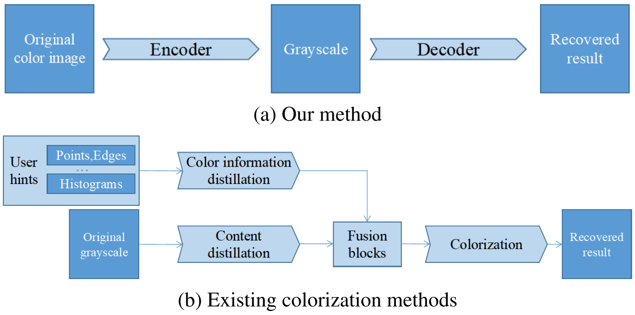
ABSTRACT Grayscale image colorization is known as an ill-posed problem because of the imbalanced matching between intensity and color values. Even given prior hints about the original color image, existing colorization methods cannot recover the original color image from grayscale faithfully. In this paper, we propose to embed color information into an invertible grayscale, such that it can be easily recovered to the original color. However, a vanilla encoding-decoding network cannot produce rich representations of color information and thus the reconstruction quality is limited. Moreover, due to the neglect of the discrimination of color information, it cannot embed color information into visually inconspicuous patterns located in the grayscale. In this paper, we propose a novel color-encoding schema, dual features ensemble network (DFENet), for the effective embedding and faithfully reconstruction. In particular, we complement the residual representations with dense representations, to integrate the ability of local residual learning and local feature fusion. Furthermore, we propose an element-wise self-attention mechanism that highlights the discriminative features and suppresses the redundant ones generated from the dual path module. Extensive experiments demonstrate the proposed method outperforms state-of-the-art methods in terms of reconstruction quality as well as the similarity between the generated invertible grayscale and its groundtruth.

INDEX TERMS Decolorization, colorization, dual features ensemble, convolutional neural network.

I. INTRODUCTION

Color-to-gray conversion is widely applied to aesthetic stylization, monochrome printing and so on. However, the converted grayscale cannot be recovered back to its original color image due to color information loss during channel reduction. Existing colorization methods either learning from a large amount of data [1] or introduce additional priors (*e.g.*, user strokes [2]), they cannot recover the original color faithfully. Data-driven colorization learns mapping from grayscale to color image directly, but this is a one-to-many mapping, the learned mapping function can never recover the original colors. On the other hand, leveraging external complemental information is too sparse for faithful color restoration.

Instead of applying external information, we convert colors into internal knowledge within the grayscale image. Intuitively, we utilize a encoder-decoder structure, to embed color information in the grayscale image, referred as invertible grayscale, while reconstructing the original colors using a decoder. We summarize the differences between our approach and existing colorization methods in Fig. 1. Though Xia *et al.* [3] use a vanilla U-net [4] for the same purpose as ours, it shows limited embedding and reconstruction performances.



(a) Our method

(b) Existing colorization methods

FIGURE 1. Different pipelines of the proposed method and existing approaches. Traditional methods need to introduce various kinds of user-guided hints, while our method automatically generates an invertible grayscale containing color information for reconstruction.

To enrich the discriminative representations of the embedding patterns, we propose a novel unified networks to implement invertible color-to-gray conversion, referred as dual

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

features ensemble network (DFENet). The proposed DFENet consists of a sequence of dual features ensemble blocks (DFEBs) that are designed to fully explore image features. Each DFEB is divided into two modules: dual path module and ensemble inference module. To distill features with affluent expressions of color information, the former one is composed of a series of assembled residual blocks and dense blocks in a dual path manner. While the latter is introduced to investigate and exploit the implicit discriminative correlation between the generated features via a combined spatial-wise and channel-wise attention strategy. Such a design allows to highlight the discriminative features and suppress the redundant ones generated from the dual path module, which is more effective for the colorization of invertible grayscale. Meanwhile it could help to produce a more natural grayscale that owns unrecognizable differences with ground truth grayscale. Extensive experiments over a large amount of images demonstrate that the proposed DFENet outperforms state-of-the-arts in terms of the recovered image quality, as well as the similarity between grayscale and its groundtruth.

The main contributions of our work can be summarized as follows:

- We propose a novel encoder-decoder network, *i.e.*, dual features ensemble network, for effectively implementing an invertible color-to-gray conversion. It is able to learn an invertible grayscale with rich and discriminative color information in the encoding stage, while faithfully recovering its original colors during decoding.
- We present to incorporate a dual features ensemble block in our DFENet, which is constructed by a dual path module and an ensemble inference module. By assembling both types of residual blocks and dense blocks in a dual path, the dual path module enables the richer color representations embedding and decoding. On the other hand, the ensemble inference module is conducive to highlight important integrated features while suppressing redundant ones, and thus assures the similarity between the generated grayscale and its groundtruth. Such a design makes the invertible grayscale more feasible in practical applications.
- Extensive experiments conducted on a large amount of images have validated the superiority of the proposed approach comparing to the state-of-the-art methods.

II. RELATED WORK

A. DECOLORIZATION

Decolorization converts color images to grayscale and it is used for many applications, such as monochrome printing, single channel image processing and stylization. Early color-to-gray methods simply generate grayscale images by acquiring the lightness channel in the specific color space (*e.g.*, CIELab color gamut, YIQ color gamut) or obtaining gray values from linear computation in RGB color space. However, these simple methods cannot precisely reflect image details. Recent methods enhance the contrast in either local-level [5] or global-level [6] features using various measures, such as high-frequency chromatic components [7], consistent gradient field notion [8], saliency regions [9], color orders with respect to the visual context [10], perceptually important features [11] [8], and gradient correlation [12]. Ancuti *et al.* [13] propose a multi-scale approach to minimize artifacts caused by the weight maps. In order to suppress the artifacts introduced by local contrast conservation process, Liu and Zhang [14] incorporate a local feature network to focus on local semantic features. In global-level contrast conservation process, [14], [15] and [16] propose to save computational costs via specific designed optimization. But none of them are invertible such that cannot be well recovered to the original colors as we do.

B. COLORIZATION

Colorization has been studied for decades, which can be categorized into two classes, user-guided colorization and automatic colorization. User-guided colorization uses hints such as scribbles [17] or histograms [18], aiming at controlling the generated color images. However, sparse scribbles cannot produce vivid colors while histograms cannot produce semantically correct colors. On the other hand, data-driven methods free users from tedious annotations and obtain a unified paradigm of colorization by learning parametric mapping from the grayscale to color image. However, as colorization is an imbalanced one-to-many problem, the generated colors tend to be average without enough diversity [1]. Liu and Zhang [19] propose a matching approach to align features from both grayscale and referred color image, achieving correct color transfer in corresponding regions. Unlike traditional colorization method, Xia *et al.* [3] propose an invertible grayscale method, such that the generated grayscale image can be easily converted back with the original colors. However, they use a vanilla U-net for embedding and decoding, preventing both the invertible grayscale and the reconstructed color image from a high image quality. We address this problem by proposing a dual features ensemble network, achieving a high embedding and decoding performance.

III. APPROACH

In this section, we propose a novel method named DFENet, which is devoted to generate a grayscale image that can efficiently recover its original colors. Our framework involves a fully convolutional encoder-decoder network with color consistency. The overall architecture of the proposed DFENet is illustrated in Fig. 2. In order to gain more informative feature representations as well as considering the nonuniform distribution of those features for generating more preferable grayscale and high quality results, we particularly design the dual features ensemble block (DFEB) as the basic unit in our network architecture. To begin with, we elaborate the network architecture in Section III-A. Then, we present the DFEB detailly in Section III-B. And the optimization function of the whole system is discussed in Section III-C.

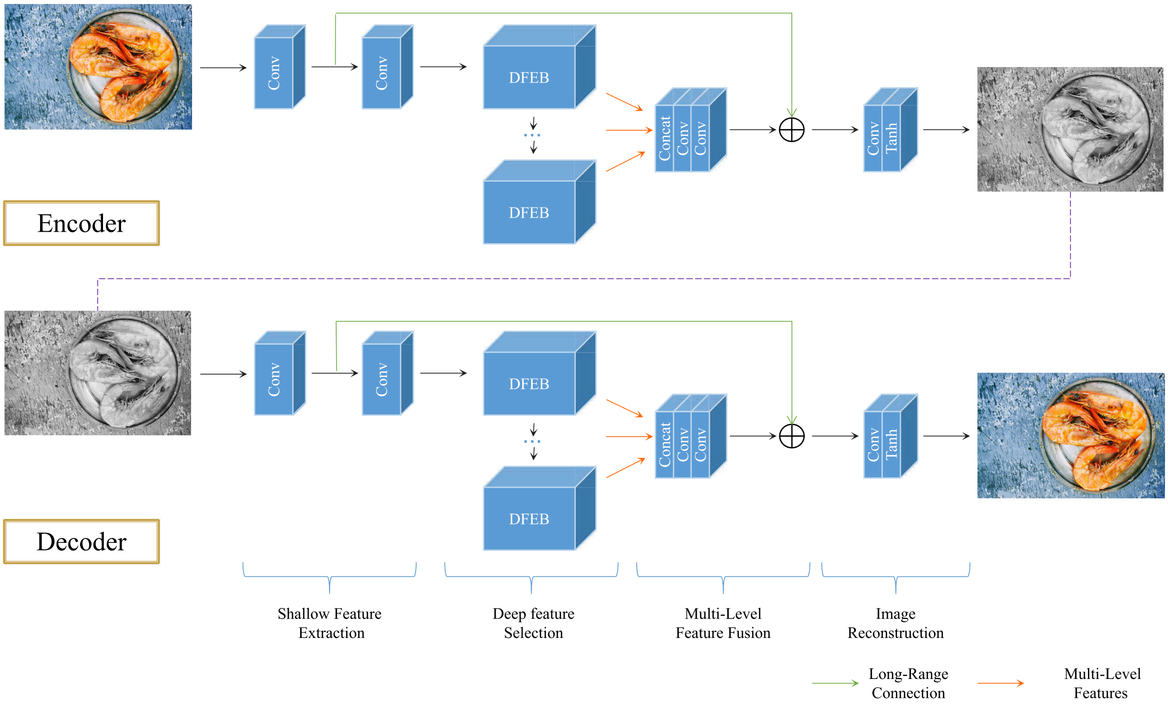


FIGURE 2. The architecture of the proposed dual features ensemble network.

A. NETWORK ARCHITECTURE

Our network DFENet involves an encoder-decoder structure that combines the decolorization and colorization procedures in a closed loop. Given an input color image $\boldsymbol{I}$ in RGB color space, the encoder network aims at generating an invertible grayscale image $\boldsymbol{G}$. Regarding the decoder network, it attempts to recover the original colors of grayscale $\boldsymbol{G}$ and the reconstruction is denoted as $\boldsymbol{O}$. We divide each of such subnetworks into four stages, that is: 1) Shallow feature extraction; 2) Dual feature selection; 3) Multi-level feature fusion; 4) Image reconstruction. Note that the encoder and decoder adopt a similar architecture except for the channel numbers of the entrance and exit of each subnetwork, *i.e.*, color image is corresponded to 3, while grayscale is corresponded to 1. Hence we mainly discuss the architecture of our encoder for the sake of simplicity.

1) SHALLOW FEATURE EXTRACTION

As illustrated in Fig. 2, in the very beginning of our encoder, two flat convolutional layers with a filter size of 3 × 3 are leveraged to extract shallow features of the input color image $\boldsymbol{I}$, which is formulated as follows:

$$\boldsymbol{F}^1_{\text{conv}} = E^1_{\text{conv}}(\boldsymbol{I}), \tag{1}$$

where $E^1_{\text{conv}}(\cdot)$ indicates the first flat convolutional layer in our encoder. Note that we link the shallow features $\boldsymbol{F}^1_{\text{conv}}$ with the layer next to the output layer of encoder network. Such a long-range connection design could facilitate the propagation of low-level information, and benefits the optimization procedure from residual learning as well.

Then, the shallow features $\boldsymbol{F}^1_{\text{conv}}$ is fed into the second flat convolutional layer $E^2_{\text{conv}}(\cdot)$, that is

$$\boldsymbol{F}^2_{\text{conv}} = E^2_{\text{conv}}(\boldsymbol{F}^1_E). \tag{2}$$

$\boldsymbol{F}^2_{\text{conv}}$ is utilized as the input to subsequent encoding layers.

2) DUAL FEATURE SELECTION

The generated shallow features would then be fed into several dual features ensemble blocks (DFEBs). In each DFEB, we assemble several residual blocks [20] and dense blocks [21], and highlight more discriminative features via a self-contained attention mechanism. The detailed architecture of the proposed DFEB will be explained in Section III-B. The benefits of the dual feature selection stage are two-folds: first, the dual path design facilitates the re-exploration and re-usage of the features; Second, the attention mechanism provides a more reasonable distribution of the features such that the redundant features are suppressed and the discriminative features can be better exploited.

3) MULTI-LEVEL FEATURE FUSION

During this stage, we fuse the multi-level features generated from different DFEBs for the reconstruction of the final results. Compared with single-level features, such a strategy could help to excavate a more comprehensive feature representation and thus improve the performance of our approach. Specifically, those hierarchical features are first concatenated and then fed into two consecutive convolutional layers $E^3_{\text{conv}}(\cdot)$, $E^4_{\text{conv}}(\cdot)$, respectively with filter sizes

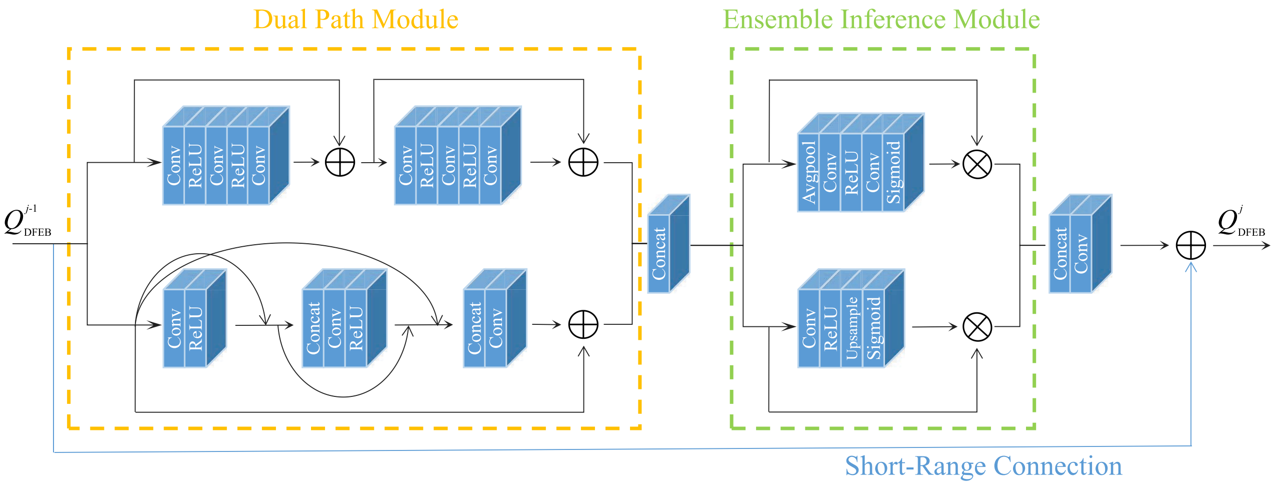


FIGURE 3. Illustration of the proposed dual features ensemble block.

of 1 × 1, 3 × 3. That is,

$$\boldsymbol{F}_{\text{fused}} = E^4_{\text{conv}}(E^3_{\text{conv}}([F^1_{\text{DFEB}}, \cdots, F^J_{\text{DFEB}}])) + \boldsymbol{F}^1_{\text{conv}}, \quad (3)$$

where $[\cdot]$ denotes concatenation operation, and J indicates the total number of DEFBs.

4) IMAGE RECONSTRUCTION

Once the fused features are obtained, the encoder network would generate the grayscale $\boldsymbol{G}$ (or colorize the grayscale in regard to decoder) by the following formation:

$$\boldsymbol{G} = \text{Tanh}(E^5_{\text{conv}}(\boldsymbol{F}_{\text{fused}})). \quad (4)$$

B. DUAL FEATURES ENSEMBLE BLOCK

Conventional encoder-decoder network are extensively utilized in many image processing tasks and usually constructed in a single path topology. On the contrary, we are interested in exploring a different network topology for a better feature representation. Inspired by ensemble learning, we investigate a dual path design of basic blocks in our model, namely Dual Features Ensemble Block (DFEB). Each DFEB is consist of two modules, *i.e.*, dual path module and ensemble inference module. The former one is constructed by two types of blocks, that is, residual blocks and dense blocks, while the latter is utilized for the inference of the ensembled features. The detailed structure of the proposed DFEB is illustrated in Fig. 3. And the output of the jth DFEB is formulated as follows:

$$\boldsymbol{F}^j_{\text{DFEB}} = \boldsymbol{A} \cdot [\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}] + \boldsymbol{F}^{j-1}_{\text{DFEB}}, \quad (5)$$

where Q is the total number of the adopted residual blocks, K is the number of dense blocks. Thus $\boldsymbol{F}^Q_{\text{res}}$ and $\boldsymbol{F}^K_{\text{dense}}$ denote the feature maps generated from the last residual block and dense block in the DFEB respectively. Note that when $j = 1$, $\boldsymbol{F}^0_{\text{DFEB}}$ indicates the feature maps $\boldsymbol{F}^2_{\text{conv}}$. And $\boldsymbol{A}$ is generated by our attention mechanism.

From a perspective of feature representation, the superiority of residual blocks is that achieves a reuse of the preceding low-level features, due to a residual connection. Dense blocks, on the other hand, could keep exploring novel features because of the dense connection. To exploit different advantages of these two types of blocks, we attempt to arrange them in a single DFEB via a dual path manner.

Instead of simply concatenating the generated feature maps, we subsequently consider to delve into a better distribution of them for improving the performance. Specifically, we introduce an attention mechanism into the DFEB for better evaluating and exploiting the different importance of the concatenated feature maps. The proposed attention technique contains two parts: spatial-wise attention and channel-wise attention. And Eq. (5) can be reformulated as

$$\boldsymbol{F}^j_{\text{DFEB}} = \text{Conv}_{1\times1}[\boldsymbol{A}_{\text{s}} \cdot [\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}], \boldsymbol{A}_{\text{c}} \cdot [\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}]] + \boldsymbol{F}^{j-1}_{\text{DFEB}}, \quad (6)$$

where $\boldsymbol{A}_s$ indicates the spatial-wise attention, and $\boldsymbol{A}_c$ indicates the channel-wise attention. Regarding the calculation of spatial-wise attention, the concatenated feature maps $[\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}]$ are first fed into a stride-2 convolutional layer with a filter size of 3 × 3, and then upsampled by a corresponded deconvolutional layer for a size match, which is formulated as follows:

$$\boldsymbol{A}_s = \text{Sigmoid}(\text{Deconv}(\text{ReLU}(\text{Conv}_{3\times3}([\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}])))). \quad (7)$$

Note that for alleviating the computational burden, we use a combination of stride-2 convolutional and deconvolutional layers here rather than a general choice of several stride-1 convolutional layers to obtain the spatial-wise attention. Applying spatial attention to deconvlutional layers can further regularize the features response in the upsampling process.

To aggregate the spatial information and gain the channel-wise attention $\boldsymbol{A}_c$, we first use average pooling to shrink each channel of the feature maps $[\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}]$ along the spatial

TABLE 1. Ablation study with respect to different components of the proposed dual features ensemble block.

Structure		1	2	3	4	5	6	7
Dual path	ResBlock	√	×	√	√	√	√	√
	DenseBlock	×	√	√	√	√	√	√
Attention mechanism	Channel Att.	×	×	×	√	×	√	√
	Spatial Att. (D)	×	×	×	×	√	√	×
	Spatial Att. (C)	×	×	×	×	×	×	√
Recovered Color Image		30.575/0.969	34.658/0.977	35.815/0.980	36.091/0.982	38.327/0.987	38.488/0.988	37.821/0.986
Generated Gray Image		36.477/0.977	36.283/0.978	32.478/0.951	37.860/0.986	40.082/0.991	40.296/0.992	38.327/0.988

dimension, and then fed it into two consecutive convolutional layers. That can be written as

$$\boldsymbol{A}_c = \text{Sigmoid}(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(\text{Avgpool} \\ ([\boldsymbol{F}^Q_{\text{res}}, \boldsymbol{F}^K_{\text{dense}}]))))). \tag{8}$$

C. OPTIMIZATION FUNCTION

The whole optimization function is defined by three parts, that is, luminance consistency loss $\mathcal{L}_l$, color consistency loss $\mathcal{L}_c$ and perceptual loss $\mathcal{L}_p$. $\mathcal{L}_l$ forces a similarity between the generated grayscale and the groundtruth one. $\mathcal{L}_c$ constrains the color accuracy. And $\mathcal{L}_P$ serves for ensuring a better visual perception of the colorized results. Hence, the overall objective $\mathcal{L}$ is formulated as

$$\mathcal{L} = \lambda_1\mathcal{L}_l + \lambda_2\mathcal{L}_c + \lambda_3\mathcal{L}_p, \tag{9}$$

where λ_1, λ_2, λ_3 respectively denote the balance factors.

1) LUMINANCE CONSISTENCY LOSS

We first consider that the invertible grayscale should be highly similar to the groundtruth. In this case, the embedded color information are more unrecognizable, and thus be more feasible for practical applications. Given a groundtruth luminance image $\boldsymbol{Y}$, we introduce a luminance consistency loss to confine the generated grayscale $\boldsymbol{G}$, which is formulated as follows:

$$\mathcal{L}_l = \frac{1}{M}\sum_{i=1}^{M} \|\boldsymbol{G}^{(i)} - \boldsymbol{Y}^{(i)}\|_1, \tag{10}$$

where $\boldsymbol{G}^{(i)}$ denotes the ith generated grayscale image, and $\boldsymbol{Y}^{(i)}$ denotes the corresponded groundtruth. M indicates the total number of the training samples, $\|\cdot\|_1$ represents the L_1 norm. Note that we obtain the groundtruth $\boldsymbol{Y}$ by extracting the luminance channel from the corresponded color image in a CIELab color space.

2) COLOR CONSISTENCY LOSS

To faithfully recover the original colors, we then utilize a color consistency loss in our objective that constrains the restored results $\boldsymbol{O}$ via utilizing the input color image $\boldsymbol{I}$, that is

$$\mathcal{L}_c = \frac{1}{M}\sum_{i=1}^{M} \|\boldsymbol{O}^{(i)} - \boldsymbol{I}^{(i)}\|_1. \tag{11}$$

Such a pixel-wise similarity constraint can effectively improve the color accuracy of the outputs.

3) PERCEPTUAL LOSS

Another problem we concern about is that just keeping a pixel-wise consistency could not guarantee a fine visual perception of the recovered results. To exploit the visually important information for a better reconstruction, we also adopt a perceptual loss [22] $\mathcal{L}_p$, which is defined by

$$\mathcal{L}_p = \frac{1}{M}\sum_{i=1}^{M} \|V(\boldsymbol{O}^{(i)}) - V(\boldsymbol{I}^{(i)})\|_F, \tag{12}$$

where $V(\cdot)$ indicates the specified layers of VGG [23] network, and $\|\cdot\|_F$ denotes the Frobenius norm.

D. IMPLEMENTATION DETAILS

In the proposed DFEB (see Fig. 3), three convolutional layers followed by the ReLU activation function are utilized to construct each dense block. And the cascaded dense blocks that considered multi-level information are beneficial for feature reuse. In regard to each residual block, we leverage three bottleneck-based residual layers [20] to refine features. The structure of the decoder is similar to the encoder, except for the number of input and output channels. As a result, when the generated grayscale is fed into the decoder, the number of the input channel is set to 1 whereas that of output channels is set to 3.

IV. EXPERIMENTS

A. TRAINING DETAILS

We implement the proposed method in Pytorch on a PC with a Nvidia Geforce GTX 1080Ti GPU and a Intel(R) Core(TM) i7-6859K CPU @ 3.60GHZ. We use the Pascal VOC 2012 dataset [24] for training and testing following [3]. This dataset contains 17125 color images with different contents. We divide the dataset randomly into training part containing 13758 color images, and the remained images are used for testing. Input images are resized to 256 × 256 × 3 resolution during training. However, arbitrary sizes of images can be processed during testing.

In our experiments, we train the proposed network from scratch with a batch size of 8 in 200 epochs, and the Adam solver is adopted to optimize model. The learning rate is set to

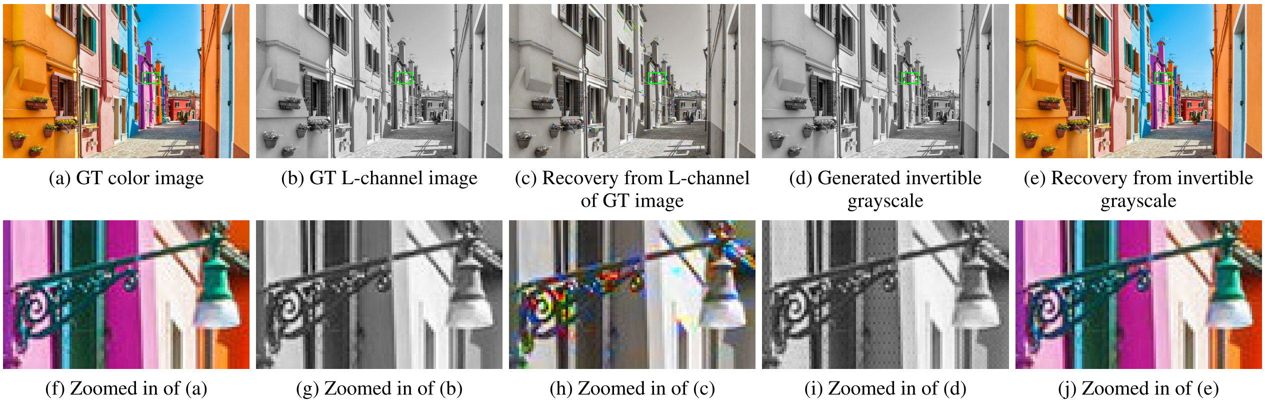


FIGURE 4. Evaluation on the generated invertible grayscale. Our method encodes color information into grayscale image with inconspicuous patterns (i) (comparing to (g)), while it can recover faithful colors (e). Feeding the raw grayscale image to the network generates random colors (c).

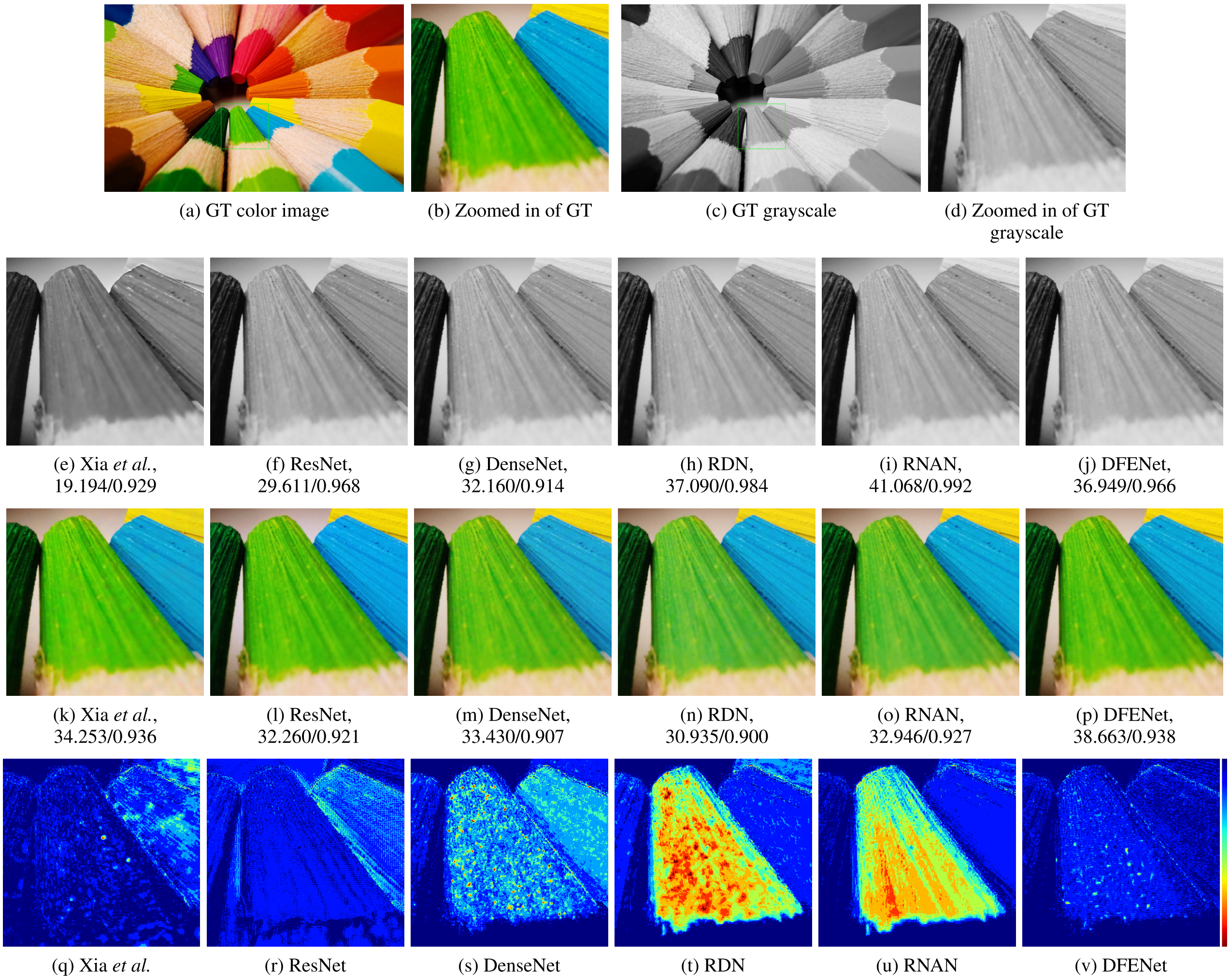


FIGURE 5. Qualitative comparisons with state-of-the-arts in color-encoding patterns and recovered image quality. The second row shows the grayscale results, the third row shows the reconstructions, and the last row displays the different maps. PSNR/SSIM are shown in subtitles.

0.0001 and unchanged in the first 100 epochs. Then it decays to zero linearly in the following 100 epochs. This setting contributes to find the sub-optimal solution space fast at the beginning and then fine-tunes the parameters in a smaller step. For the hyper-parameters in the loss function, we empirically set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in all of our experiments.

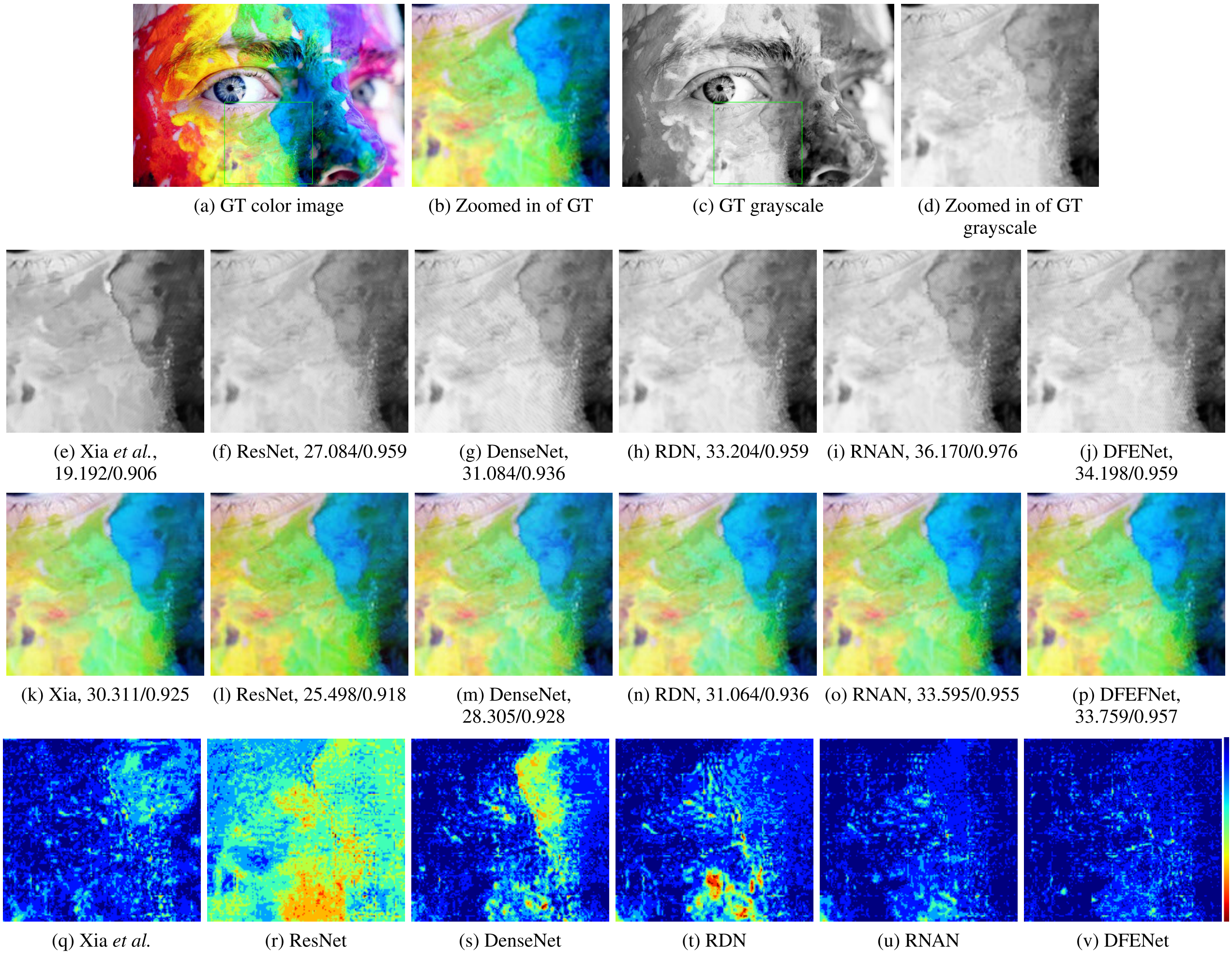


FIGURE 6. **Qualitative comparisons with state-of-the-arts in color-encoding patterns and recovered image quality. The second row shows the grayscale results, the third row shows the reconstructions, and the last row displays the different maps. PSNR/SSIM are shown in subtitles.**

In practice, we observe that the value of $\mathcal{L}_p$ is about one order of magnitude higher than the ones of $\mathcal{L}_c$ and $\mathcal{L}_l$. As a result, the optimization of the whole system is more influenced by $\mathcal{L}_p$ under this parameter setting. This implies the importance of a fine visual perception of the reconstructions.

B. METRICS

To measure the similarity between two images, three metrics are widely used: mean absolute error (MAE), peak signal to noise ratio (PSNR), structural similarity (SSIM). These three metrics are used for quantitative evaluations between the recovered color image and original color image (color consistency), and between the generated grayscale image and L-channel of the original image (luminance consistency).

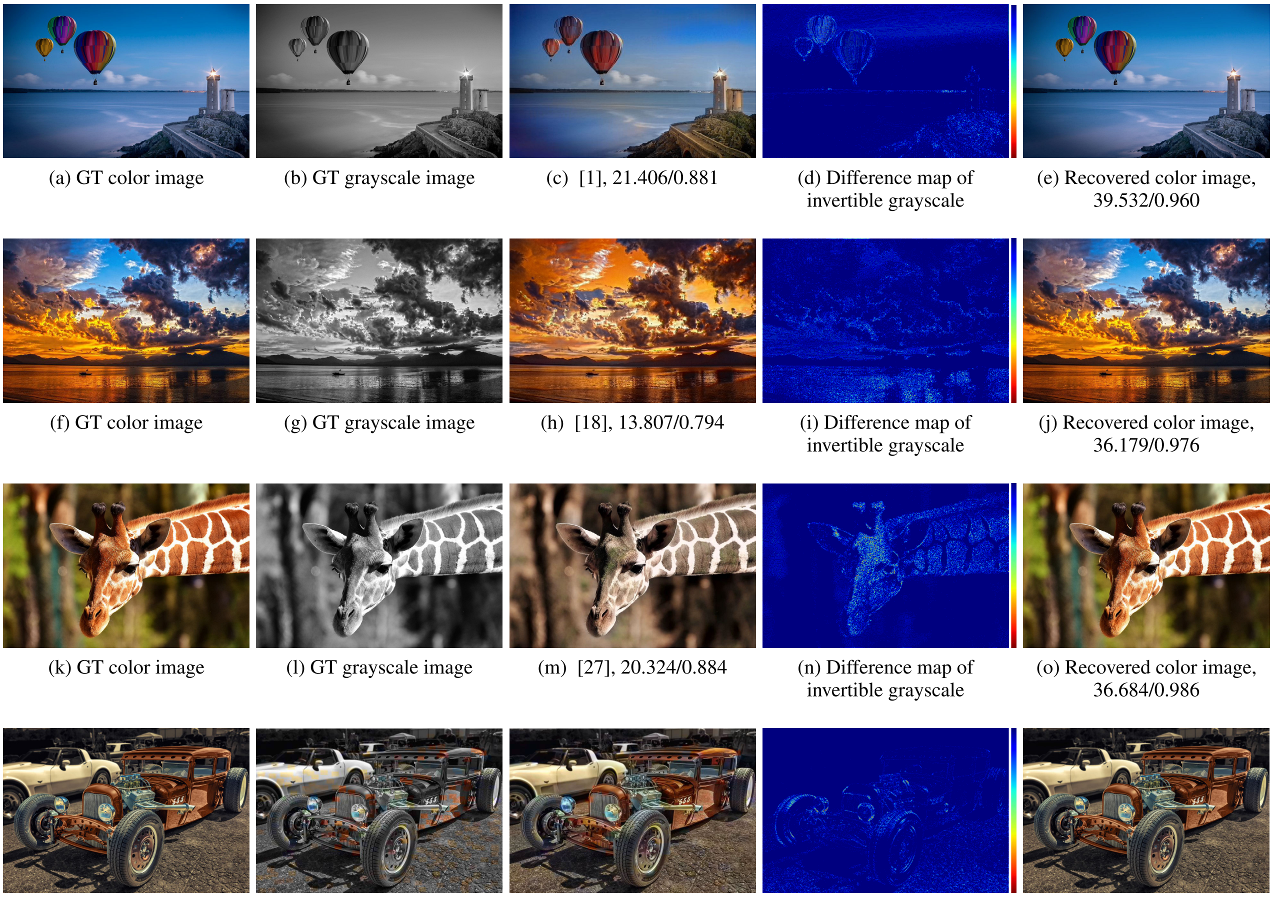
C. ABLATION STUDY

1) DUAL FEATURES ENSEMBLE BLOCK ANALYSIS

To demonstrate the effectiveness of our DFEB, we break down the proposed DEFB into different components, the ResBlock and DenseBlock in the dual path, and the spatial- and channel-wise attention modules. We compare different combinations of these components in Table 1. We can see that ResBlock and DenseBlock show different responses for the task of color recovery, and particularly ResBlock performs poorly in this task. However, our dual path ensemble demonstrates superior performance over the individual path. By introducing attention mechanism, structures #4, #5, #6, #7 obtain higher scores in PSNR and SSIM. Both two attention mechanisms show effectiveness to luminance and color consistencies. Note that in Section III-B (Eq. (7)) we apply spatial attention in deconvolutional layer rather than convolutional ones to regularize the features response in the upsampling process. We compare to the traditional spatial attention on the convolutional layers (structure #7), and our design indeed brings superior performance (structure #6) over the traditional design as regularizing the upsampling process is vital for image reconstruction.

2) OBJECTIVE FUNCTION ANALYSIS

We further examine the contributions of different loss functions. As shown in Table 2, adding the color consistency loss


(a) GT color image

(b) GT grayscale image

(c) [1], 21.406/0.881

(d) Difference map of invertible grayscale

(e) Recovered color image, 39.532/0.960

(f) GT color image

(g) GT grayscale image

(h) [18], 13.807/0.794

(i) Difference map of invertible grayscale

(j) Recovered color image, 36.179/0.976

(k) GT color image

(l) GT grayscale image

(m) [27], 20.324/0.884

(n) Difference map of invertible grayscale

(o) Recovered color image, 36.684/0.986

(p) GT color image

(q) Input image with hints

(r) [17], 28.350/0.986

(s) Difference map of invertible grayscale

(t) Recovered color image, 39.757/0.994

FIGURE 7. Comparisons with user-guided and automatic colorization methods. [1], [18] and [27] are data-driven colorization methods, while [17] involves user-guided hints. Both methods cannot restore the original colors faithfully as we do. Image (d), image (i), image (n) and image (s) respectively visualize indistinguishable embedding patterns of our invertible grayscale.

TABLE 2. Ablation study with respect to different loss functions.

Method	Recovered Color Image		Generated Gray Image	
	PSNR	SSIM	PSNR	SSIM
$\mathcal{L}_c$	39.741	0.991	9.130	0.098
$\mathcal{L}_c + \mathcal{L}_p$	39.034	0.989	14.043	0.626
$\mathcal{L}_c + \mathcal{L}_l$	37.004	0.984	48.534	0.999
$\mathcal{L}_c + \mathcal{L}_p + \mathcal{L}_l$	38.488	0.988	40.296	0.992

$\mathcal{L}_c$ only leads to the best recovery result, since it is a simple reconstruction task that ignores the intermediate grayscale image. By adding a luminance consistency loss $\mathcal{L}_l$, the generated grayscale achieves a high similarity to the GT luminance, while the performance of the recovery drops by 2.7dB of PSNR. This implies that ensuring a proper intermediate form introduces difficulties in the color recovery. The perceptual loss $\mathcal{L}_p$ helps achieving a good balance between color consistency and luminance consistency. All these results also demonstrate the effectiveness of our network design, and we can perform well without complex loss functions.

3) EMBEDDING AND RECOVERY ANALYSIS

Our model embeds color information into inconspicuous patterns within the invertible grayscale. As shown in Fig. 4, the embedding patterns show grid-like structure that almost invisible to human. Furthermore, the recovered results are very similar to the original inputs. To demonstrate the unique properties of our embedding patterns, we also feed the original luminance channel to the network. Fig. 4 (h) shows that given a clean grayscale, the proposed network will assign random colors according to the grayscale distributions.

D. COMPARISON WITH STATE-OF-THE-ARTS

The work of Xia *et al.* [3] is the only one that shares the same spirit to ours for generating invertible grayscale. This work uses a vanilla U-net [4], and therefore cannot achieve a high quality of color information embedding and recovery. To better evaluate the proposed method, we further compare to advanced feature extraction convolutional block designs, *i.e.*, ResNet [20], DenseNet [21], RDN [25], and RNAN [26]

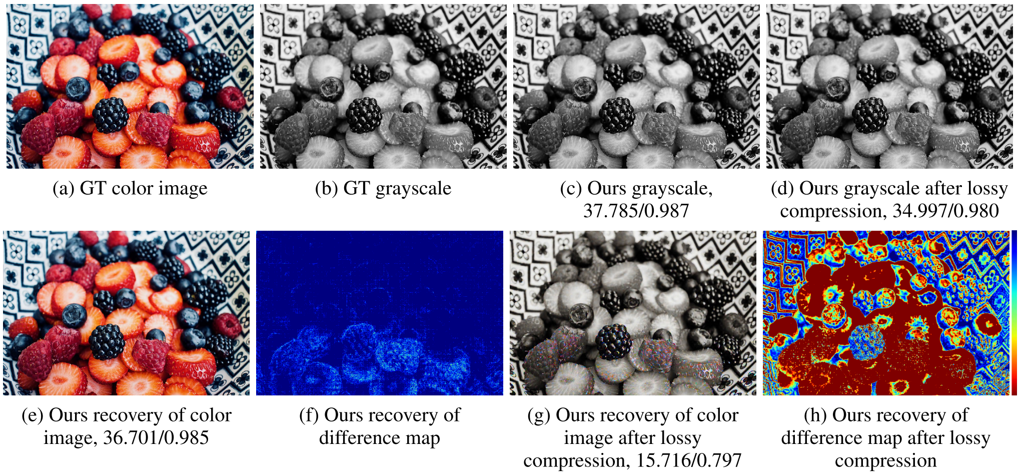

(a) GT color image
(b) GT grayscale
(c) Ours grayscale, 37.785/0.987
(d) Ours grayscale after lossy compression, 34.997/0.980
(e) Ours recovery of color image, 36.701/0.985
(f) Ours recovery of difference map
(g) Ours recovery of color image after lossy compression, 15.716/0.797
(h) Ours recovery of difference map after lossy compression

FIGURE 8. **Applying lossy compression (JPEG compression with the quality rate of 70% is used in this example) on our generated grayscale will destroy the embedding patterns.**

TABLE 3. **Quantitative comparison with state-of-the-art invertible grayscale and other advanced feature extraction convolutional block designs.**

Method	Recovered Color Image		Generated Gray Image	
	PSNR	SSIM	PSNR	SSIM
Xia *et al.* [3]	36.323	0.984	33.007	0.978
ResNet [20]	30.726	0.970	30.187	0.969
DenseNet [21]	37.935	0.987	37.707	0.986
RDN [25]	37.897	0.986	38.928	0.989
RNAN [26]	38.014	0.987	37.875	0.987
Ours	**38.488**	**0.988**	**40.296**	**0.992**

by using their blocks in the U-net structure. Quantitative comparisons of these models are shown in Table 3. We can see that the vanilla U-net used in Xia *et al.* [3] cannot recover the color image well, and they also cannot embed color information inconspicuously. Regarding ResNet, similar observation of our ablation study can be found in here, demonstrating that merely utilizing residual learning is not suitable for color information embedding. DenseNet, RDN and RNAN achieve better color consistency and luminance consistency than the former two models. However, the proposed method achieves superior performance against all these competitors. Note that RNAN gains a better grayscale but a worse reconstruction in contrast to DFENet, which implies a lack of the color information in the generated grayscale. Qualitative comparisons are shown in Fig. 5 and 6. We can see that the proposed method can achieve the least noticeable embedding patterns, meanwhile the highest reconstruction quality.

E. COMPARISON WITH TRADITIONAL COLORIZATION

We also compare to traditional colorization methods. Fig. 7 shows the comparisons of user-guided and automatic colorization methods. Due to the one-to-many nature of automatic colorization, existing methods [1], [18] and [27] tend to produce average or over-saturated colors, which obviously cannot satisfy the requirement of users. On the other hand, user-guided method relies heavily on accurate and dense user strokes. And it can be observed that though given an input image with rather dense user hints [17], the result is still unsatisfied. In contrast, the proposed method can achieve a faithful color recovery.

F. LIMITATIONS

Despite of the effectiveness of the proposed DFENet validated in the above experiments, it is limited in recovering color image from a lossy compressed grayscale. That is, the generated grayscale cannot tolerate information damages caused by lossy image compression methods such as JPEG. Given a grayscale generated by our encoder, we first compress it with JPEG, and then feed the compressed grayscale into our decoder, the visual results are shown in Fig. 8. It can be observed that the reconstruction of the compressed grayscale (see Fig. 8 (g)) suffers a lower recovery quality. This may be concluded to the importance of the embedding inconspicuous patterns in the grayscale for colorization, and thus could not be damaged. In the future, we will explore the potential improvements of our methods, in regard to the resistance of the noises introduced by lossy compression strategies.

V. CONCLUSION

In this paper, we propose an invertible color-gray conversion method. In order to capture rich color and contextual representations, we propose a novel color-encoding schema, dual features ensemble network (DFENet). Specifically, we integrate the residual representations with dense representations, extracting features in the way of residual learning and local feature fusion. Furthermore, we present an element-wise self-attention mechanism that highlights the discriminative features in both downsampling and upsampling processes. Extensive experiments demonstrate the proposed method achieves superior performance against state-of-the-art methods in terms of recovery of color consistency and embedding luminance consistency.

ACKNOWLEDGMENT

(Taizhong Ye and Yong Du are co-first authors.)

REFERENCES

[1] R. Zhang, P. Isola, and A. A. Efros, ''Colorful image colorization,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 649–666.

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

[3] M. Xia, X. Liu, and T.-T. Wong, ''Invertible grayscale,'' *ACM Trans. Graph.*, vol. 37, no. 6, p. 246, Jan. 2019.

[4] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[5] A. A. Gooch, S. C. Olsen, J. Tumblin, and B. Gooch, ''Color2Gray: Salience-preserving color removal,'' *ACM Trans. Graph.*, vol. 24, no. 3, pp. 634–639, 2005.

[6] J. G. Kuk, J. H. Ahn, and N. I. Cho, ''A color to grayscale conversion considering local and global contrast,'' in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 513–524.

[7] R. Bala and R. Eschbach, ''Spatial color-to-grayscale transform preserving chrominance edge information,'' in *Proc. Color Imag. Conf.*, vol. 2004, no. 1. Springfield, VA, USA: Society for Imaging Science and Technology, 2004, pp. 82–86.

[8] L. Neumann, M. Čadik, and A. Nemcsics, ''An efficient perception-based adaptive color to gray transformation,'' in *Proc. 3rd Eurograph. Conf. Comput. Aesthetics Graph., Vis. Imag.* Aire-la-Ville, Switzerland: Eurographics Association, 2007, pp. 73–80.

[9] H. Du, S. He, B. Sheng, L. Ma, and R. W. H. Lau, ''Saliency-guided color-to-gray conversion using region-based optimization,'' *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 434–443, Jan. 2015.

[10] C. Lu, L. Xu, and J. Jia, ''Contrast preserving decolorization with perception-based quality metrics,'' *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 222–239, Nov. 2014.

[11] K. Smith, P.-E. Landes, J. Thollot, and K. Myszkowski, ''Apparent greyscale: A simple and fast conversion to perceptually accurate images and video,'' *Comput. Graph. Forum*, vol. 27, no. 2, pp. 193–200, Apr. 2008.

[12] Q. Liu, P. X. Liu, W. Xie, Y. Wang, and D. Liang, ''GcsDecolor: Gradient correlation similarity for efficient contrast preserving decolorization,'' *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2889–2904, Sep. 2015.

[13] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert, ''Image and video decolorization by fusion,'' in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 79–92.

[14] S. Liu and X. Zhang, ''Image decolorization combining local features and exposure features,'' *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2461–2472, Oct. 2019.

[15] M. Song, D. Tao, C. Chen, X. Li, and C. W. Chen, ''Color to gray: Visual cue preservation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1537–1552, Sep. 2010.

[16] Z. Ji, M.-E. Fang, Y. Wang, and W. Ma, ''Efficient decolorization preserving dominant distinctions,'' *Vis. Comput.*, vol. 32, no. 12, pp. 1621–1631, Dec. 2016.

[17] 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,'' 2017, *arXiv:1705.02999*. [Online]. Available: http://arxiv.org/abs/1705.02999

[18] C. Xiao, C. Han, Z. Zhang, J. Qin, T. Wong, G. Han, and S. He, ''Example-based colourization via dense encoding pyramids,'' *Comput. Graph. Forum*, vol. 39, no. 1, pp. 20–33, Feb. 2020.

[19] S. Liu and X. Zhang, ''Automatic grayscale image colorization using histogram regression,'' *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1673–1681, Oct. 2012.

[20] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[22] J. Johnson, A. Alahi, and L. Fei-Fei, ''Perceptual losses for real-time style transfer and super-resolution,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[23] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ''The Pascal visual object classes (VOC) challenge,'' *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[25] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, ''Residual dense network for image restoration,'' 2018, *arXiv:1812.10477*. [Online]. Available: http://arxiv.org/abs/1812.10477

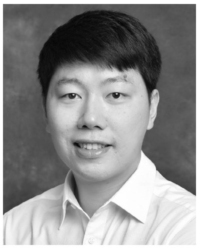
[26] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, ''Residual non-local attention networks for image restoration,'' in *Proc. ICLR*, 2019, pp. 1–18.

[27] 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.

TAIZHONG YE received the B.Sc. degree in communication engineering from South China Normal University, Guangzhou, China, in 2017. He is currently pursuing the M.Sc. degree in computer science and technology with the South China University of Technology. His research interests include image processing and computer vision.

YONG DU received the B.Sc. and M.Sc. degrees from Jiangnan University and the Ph.D. degree from the South China University of Technology. He is currently an Assistant Professor with the Department of Computer Science and Technology, Ocean University of China. His research interests include computer vision and image processing.

JUNJIE DENG received the B.Sc. degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, where he is currently pursuing the M.Sc. degree. His research interests include image processing and computer vision.



SHENGFENG HE (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Macau University of Science and Technology and the Ph.D. degree from the City University of Hong Kong. He was a Research Fellow with the City University of Hong Kong. He is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, computer graphics, and deep learning.

• • •