Zhang W, Fang CW, Li GB. Automatic colorization with improved spatial coherence and boundary localization. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(3): 494–506 May 2017. DOI 10.1007/s11390-017-1739-6

Automatic Colorization with Improved Spatial Coherence and Boundary Localization

Wei Zhang [1], *Member, IEEE*, Chao-Wei Fang [1], and Guan-Bin Li [2,*], *Member, IEEE*

[1] *Department of Computer Science, The University of Hong Kong, Hong Kong, China*
[2] *School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China*

E-mail: wzhang2@cs.hku.hk; chwfang@connect.hku.hk; liguanbin@mail.sysu.edu.cn

Received December 25, 2016; revised February 26, 2017.

Abstract Grayscale image colorization is an important computer graphics problem with a variety of applications. Recent fully automatic colorization methods have made impressive progress by formulating image colorization as a pixel-wise prediction task and utilizing deep convolutional neural networks. Though tremendous improvements have been made, the result of automatic colorization is still far from perfect. Specifically, there still exist common pitfalls in maintaining color consistency in homogeneous regions as well as precisely distinguishing colors near region boundaries. To tackle these problems, we propose a novel fully automatic colorization pipeline which involves a boundary-guided CRF (conditional random field) and a CNN-based color transform as post-processing steps. In addition, as there usually exist multiple plausible colorization proposals for a single image, automatic evaluation for different colorization methods remains a challenging task. We further introduce two novel automatic evaluation schemes to efficiently assess colorization quality in terms of spatial coherence and localization. Comprehensive experiments demonstrate great quality improvement in results of our proposed colorization method under multiple evaluation metrics.

Keywords automatic colorization, deep learning, conditional random field (CRF), color transform, quality evaluation

1 Introduction

Image colorization aims to convert grayscale images into colorful ones. This task has attracted a lot of research in computer graphics due to its practical application values, such as colorizing old photographs and assisting creative work[1-15]. With user-assisted scribbles[1-5] or reference color images[7-14], traditional research mainly focuses on developing better colorization systems with less user interactions and time consumption. Interestingly, human can effortlessly judge suitable colors for different regions just by a quick glance of a grayscale image. In order to make this process possible for a colorization system, recent research focuses on fully-automatic colorization techniques. Ideally, an automatic colorization system takes grayscale images as input and generates visually plausible color versions directly. This problem can be readily formulated as a pixel-wise regression problem in computer version and the effectiveness has been proven by Cheng *et al.*[16] and Dahl[17]. Over the past few months, Larsson *et al.*[18] and Zhang *et al.*[19] both introduced classification-based colorization frameworks built on deep convolutional neural networks (DCNNs). In order to produce colorization results with higher saturation and fidelity, they both learned pixel-wise labeling models over discrete color bins. Furthermore, Zhang *et al.*[19] applied class re-balance during training and achieved the state of the art. Though significant improvement has been made in [18-19], there are obvious drawbacks existing in maintaining color consistency in homogeneous regions as well as in precisely distinguishing colors near region boundaries. For example in Fig.1, Figs.1(a) and 1(d) show the top-1 color predictions produced by [19]. Inconsistent color assignments

Regular Paper
Special Section of CVM 2017
This work was partially supported by Hong Kong Research Grants Council under General Research Funds (HKU17209714).
*Corresponding Author

(a) (b) (c) (d) (e) (f)

Fig.1. Comparison with leading method proposed by Zhang *et al.*[19] (a) (d) Top-1 color predictions results by Zhang *et al.*[19] (b) (e) Final results of Zhang *et al.*[19] by calculating the expectation over color bins. (c) (f) Our results generated by our method which improve spatial consistency using CRF and chromatic resolution by color transform CNN.

can be commonly observed even in simple images with single object and uncomplicated background. Furthermore, the poor localization of object boundaries often leads to color bleeding. Hence, even though the color prediction for major parts of images is correct, people can easily discover unreal colorization from such phenomena. Obviously, such existing shortages reveal that high-quality automatic colorization remains a challenging task.

In this paper, we propose a novel pipeline which aims to improve the spatial coherence and boundary localization in colorization. Fully-connected conditional random field (CRF) has been widely used to improve the spatial accuracy in general DCNN-based image labeling tasks, such as semantic segmentation, in which inputs are usually color images. However, colorization models only take grayscale images as input without any chromatic information, making it extremely hard to capture edge details and local consistency for a fully-connected CRF. To address this problem, we present a boundary-guided local CRF which is capable of improving pixel-wise color labeling results of DCNNs. Since the ultimate goal of colorization is to predict suitable color values for each pixel instead of fixed color category, one essential step at the end is to infer continuous values from labeling results. Current colorization system simply calculates the expectation[19] or takes the median value[18] over histogram bins. However, as shown in Figs.1(b) and 1(e), such method brings no improvement but a blurry effect. In this paper, we introduce a color transform CNN to learn a better inference. As shown in Figs.1(c) and 1(f), our final results achieve significantly improved quality both locally and globally.

We further introduce two novel evaluation schemes to automatically evaluate colorization quality in terms of regional consistency and boundary localization respectively on large datasets.

In summary, our contributions in this paper can be summarized as follows.

- We introduce a novel automatic colorization framework based on DCNN. With a boundary-guided local CRF and a color transform CNN as post-processing steps, our system not only is capable of cap-

turing detailed edge information from grayscale images to facilitate better color labeling, but also learns a mapping from labeling results valued in fixed color bins to final color values.

• We develop two novel schemes for efficient quality evaluations through a large number of automatic colorization results. Our proposed evaluation schemes reflect human's common preference for high-quality colorizations. Experimental results illustrate our colorization method achieves much better performance than previous methods under both evaluation schemes.

2 Related Work

We review recent automatic colorization systems and existing evaluation criteria in this section.

Automatic Colorization. Fully automatic colorization could be formulated as a pixel-wise prediction problem which is targeted at transforming one gray image to its color version. Cheng *et al.*[16] first attacked this problem by exploring a combination of different levels of handcrafted features for each pixel. They attempted to predict chromatic values using neural network with L_2 regression loss. Recently, end-to-end deep CNN features demonstrate considerably superior performance compared with traditional handcrafted ones in extensive vision tasks[20-26]. Hence it is not surprising with state-of-the-art deep CNN architectures, Dahl[17] obtained better results than [16] even though the loss function is still L_2 regression. Iizuka *et al.*[27] generated promising results on scene-centered photographs through taking advantage of a two-steam DCNN architecture where a joint training strategy is used to fuse scene classification cues. More recently, [18-19] further address the underlying label uncertainty problem in automatic color prediction by formulating colorization as a pixel-wise labeling problem instead of regression. Typically, they first divide color space into discrete bins and then trained a DCNN to predict the probability distribution over bins. Their DCNN architectures are close in spirit: Larsson *et al.*[18] used hyper-columns[28] of multiple feature maps while Zhang *et al.*[19] adopted dilated convolutions[29] and layer concatenation. In addition, Zhang *et al.*[19] figured out the problem of extremely unbalanced distribution over color bins through introducing class-rebalancing during training, which helps to boost the performance to the state of the art.

Evaluation Criteria. The ultimate goal of an automatic evaluation is to determine the consistency of the generated colorization results with human expectation. It is a non-trivial task specially because in most cases, there exist multiple reasonable color schemes which appear to be both realistic and vibrant for one grayscale image. Though each grayscale image in the testing set[18-19] has corresponding color version as ground-truth, simply expecting exactly the same colorization results as ground-truth is overly strict. Here we summarise existing evaluation criteria for automatic colorization systems.

Direct pixel-wise comparison:

• *PSNR*: peak signal-to-noise ratio in RGB color space[16,18];

• *RMSE and Raw Accuracy*: root mean square error in a, b color space over all pixels[18-19];

• *Rebalanced Raw Accuracy*: re-weights the raw pixel distance inversely by color class probability[19].

Semantic interpretability:

• *Image Classification*: feeds automatically colorized images and corresponding ground-truth images respectively to an off-the-shelf image classifier which is trained on real color images; compares their results in classification accuracy[19].

User-assisted evaluation:

• *Naturalness*: users are asked to answer "Does this image look natural to you?" after observing each sample image within a limited time[27];

• *Color Turing Test*: one real image and the recolorized counterpart are presented to participants together, and then the participants are asked to point out the fake one[19].

Among existing evaluation criteria, evaluation methods based on direct pixel-wise comparison expect same color values as ground-truth; thus they are overly strict. Moreover, it is hard to prove the consistency between image classification accuracy and colorization quality. User-involved studies directly reflect human's observation but are too costly when dealing with large datasets. As far as we know, no existing criterion has taken into consideration common artifacts like regional inconsistency and color bleeding which are widespread in current colorization models. To address these limitations, we propose two novel evaluation schemes which consider regional inconsistency and color bleeding artifacts into quality evaluation of colorization.

3 Algorithm

Our full pipeline can be formulated as a function F, which maps a single channel input image $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 1}$

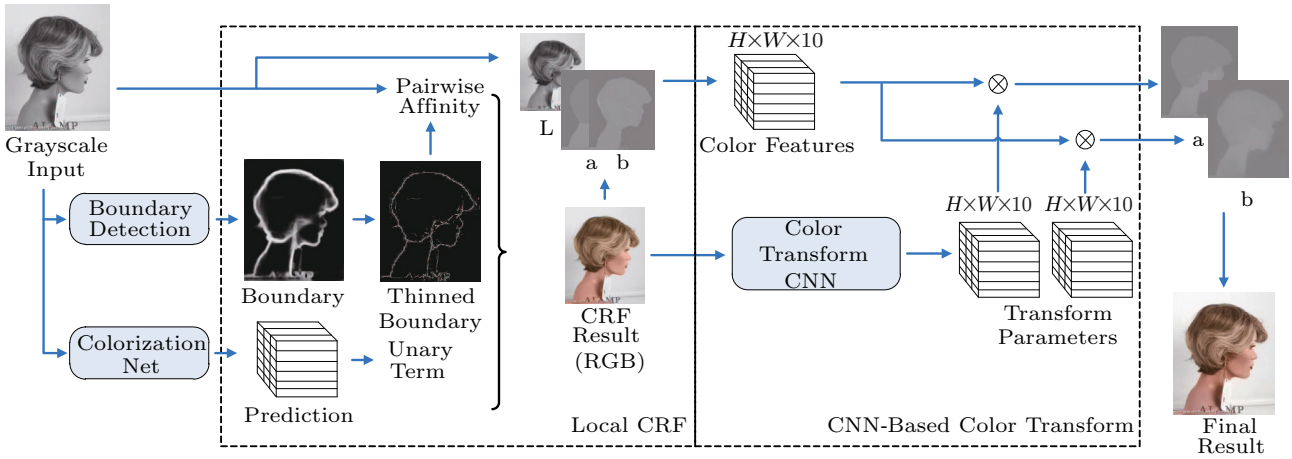
to two color channels a, b $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{H\times W\times 2}$ in CIELab color space, where H, W are spatial dimensions:

$$\widehat{\boldsymbol{Y}} = F(\boldsymbol{X}).$$

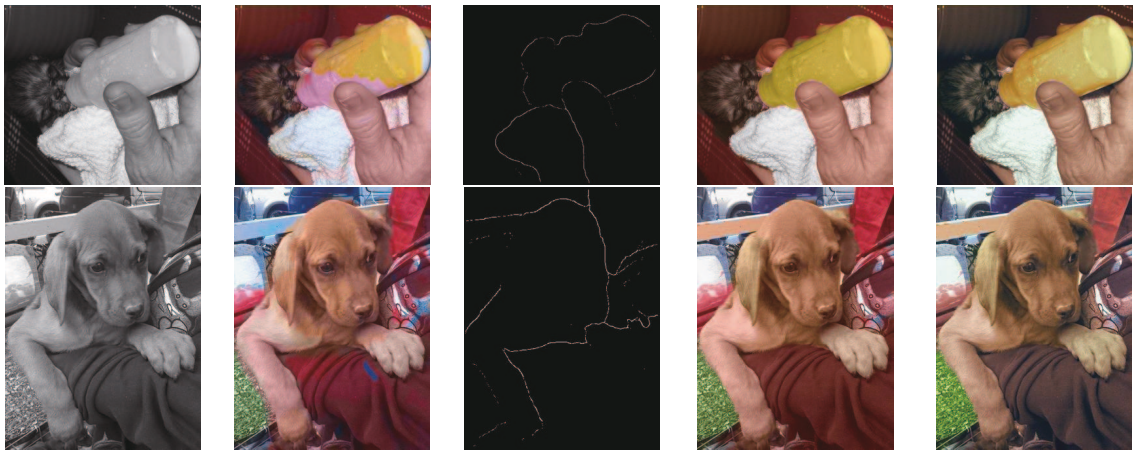
We denote $\boldsymbol{Y}$ as the a, b channels of the ground-truth color image. $\widehat{\boldsymbol{Y}}$ is expected to be close to $\boldsymbol{Y}$ during the learning procedure. The overview of the proposed pipeline is illustrated in Fig.2. It consists of three main phases: an initial colorization model, boundary-guided conditional random field (CRF), and a color transform convolutional neural network (CNN). Firstly, the target grayscale image is fed to a classification-based colorization model which generates a probability distribution over discrete color bins for each pixel. Secondly, we calculate the unary term of CRF based on the predicted probability distribution. When calculating the pairwise affinity of CRF, we involve boundary cues which are obtained by an off-the-shelf boundary detector from the input grayscale image. With such boundary-guided CRF, we aim to improve the spatial coherence of the initial labeling result. Subsequently, to infer final continuous color values, we adopt a CNN to learn a transformation from discrete color bins to continuous color values. In the reminder of this section, we will elaborate the three phases sequentially.

3.1 Initial Colorization

To obtain a good initial prediction result over color bins, we adopt a state-of-the-art deep colorization CNN[19] as our initial colorization model. Following [19], we divide a, b two-dimensional plane into $Q = 313$



(a)



(b)

Fig.2. (a) Overview of our full pipeline for automatic colorization. (b) Example images of each phase in our pipeline.

bins and learn a pixel-wise classification model from input grayscale image via deep CNN in an end-to-end manner:

$$\widehat{\boldsymbol{P}} = G(\boldsymbol{X}),$$

where G denotes the deep CNN and $\widehat{\boldsymbol{P}} \in [0,1]^{H\times W\times Q}$ represents the predicted probability distribution over color bins. Unlike in [19] where the expectations over color bins are calculated as the final color values, we further adopt a boundary-guided CRF to improve spatial coherence in labeling result.

3.2 Boundary-Guided CRF

Let $\phi(i)$ be the expected color label for pixel p_i, and $\widehat{P}_{i,\phi(i)}$ be the probability of assigning color label $\phi(i)$ to pixel p_i. $\widehat{P}_{i,\phi(i)}$ can be predicted by the initial colorization DCNN. The standard energy function of local CRF is defined as follows.

$$E(\phi) = E_{\text{unary}} + \gamma E_{\text{pair}}, \tag{1}$$

where

$$E_{\text{unary}} = -\sum_i \log \widehat{P}_{i,\phi(i)}, \tag{2}$$

$$E_{\text{pair}} = \sum_i \sum_{p_j \in N_{h(p_i)}} \omega_{ij}\tau(\phi(i),\phi(j)), \tag{3}$$

where

$$\omega_{ij} = \exp\left(-\frac{\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2}{2\sigma^2}\right), \tag{4}$$

$$\tau(\phi(i),\phi(j)) = \begin{cases} 1, \text{if } \phi(i) \neq \phi(j), \\ 0, \text{otherwise}, \end{cases} \tag{5}$$

where $N_{h(p_i)}$ is an $h \times h$ neighborhood centered at pixel p_i, and $\boldsymbol{f}_i$ represents the feature of pixel p_i in the target image.

The unary term in (2) is calculated based on $\widehat{P}_{i,\phi(i)}$ while the pairwise term in (3) models the spatial coherence of current labeling scheme, which is conventionally measured by the similarity between neighboring pixels. By this probabilistic graphical model, it is much desired to suppress artifact color predictions in homogeneous regions and keep color boundaries aligned with intrinsic changes perceived from grayscale input. However, due to the lack of chromatic values, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ become scalars, and thus the distance space of $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ is seriously compressed compared with that in color images. This significantly increases the difficulty in discovering detailed edges from an image with only one gray channel. With this consideration, we modify conventional pairwise affinity in (4)∼(6) to explicitly involve boundary cues:

$$\omega_{ij} = \min\left(\exp\left(-\frac{\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2}{2\sigma^2}\right), (1-\hat{g}_{ij})^{\rho}\right), \tag{6}$$

where ρ is a constant and $\hat{g}_{ij}$ represents the maximum boundary response value along the line segment between pixels p_i and p_j. To obtain salient boundaries in grayscale images, we fine-tune the model of state-of-the-art boundary detector HED[26] using grayscale images and adopt an edge-thinning operation.

$$\tau(\phi(i),\phi(j)) = \|\phi(i) - \phi(j)\|_2^2. \tag{7}$$

Furthermore, since class labels are coordinates in a two-dimensional space spanned by a, b chromatic channels, we model the relation between different labels using the Euclidean distance shown in (7) instead of (5). We obtain the color labeling results by minimizing the energy function in (1) using graphcuts[30]. In our implementation, we set $\lambda = 0.5$ and $\rho = 10$.

3.3 CNN-Based Color Transform

As aforementioned, color labeling should not be treated as the final result for colorization due to the fact that realistic color images are composed of a large variety of color values in nearly continuous space. Thus inferring continuous values from CRF labeling results is important for final colorization quality. In this subsection, we introduce a CNN-based color transform model accounting for final color inference.

As shown in the last phase in Fig.2, proposed color transform model takes CRF result images in RGB color space as input and outputs two transform parameter cubes for a, b chromatic channels respectively. Compared with deep colorization network which consists of 22 convolutional layers, our proposed color transform CNN only requires four convolutional hidden layers, which slightly increases the computational complexity but gains obviously better color inference results than previous methods. The detailed architecture of proposed color transform CNN is shown in Fig.3 and Table 1. Specifically, first three hidden layers are shared while the last layer contains two separate branches for learning transform parameter cubes of a, b color channels respectively. We adopt the parametric rectified linear unit (PReLU)[31] as the non-linear activation function between convolutional layers.

Let Φ be the color transform CNN. Let $\mathcal{I}_{\text{rgb}}$ and $\mathcal{I}_{\text{Lab}}$ represent the initial colorized image outputted by boundary-guided CRF in RGB and CIELab color space respectively. We use $\boldsymbol{U}$ to denote the feature

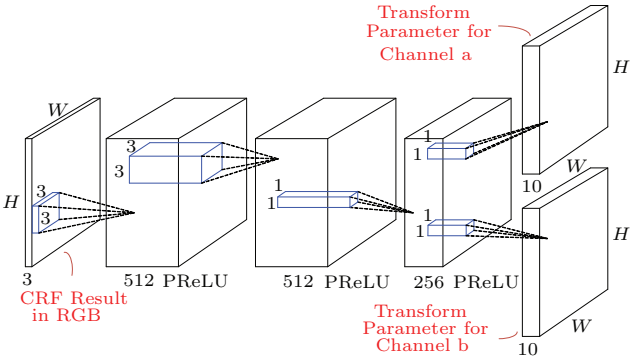


Fig.3. Architecture of color transform CNN.

Table 1. Color Transform CNN Architecture

Layer	Kernel	Stride	Dilation	Output
data				3
conv1	3	1	1	512
conv2	3	1	1	512
conv3	1	1	1	256
conv4_a/b	1	1	1	10

extracted from $\mathcal{I}_{\text{Lab}}$, in which a quadratic color feature $(l^2, a^2, b^2, l\times a, l\times b, a\times b, l, a, b, 1)^{\text{T}}$ is calculated for each pixel. Therefore, $\boldsymbol{U}$ is an $(H\times W\times 10)$-dimensional feature cube, where H, W are spatial dimensions of $\mathcal{I}_{\text{Lab}}$.

The proposed color transform CNN can be formulated as:

$$\boldsymbol{Q}^{\text{a}} = \Phi(\{\Theta_1, \Theta_2, \Theta_{\text{a}}\}, \mathcal{I}_{rgb}),$$
$$\boldsymbol{Q}^{\text{b}} = \Phi(\{\Theta_1, \Theta_2, \Theta_{\text{b}}\}, \mathcal{I}_{rgb}),$$

where $\boldsymbol{Q}^{\text{a}}$ and $\boldsymbol{Q}^{\text{b}}$ are $(H\times W\times 10)$-dimensional color transform parameter cubes for channels a, b respectively. Θ_1 and Θ_2 represent the weights of the first two convolutional layers respectively while Θ_{a} and Θ_{b} are the weights of the last layer for channels a, b respectively.

The loss function is defined as follows:

$$L = \frac{1}{2}\sum_{i}^{H}\sum_{j}^{W}((\sum_{k=1}^{10}\boldsymbol{Q}^{\text{a}}_{i,j,k}\boldsymbol{U}_{i,j,k} - \boldsymbol{Y}^{\text{a}}_{i,j})^2 + (\sum_{k=1}^{10}\boldsymbol{Q}^{\text{b}}_{i,j,k}\boldsymbol{U}_{i,j,k} - \boldsymbol{Y}^{\text{b}}_{i,j})^2),$$

where $\boldsymbol{Y}^{\text{a}}$ and $\boldsymbol{Y}^{\text{b}}$ represent a, b channels of ground-truth color image respectively. We implement this color transform CNN based on the popular open source framework Caffe[32] and solve the energy minimization using standard stochastic gradient descent with learning rate 0.001 and weight decay 0.000 5.

Fig.4 shows examples of our colorization results with and without applying CNN-based color transform. The comparison shows CNN-based color transform infers continuous chromatic values from discrete labeling results without significantly shifting color values, making our final colorization more natural and realistic.

(a)

(b)

Fig.4. Colorization results with and without CNN-based color transform. (a) Color labeling results from CRF. (b) Colorization results from color transform CNN.

4 Spatial-Consistency Evaluation

In this section, we introduce two novel evaluation schemes to efficiently measure color consistency and boundary contrast consistency in generated colorful images.

4.1 Regional Color Consistency

Regional color consistency is one of the key factors for plausible colorization results. In real color images, different levels of chromatic variations can be observed in variant semantic regions: some regions are visually homogeneous while others may contain textures or be photographed under complex lighting environment. Though knowing nothing about the ground-truth, people can effortlessly judge unreal colorization once capturing inconsistent colors in homogeneous regions. Such observation indicates that measuring color variations in homogeneous regions is a promising strategy for evaluating colorization quality. Based on this, we propose to evaluate colorization quality by sampling a pixel group in each homogeneous region, and then calculating its color variation.

One example of the proposed evaluation scheme is shown in Fig.5. For each testing image, the ground-truth color version is transformed to CIELab color space, in which the L channel represents lightness and a, b represent color-opponent dimensions. We first perform graph-based segmentation[33] to generate superpixels. Note that in order to weaken the interference of lightness variations, graph-based segmentation is operated on pixelwise color vectors $\boldsymbol{f}_c = (k \times l, a, b)^{\mathrm{T}}$ where k is used for controling the weakening degree of the L channel. We collect one representative pixel which is spatially closest to its centroid, from each superpixel. Due to large color variation in the whole image, it is unreasonable to directly calculate color consistency over all representative pixels. Thus, we merely evaluate on pixel groups which locate in homogeneous regions.

Subsequently, we perform a fast hierarchical segmentation using MCG[34] to each ground-truth image and discard small segments which are below the average spatial size. In each remaining segment, we select one group from representative pixels with an identical color value that is closest to the mean value of this segment in a, b channels. We further discard pixel groups which contain fewer than S_r pixels. Finally, we repeat the same selection scheme for all testing images, thus obtaining all pixel groups.

We denote the i-th pixel group in image j as G_{ij}. Our regional color consistency evaluation is defined as follows:

$$W_r = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\widehat{\sigma}_{i,j}^{r},$$
$$\widehat{\sigma}_{i,j}^{r} = \widehat{\sigma}_{a,i,j} + \widehat{\sigma}_{b,i,j},$$

where

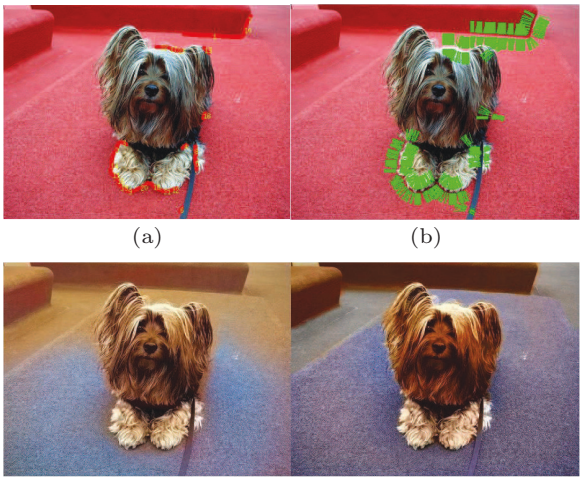
$$\widehat{\sigma}_{c,i,j} = \sqrt{\frac{1}{P_{i,j}}\sum_{p_l \in G_{ij}}(\widehat{c}_{p_l} - \widehat{\mu}_{c,i,j})^2},$$
$$\widehat{\mu}_{c,i,j} = \frac{1}{P_{i,j}}\sum_{p_l \in G_{ij}}\widehat{c}_{p_l}, \quad c \in \{a, b\},$$

where $\widehat{\sigma}_{c,i,j}$ denotes the standard deviation of pixel group G_{ij} in channel a (or b) in the generated colorization image j. $P_{i,j}$ denotes the pixel number in group G_{ij} and $\widehat{c}_{p_l}$ is the color value of pixel p_l in channel a (or b). M is the number of total testing images and N_j is the number of pixel groups in image j. We use $k = 0.3$ and $S_r = 10$ in our evaluation.

Fig.5. One example of color consistency evaluation. (a) Ground-truth color image. (b) Superpixels generated by graph-based segmentation on pixelwise color vectors $\boldsymbol{f}_c = (k \times l, a, b)^{\mathrm{T}}$ (shown in random colors). (c) Hierarchical image segmentation results (shown in random colors). (d) Selected point group for consistency evaluation (shown in blue asterisks). (e) State-of-the-art colorization result by [19], $\widehat{\sigma}^r = 7.91$. (f) Our result, $\widehat{\sigma}^r = 1.43$.

4.2 Boundary Localization

Poor color discrimination along boundaries is another key factor which leads to the unreality of the generated colorization results. For example, Fig.6(c) shows the generated colorization result of [19]. This model predicts suitable color for dog, sofa and carpet separately, but color bleeding across boundaries can be observed, e.g., colors for the dog bleed over its boundary. Hence, we propose another approach to evaluate colorization quality in terms of boundary localization.



(a) (b) (c) (d)

Fig.6. One example of color bleeding evaluation. (a) Selected boundary sections for evaluation from ground-truth color image. (b) Selected pixel groups (shown in green color) on both sides of boundaries. (c) State-of-the-art colorization result by [19]. (d) Our result.

Fig.6 illustrates one example of our boundary localization evaluation scheme. For each ground-truth color image, we obtain its boundary map via HED[26] boundary detector, and then threshold the resulted boundary map by 0.5 followed by an edge thinning operation. We subsequently divide boundaries into short sections using a boundary subdivision method[34] which involves a recursive boundary breaking procedure. Among the results, boundary sections with less than S_b pixels are not considered. Afterwards, for each pixel located on resulted boundary sections, we sample r pixels along the local boundary normal on both sides. In this way, two pixel groups are sampled on both sides of each selected boundary section. We calculate the variation of the i-th boundary section as below,

$$\sigma_{\mathrm{L},i} = \sigma_{\mathrm{L},i,a} + \sigma_{\mathrm{L},i,b},$$
$$\sigma_{\mathrm{R},i} = \sigma_{\mathrm{R},i,a} + \sigma_{\mathrm{R},i,b},$$
$$\sigma_i = \sigma_{\mathrm{L},i} + \sigma_{\mathrm{R},i},$$

where $\sigma_{L,i}$ and $\sigma_{R,i}$ represent the stand deviation of selected pixel group from the left and the right hand side of boundary section i respectively. The subscripts a, b denote color channels in CIELab color space respectively. We further filter out boundary sections with large variations by setting a low threshold T_{std} for σ_i. Resulted boundary sections and sampled pixel groups on both sides provide the target locations for evaluation.

Finally, we define a localization evaluation criterion by calculating the color variation in colorized images according to target evaluation locations:

$$W_b = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\widehat{\sigma}_{i,j},$$

where N_j denotes the number of boundary sections selected in image j and M denotes the number of images for evaluation. We set $S_b = 10$, $r = 10$ and $T_{std} = 50$ in our evaluation.

5 Experimental Results

5.1 Experimental Settings

We use the same dataset as in [19]: all the training images (around 1.3M) in ImageNet[35] are used as our training data, while the first 2k and the last 10k images in the validation set of ImageNet are used as validation and testing data respectively in our experiments. For each testing image, it takes around 5 seconds to calculate the CRF labeling result and less than 1 second for all remaining steps in our proposed pipeline using a NVIDIA Titan X GPU.

Since single evaluation criterion is not adequate for a comprehensive measurement of the colorization quality, we use multiple criteria to draw an overview of colorization results. Besides the proposed evaluation methods for color consistency and color bleeding, we further adopt the raw accuracy measure method[19] to provide a pixel-wise comparison with the corresponding ground-truth color images.

5.2 Raw Accuracy Evaluation Results

We employ the rebalanced variant of the AuC (area under curve) CMF (conditional mass function) metric[19] for a raw accuracy evaluation. This measurement first calculates the rebalanced pixel-wise Euclidean distance in a, b color channels between colorized image and its corresponding ground-truth color image. Then the curve is formed by calculating the percentage

of pixels within certain distance from 0 to 150. Hence, higher AuC score indicates smaller distance between colorization results and ground-truths. The evaluation results are shown in Fig.7. As shown in Fig.7, except [17] and grayscale images, the AUC scores of all the other three methods (including ours) are almost equal, with variation within 0.5%. Thus we conclude that regional color inconsistency and color bleeding phenomena, which apparently bring great visual defect, cannot be reflected under the AuC metric. A visual comparison is demonstrated in Figs.8 and 9.

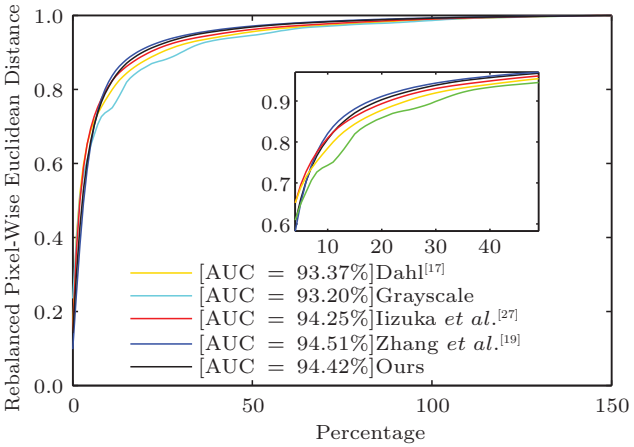


Fig.7. Comparing rebalanced Euclidean distance in a, b color channels. A sub-figure that locates in the up-right corner shows the enlarged curve.

5.3 Consistency Evaluation Results

Following the proposed point group selection scheme introduced in Subsection 4.1, 5 525 images are selected from testing images for evaluation. Then our proposed color consistency measurement is performed on locations of selected pixel groups in each testing image.

Table 2 lists the evaluation results using the proposed color consistency criterion. As shown in the table, our colorization results significantly achieve the lowest color variations in homogeneous regions when compared with other leading colorization methods. Besides, we observe that the CNN-based color transform slightly improves the regional color consistency, although it is not initially designed for this purpose.

Table 2. Color Consistency Evaluation Results

Name	W_r
Dahl[17]	3.71
Iizuka *et al.*[27]	3.16
Zhang *et al.*[19]	6.54
Ours without color transform	2.84
Our full pipeline	**2.63**

This can be understood from the additional spatial smoothness that CNN-based color transform introduces.

5.4 Boundary Localization Evaluation Results

Based on the proposed selection criterion introduced in Subsection 4.2, 21 156 boundary sections are selected for our color bleeding evaluation. The results of quantitative measurement on the selected boundary sections are illustrated in Table 3.

Table 3. Boundary Localization Evaluation Results

Name	W_b
Dahl[17]	5.20
Iizuka *et al.*[27]	5.50
Zhang *et al.*[19]	8.57
Ours without color transform	5.33
Our full pipeline	**5.16**

Table 3 reveals that our colorization results give rise to much smaller chromatic variations on both sides of selected boundary sections. Meanwhile, we also observe CNN-based color transform slightly benefits boundary localization.

In general, our method achieves significant improvement in colorization quality by enhancing regional color consistency and eliminating color bleeding.

5.5 User Study

In order to evaluate the colorization results quantitatively, we perform a user study. The user study is conducted by showing a pair of colorized images at a time, which are generated from one grayscale image using our proposed method and the method of Zhang *et al.*[19] respectively. The user is asked to choose "Which image looks more natural to you?" after comparing the two results. Unlike the user studies of Zhang *et al.*[19] and Iizuka *et al.*[27], which only allow users to take a quick glance of each image pair, we do not limit the time for each comparison and encourage users to combine their gut feeling with detailed observations. Besides, we provide a third option "hard to decide" for each comparison in case users could not decide their preference. We invite 16 different participants in our user study, each showing 40 pairs of different images. All images are randomly chosen from testing dataset and randomly shown on the left or right hand side to avoid bias.

Fig.10 shows the result of the user study. We can see our method receives more users' satisfaction than

(a) (b) (c) (d) (e)

Fig.8. Comparison with leading automatic colorization methods. (a) Dahl[17]. (b) Iizuka *et al.*[27] (c) Zhang *et al.*[19] (d) Ours. (e) Ground-truth.

Fig.9. Comparison with leading automatic colorization methods. (a) Grayscale input. (b) Dahl[17]. (c) Iizuka *et al.*[27] (d) Zhang *et al.*[19] (e) Ours.

state-of-the-art colorization method. Besides, during our user study, several participants mention that they clearly decide their preference by capturing the color bleeding phenomena, which is consistent with our proposed evaluation schemes.

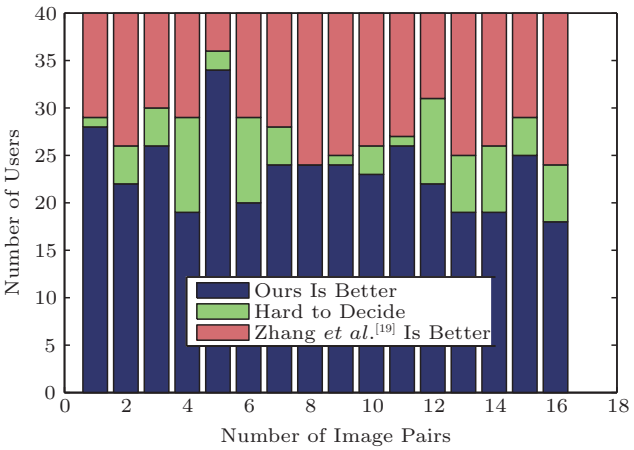


Fig.10. Results of user study. We compare our method and a state-of-the-art method from Zhang *et al.*[19] in terms of naturalness.

6 Conclusions

In this paper, we proposed post-processing steps for automatic image colorization, which involve a boundary-guided CRF and a CNN-based color transform model. Extensive experimental results demonstrated that our proposed methods greatly improve the colorization quality compared with current leading methods. To prove the effectiveness of our proposed methods, we further introduced two novel evaluation schemes to quantitatively measure the quality of automatically colorized images in terms of color consistency and boundary localization.

References

[1] Levin A, Lischinski D, Weiss Y. Colorization using optimization. *ACM Transactions on Graphics* (*TOG*), 2004, 23(3): 689-694.

[2] Huang Y C, Tung Y S, Chen J C, Wang S W, Wu J L. An adaptive edge detection based colorization algorithm and its applications. In *Proc. the 13th Annual ACM International Conference on Multimedia*, Nov. 2005, pp.351-354.

[3] Luan Q, Wen F, Cohen-Or D, Liang L, Xu Y Q, Shum H Y. Natural image colorization. In *Proc. the 18th Eurographics Conference on Rendering Techniques*, Jun. 2007, pp.309-320.

[4] Qu Y,Wong T T, Heng P A. Manga colorization. *ACM Transactions on Graphics* (*TOG*), 2006, 25(3): 1214-1220.

[5] Zhao H L, Nie G Z, Li X J, Jin X G, Pan Z G. Structure-aware nonlocal optimization framework for image colorization. *Journal of Computer Science and Technology*, 2015, 30(3): 478-488.

[6] Sheng B, Sun H, Magnor M, Li P. Video colorization using parallel optimization in feature space. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(3): 407-417.

[7] Welsh T, Ashikhmin M, Mueller K. Transferring color to greyscale images. *ACM Transactions on Graphics* (*TOG*), 2002, 21(3): 277-280.

[8] Irony R, Cohen-Or D, Lischinski D. Colorization by example. In *Proc. Eurographics Symp. Rendering Techqiques*, June 29-July 1, 2005, pp.201-210.

[9] Charpiat G, Hofmann M, Schölkopf B. Automatic image colorization via multimodal predictions. In *Proc. the 10th European Conference on Computer Vision*, Oct. 2008, pp.126-139.

[10] Liu X, Wan L, Qu Y, Wong T T, Lin S, Leung C S, Heng P A. Intrinsic colorization. *ACM Transactions on Graphics* (*TOG*), 2008, 27(5): 152:1-152:9.

[11] Gupta R K, Chia A Y S, Rajan D, Ng E S *et al.* Image colorization using similar images. In *Proc. the 20th ACM International Conference on Multimedia*, Oct.29-Nov.2, 2012, pp.369-378.

[12] Jin S Y, Choi H J, Tai Y W. A randomized algorithm for natural object colorization. *Computer Graphics Forum*, 2014, 33(2): 205-214.

[13] Chia A Y S, Zhuo S, Gupta R K, Tai Y W, Cho S Y, Tan P, Lin S. Semantic colorization with Internet images. *ACM Transactions on Graphics* (*TOG*), 2011, 30(6): 156:1-156:8.

[14] Deshpande A, Rock J, Forsyth D. Learning large-scale automatic image colorization. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.567-575.

[15] Li X, Zhao H, Nie G, Huang H. Image recoloring using geodesic distance based color harmonization. *Computational Visual Media*, 2015, 1(2): 143-155.

[16] Cheng Z, Yang Q, Sheng B. Deep colorization. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.415-423.

[17] Dahl R. Automatic colorization. http://tinyclouds.org/colorize/, Aug. 2016.

[18] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. In *Proc. European Conference on Computer Vision*, Oct. 2016, pp.577-593.

[19] Zhang R, Isola P, Efros A A. Colorful image colorization. In *Proc. European Conference on Computer Vision*, Oct. 2016, pp.649-666.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556, Aug. 2016.

[21] Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.447-456.

[22] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.1520-1528.

[23] Li G, Yu Y. Deep contrast learning for salient object detection. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.478-487.

[24] Li G, Yu Y. Visual saliency based on multiscale deep features. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.5455-5463.

[25] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems*, Dec. 2015, pp.91-99.

[26] Xie S, Tu Z. Holistically-nested edge detection. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.1395-1403.

[27] Iizuka S, Simo-Serra E, Ishikawa H. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics* (*TOG*), 2016, 35(4): 110:1-110:11.

[28] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.1520-1528.

[29] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122, 2015. https://arxiv.org/abs/1511.07122, Aug. 2016.

[30] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(11): 1222-1239.

[31] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. the IEEE International Conference on Computer Vision*, Dec. 2015, pp.1026-1034.

[32] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In *Proc. the 22nd ACM International Conference on Multimedia*, Nov. 2014, pp.675-678.

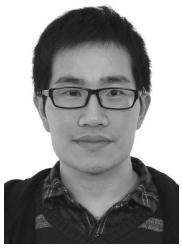
[33] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004, 59(2): 167-181.

[34] Arbeláez P, Pont-Tuset J, Barron J T, Marques F, Malik J. Multiscale combinatorial grouping. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp.328-335.

[35] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M *et al*. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, 115(3): 211-252.

Wei Zhang is a Ph.D. candidate at the Department of Computer Science, The University of Hong Kong, Hong Kong. He received his B.E. degree in automation from Chongqing University, Chongqing, in 2010, and M.S. degree in pattern recognition and artificial intelligence from Huazhong University of Science and Technology, Wuhan, in 2013. His research interest covers image colorization and boundary detection.

Chao-Wei Fang is a Ph.D. candidate at the Department of Computer Science, The University of Hong Kong, Hong Kong. He received his B.E. degree in automation from Xi'an Jiaotong University, Xi'an, in 2013. His current research interests include image segmentation and image processing.



Guan-Bin Li received his Ph.D. degree in computer science from The University of Hong Kong, Hong Kong, in 2016. He is currently a researcher in the School of Data and Computer Science, Sun Yat-sen University, Guangzhou. He is a recipient of Hong Kong Postgraduate Fellowship. His current research interests include computer vision, image processing, and deep machine learning.