Example-Based Colourization Via Dense Encoding Pyramids

Article *in* Computer Graphics Forum · April 2019

DOI: 10.1111/cgf.13659

CITATIONS
2

READS
292

7 authors, including:

Chufeng Xiao
City University of Hong Kong
2 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Chu Han
The Chinese University of Hong Kong
9 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Jing Qin
The Hong Kong Polytechnic University
210 PUBLICATIONS **3,608** CITATIONS
SEE PROFILE

Tien-Tsin Wong
The Chinese University of Hong Kong
207 PUBLICATIONS **3,666** CITATIONS
SEE PROFILE

DOI: 10.1111/cgf.13659

COMPUTER GRAPHICS *forum*
Volume 00 (2019), number 00 pp. 1–14

Example-Based Colourization Via Dense Encoding Pyramids

Chufeng Xiao[1,*], Chu Han[2,*], Zhuming Zhang[2], Jing Qin[3], Tien-Tsin Wong[2], Guoqiang Han[1] and Shengfeng He[1,#]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
chufengxiao@outlook.com, {csgqhan, hesfe}@scut.edu.cn
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
{chan, zhangzm, ttwong}@cse.cuhk.edu.hk
[3]School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
harry.qin@polyu.edu.hk

Abstract

We propose a novel deep example-based image colourization method called dense encoding pyramid network. In our study, we define the colourization as a multinomial classification problem. Given a greyscale image and a reference image, the proposed network leverages large-scale data and then predicts colours by analysing the colour distribution of the reference image. We design the network as a pyramid structure in order to exploit the inherent multi-scale, pyramidal hierarchy of colour representations. Between two adjacent levels, we propose a hierarchical decoder–encoder filter to pass the colour distributions from the lower level to higher level in order to take both semantic information and fine details into account during the colourization process. Within the network, a novel parallel residual dense block is proposed to effectively extract the local–global context of the colour representations by widening the network. Several experiments, as well as a user study, are conducted to evaluate the performance of our network against state-of-the-art colourization methods. Experimental results show that our network is able to generate colourful, semantically correct and visually pleasant colour images. In addition, unlike fully automatic colourization that produces fixed colour images, the reference image of our network is flexible; both natural images and simple colour palettes can be used to guide the colourization.

Keywords: image and video processing, image processing

ACM CCS: I.3.3 [Computer Graphics]: Picture/Image; Computing Methodologies: Neural Networks, Computational Photography

1. Introduction

The goal of the image colourization is to turn a greyscale image into a colour image. It is motivated by the demand for image editing to restore the colour of old pictures and videos. However, to recover the already lost colour information is challenging. Additional information like user interactions or a colour image as a reference may be an alternative solution for this ill-posed problem. Existing image colourization techniques can be summarized into two main categories, traditional colourization with user interactions, learning-based automatic colourization.

At the early stage, most of the colourization methods, e.g. [QWH06, LWCO*07, RKB04, LLW04, XYJ13], were proposed with the requirement of user interactions, like colour scribbles or manual segmentation. Guided by the user input, an optimization process was used to propagate the colours. It is good for users to control the colours of what they want. However, the optimization process is usually computation-intensive and time-consuming. Moreover, colour leakage and drifting may often occur at the regions with open boundaries. Since such simple manual hints are not able to deliver rich enough colour information, some works [WAM02, ICOL05] proposed to colourize images guided by a reference image. However, traditional example-based methods were trapped by the similarity between the reference and the greyscale images. Once the discrepancy of reference and greyscale images is large, these methods may not be able to guarantee a visually reasonable result.

With the evolution of CNN, more data-driven automatic colourization methods were proposed. The approaches [CYS15, ZIE16,

*Joint first authors.
#Corresponding author.

(a) Original (b) [ZIE16] (c) [LMS16] (d) [ZZI*17] (e) [HCL*18] (f) Ours

Figure 1: *Previous deep learning-based methods, e.g. automatic methods (b and c) and example-based methods (d and e), cannot fully explore the local–global context information, leading to unsatisfactory results. Our dense encoding pyramids (f) is able to produce plausible results with varied colours according to the reference images or colour palette. More importantly, it is more user-friendly than others since users can easily manipulate the colour palettes for desired styles, like the second row. Our results are more faithful and have less colour drifting (first row of d) than results from other methods. Images that are placed at the bottom left corner are the reference images.*

LMS16] presented the fully automatic colourization network by learning from a large-scale colour image data. Iizuka *et al.* [ISSI16] learned the semantic information from an image and automatically colourized the image according to the global priors and local image features. However, the learning-based automatic methods are not flexible for generating images with desired colour distributions. Zhang *et al.* [ZZI*17] proposed a deep user-guided colourization method, but such user interactions have a latent requirement that users should have good enough art sense for choosing a suitable colour scheme, especially for realistic images. Otherwise, the colourized image maybe unnatural. Furthermore, existing deep colourization networks cannot explore the local–global context well, which leads to semantically wrong and dull colourization results (see Figure 1).

In this paper, we propose a novel deep example-based colourization method, which is called the Dense Encoding Pyramids Network (DEPN). The key idea of our method is that the image colour priors always come from the image itself. Our network performs colourization by mapping the colour distribution from a reference image to a greyscale image. We design our network as a pyramid form in order to leverage the pyramidal shape of the colour representations hierarchy. Furthermore, we propose a novel Parallel Residual Dense Block (PRDB) for exploring richer local–global context information within the network. A Hierarchical Decoder–Encoder Filter (HDEF) is proposed to aggregate the colour distribution results between two adjacent levels. Extensive experiments are conducted to evaluate the performance of our method. Experimental results demonstrate the proposed method outperforms the existing methods quantitatively and qualitatively, and it shows a clear preference in the user study over previous methods.

2. Related Works

2.1. Traditional colourization

To restore the lost information from a greyscale image, traditional colourization methods require some guidance from users, either user interactions or a reference image. Levin *et al.* [LLW04] proposed an optimization-based method with the requirement of input colour scribbles. Qu *et al.* [QWH06] proposed a user interactive method for manga colourization. Sýkora *et al.* [SBv04] presented a colourization method for hand-drawn cartoons. Luan *et al.* [LWCO*07] applied texture information for better colour propagation. The above methods apply optimization according to the user interactions and then propagate the colours. They allow users to take full control of the colours. However, the optimization process is computationally intensive. In addition, manually drawn scribbles are not able to deliver rich enough colour information.

Some colour transfer methods [TJT05, RAGS01, WDK*13] can also be applied on transferring colour to a greyscale image. They establish a mapping function to map the colour distribution from one to another. Similar to colour transfer techniques, [WAM02] colourizes images guided by a reference image without user intervention. They transfer the colour from the reference image to target image by matching luminance and texture information between them. Ironi *et al.* [ICOL05] introduced a method to colourize greyscale images by transferring colour from a segmented example image. Gupta *et al.* [GCR*12] proposed an example-based colourization method by leveraging superpixel representation to guide the colourization process. However, their method requested the reference image has to be similar to the greyscale image. Bugeau *et al.* [BTP14] proposed a variational approach to select the best colour among a set of colour candidates while simultaneously ensuring the local spatial coherency of the reconstructed colour information. However, traditional example-based methods were trapped by the similarity between the reference image and the greyscale image. Once the style discrepancy of the reference image and the greyscale image is large, these methods may not be able to guarantee a visually reasonable result.

2.2. Learning-based colourization

With the evolution of CNN, more automatic colourization methods were proposed. Cheng *et al.* [CYS15] and Zhang *et al.* [ZIE16]

presented fully automatic colourization networks by leveraging large-scale colour image data [ZLX*14, RDS*15]. However, these fully automatic methods without any other hints highly rely on the training data. Once the input image is not covered by the training data, it may fail to generate a reasonable and visually pleasant result. Iizuka *et al.* [ISSI16] learned the semantic information from an image and automatically colourized the image according to the global priors and local image features. Larsson *et al.* [LMS16] leveraged deep semantic information associated with a per-pixel colour histogram to estimate the colour of each pixel. Royer *et al.* [RKL17] developed a probabilistic technique which can generate different plausible colourizations. Guadarrama *et al.* [GDB*17] utilized two networks to produce multiple colourization results automatically. A conditional PixelCNN was used to generate a low-resolution colour image while another CNN was used to give a high resolution one. But such user interactions have a latent requirement that users should have good enough art sense for choosing a suitable and harmonious colour scheme, especially for realistic images. Otherwise, the colourized image may be unnatural.

Some example-based methods are proposed to let the user control the colour distribution by providing a reference image. Zhang *et al.* [ZZI*17] proposed a deep user-guided colourization method. The network propagates user interactions by fusing low-level cues along with high-level semantic information which learned from large-scale data. However, colourization results highly depend on the quality of the reference image. In the meanwhile, it also has to be as similar as possible to the target image. He *et al.* [HCL*18] presented an example-based local colourization approach with two subnetworks. The similarity subnet was designed to find the semantic correspondence between the reference and greyscale images, then fuses multi-level warped features. The colourization subnet takes the output of the similarity subnet to generate vivid results. However, when feeding a reference image which lacks semantic information, like a simple colour palette, the whole network fails to produce plausible results. This is because the similarity subnet cannot obtain semantic and luminance information.

Our proposed network leverages large-scale data to encode colour information in our model. It is able to colourize image by showing the network only one reference image without any other manual hints. Different from the fully automatic colourization networks, our network can generate different colourization results according to different reference images.

2.3. Colour transfer

Colour transfer techniques can also achieve colourization given a reference image. Chang *et al.* [CFL*15] introduced a palette-based approach to recolour an image. It developed an accelerated clustering method to produce a colour palette from one image. They achieve colour transfer by user changes of the colour palette. Wang *et al.* [WZL*17] presented a two-stage method which includes similarity mapping and detail conservation. The similarity mapping model used super-pixel sampling and *K*-means clustering to extract the intermediate features in RGB space from the input and the reference image. Then an L0 gradient-preserving algorithm was developed to generate the transfer results by controlling the gradients of pixels within their colour regions.

Our colourization method can also be applied to colour transfer by simply removing the chrominance channel of the original image and produce diverse colour images given different references.

2.4. Pyramid structure

Pyramid structure has been widely used in solving computer vision problems. Spatial pyramid pooling methods [LSP06, GD05, YYGH09] extract image context at different scales which can reduce the computational complexity. He *et al.* [HZRS14] introduced spatial pyramid pooling into the CNN to make it possible to generate representations from arbitrarily sized images. Chen *et al.* [CPK*18] proposed atrous spatial pyramid pooling, where parallel atrous convolution layers with different sampling rates and effective fields-of-views to capture multi-scale information. Zhao *et al.* [ZSQ*17] conducted a pyramid scene parsing network which performs spatial pooling at several grid scales and achieved great performance in semantic segmentation. The proposed method, on the other hand, is designed with the different information flow and recurrent mechanism for accurate and varied colour assignment.

2.5. Residual and dense structure

Recently, some residual and dense structures are proposed to extract hierarchical features and make full use of them. Huang *et al.* [HLWvdM17] presented DenseNet to fuse features from each layer, which effectively overcome the vanishing-gradient problem with much fewer parameters. DenseNet consists of a number of dense blocks, in which any layer connects to all subsequent layers. It is an efficient extractor for many computer vision and graphics problems. Zhang *et al.* [ZTK*18] introduced a residual dense block (RDB) for image super-resolution. Through dense connected convolutional layers, RDB bridges directly from the former layers to all of the current layers, providing a contiguous memory. In addition, global features fusion is applied to integrate all the RDBs, which performs well in image super-resolution. Zhang *et al.* [ZLL*18] conducted a very deep residual channel attention network which builds a highly accurate model for image SR. The residual in residual (RIR) structure is introduced to increase the depth of the network, which helps filter the abundantly invalid information by several skip connections.

However, for image colourization, an extremely deep architecture may not be helpful. It is because both low- and high-level knowledge is necessary for generating semantically correct and vivid colour images. Thus, taking full advantage of multi-scale features is more important than learning more redundant features. In our paper, we propose a parallel structure based on RDB [ZTK*18] to extract low- and high-level information effectively in a wider network. The residual and dense structure allows the network to learn low-level information and extracts features than the normal convolution. The proposed PRDB aims at aggregating more information from the dense parallel structure.

3. Method

The architecture of the proposed DEPN is illustrated in Figure 2. It aims at more semantic-correctly map the colour distribution of

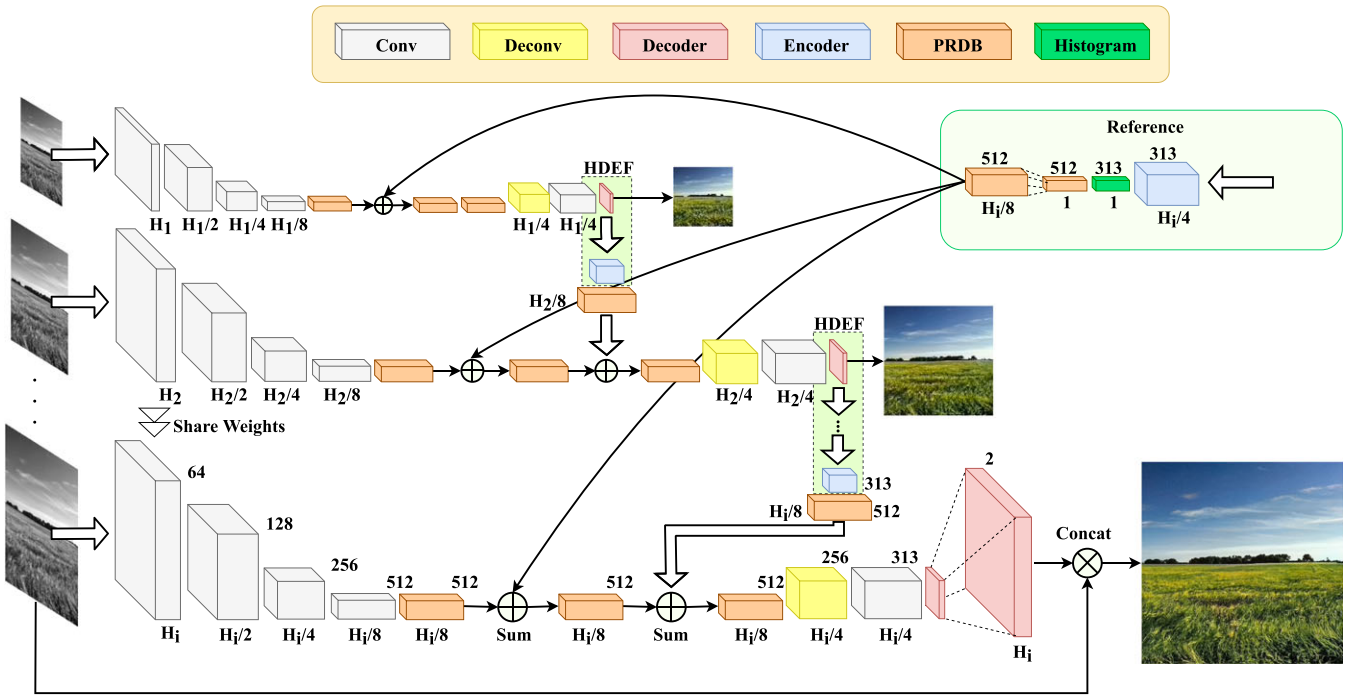


Figure 2: *Our network structure. For each level of the pyramid, the network branch takes a greyscale image under specific resolution as an input image. The deep colour distribution features of the reference image are extracted by parallel residual dense blocks and then are utilized to guide every level of the pyramid. The hierarchical decoder–encoder filter is proposed to pass the colour distributions from lower level to higher level. Note that the number of feature maps in each block at each level is all the same.*

the reference image onto the input greyscale image. We design the network as a pyramid structure in order to exploit the inherent multi-scale, pyramidal hierarchy of colour representations. The PRDBs are proposed to extract abundant features within the network by making the network wider instead of deeper, which is important to generate varied colour images. We further define an HDEF to promote the information propagation between different layers of the pyramidal network, avoiding potential error accumulation by harnessing information extracted from other layers as guidance. The colour distribution of the reference image is extracted by an encoder followed by a PRDB and then is injected into each level as the guidance of the colourization process. We shall systematically introduce the implementation, as well as the underlying thoughts, of each component of the proposed DEPN, including its objective function (Section 3.1), the network architecture (Section 3.2), the PRDBs (Section 3.3), the HDEF (Section 3.4), the reference image feature extraction (Section 3.5) and the implementation and training details (Section 3.6).

3.1. Objective function

Given a greyscale image X and a reference image Y, our network aims to learn the mapping $\mathcal{F}(X, Y)$ in order to colourize X with the colour distribution of Y. In our study, we utilize CIE Lab colour space instead of RGB because of the perceptual uniformity with respect to human colour vision. Our network regards $X \in \mathbb{R}^{H\times W\times 1}$ as the L channel then outputs the result $\hat{X} \in \mathbb{R}^{H\times W\times 2}$, which indicates the ab channels of the image. Similar to Zhang *et al.* [ZIE16], we define the colourization problem as a multinomial classification. We quantize the ab channels into grids with the size of 10 so that the number of colour pairs of ab channels can drastically decrease, which is capable of greatly reducing the computation. As a result, we keep the vector Q with 313 values in the gamut, indicating the number of ab pairs.

According to our pyramid network design, the objective of each network branch (as shown in Figure 2) is to minimize the expected classification loss over the training data set:

$$\hat{\theta}^i = \arg\min_{\theta^i}(\mathcal{L}_{cl}(\mathcal{F}^i(X^i, Y^i; \theta^i), Y^i)), \tag{1}$$

where θ_i indicate the parameters in the ith-level network branch. $\mathcal{L}_{cl}$ is the cross entropy loss function with the rebalancing weight to measure the errors between the prediction of colour distribution $\hat{Z}$ and the ground truth colour distribution $Z \in \mathbb{R}^{H\times W\times Q}$ of the reference image Y. To the end, the loss function of our network at the ith level is defined as:

$$\mathcal{L}_{cl}(\hat{Z}^i, Z^i) = -\sum_{h,w} \omega^i\left(Z^i_{h,w}\right) \sum_q Z^i_{h,w,q} \log\left(\hat{Z}^i_{h,w,q}\right), \tag{2}$$

where ω^i represents the rebalancing weight of $Z^i_{h,w}$. It is able to correct the unbalanced foreground–background distribution of ab values.

3.2. Dense encoding pyramids

The previous methods [ZIE16, ZZI*17, HCL*18] only consider a single-scale image as reference. However, the major drawback of this scheme is that they lack local–global context of the colour representations. In this paper, we propose a novel pyramidal structure, which inputs the multi-scale images and lets the features from lower levels guide the higher levels in a coarse-to-fine manner.

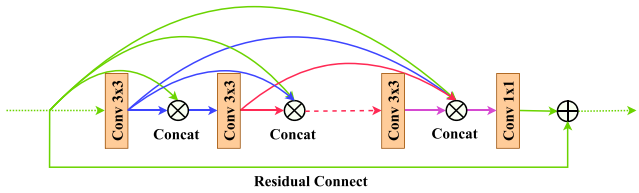
As shown in Figure 2, our DEPN consists of several network branches in a pyramid form. The main advantages brought by the pyramid structure are the multi-scale contextual information and the recurrent prediction mechanism. Although the distributions of the reference image are the same across different scales, the assigned colours are depending on a certain region of image content, which directly affected by the input resolution and receptive fields of the network. Our pyramid structure colourizes image from four different scales of contextual information. Furthermore, the recurrent mechanism predicts colour image in a coarse-to-fine manner with progressive refinement.

Here, we denote the number of pyramid levels as l. In principle, the proposed DEPN can be considered as a coarse-to-fine process for image colourization. As can be seen in Figure 2, each network branch can be regarded as a complete network for colourization at a specific resolution. For example, considering the ith level $\mathcal{F}^i$, it receives a greyscale image X^i and a reference image Y^i, which are downsampled 2^{l-i} times from X and Y, respectively. Each branch starts with shallow feature extraction by four normal convolution blocks. Then, three PRDBs (will be elaborated in Section 3.3) are inserted to extract rich features with diverse scales. Colour distributions of the reference image are injected into the first PRDB as the colour hints. After a deconvolution block, a convolution block and a decoder, the network can generate the predicted ab channels of $\hat{Y}^i$.

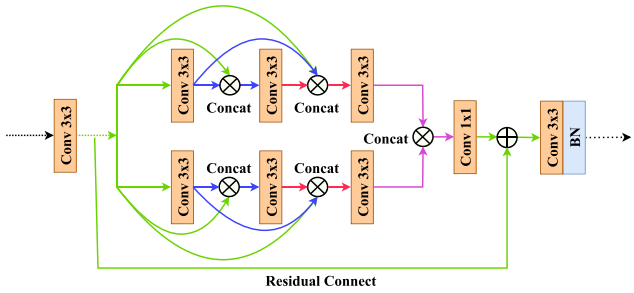
To aggregate the features between two adjacent levels, i.e. $\mathcal{F}^{i-1}$ and $\mathcal{F}^i$ ($i > 1$), we further propose an HDEF (will be elaborated in Section 3.4). HDEFs, which are associated with PRDBs, build connections between the adjacent two levels to pass on the information from the lower level to higher level. Note that there is no aggregation at the lowest level when $i = 1$. As a result, the assembled DEPN is defined as follows:

$$\mathcal{F}(G, X) = \begin{cases} \mathcal{F}^i(X^i, Y^i), i = 1 \\ \mathcal{F}^i(X^i, Y^i, \Theta(\mathcal{F}^{i-1})), i > 1, \end{cases} \tag{3}$$

where Θ indicates HDEF. We train the network progressively so that the network will be more adaptive to multi-scale colourization, confirming high robustness in high resolution. At the last level $\mathcal{F}^l$, the ab value of $\hat{X}^l$, which is decoded from Z^l, is upsampled to the resolution of the original input image by bilinear interpolation. $\hat{X}$ combining with the L channel of X forms the output colour image. In this way, our pyramid structure allows assigning accurate and semantically correct colours. We demonstrate its effectiveness quantitatively in Section 4.4.



(a) Chained residual dense block



(b) Parallel residual dense block

Figure 3: *Structure of chained residual dense block [ZTK*18] and our proposed parallel residual dense block. For simplification, the ReLU activation layers are hided in all conv 3*3 blocks.*

3.3. Parallel residual dense block

We propose the PRDB to exploit diverse features of $\mathcal{F}^i$. The proposed PRDB is used to extract the rich feature representations of images including input image, reference image and output image. It is advantageous to our colourization application compared with original residual block [HZRS16] and recently proposed chained RDB [ZTK*18] (Figure 3a).

The residual and dense structure is beneficial to memorize the low-level information and extract rich features than the normal convolution. Although these existing models [HZRS16, ZTK*18] are capable of extracting rich semantic information using deeper networks, such a deep architecture may not be helpful for our colourization process (see Section 4.4 for the relevant experiments). This is because colourization relies equally on low- and high-level knowledge for producing semantically correct and varied colour images. Taking full advantage of multi-scale features is more important than learning more redundant features. In this regard, instead of making the network deeper as the chained RDB does, we make our network wider by PRDB, which aims at aggregating more information from the parallel dense structure. As shown in Figure 3(b), unlike chained RDB, our proposed PRDB contains two densely connected structures extracting the features in parallel. The features of two branches are concatenated and fused by a 1×1 convolution layer. Since the features from two parallel branches are concatenated, our PRDB is to separately learn a hyper feature, each branch corresponds to half of it. It shortens the required information flow within the block and thus reduces the learning ambiguity. In addition, we add a convolution layer at the beginning of PRDB for trimming the foregoing block, and another convolution layer followed by the batch normalization at the end for integrating information and fast convergence.

In each PRDB, each convolution layer except the fusion layer is set to kernel size of 3×3 and padding size of 1 to fix the size and followed by the ReLU [GBB11]. The fusion layer is set to a kernel size of 1×1 without the ReLU layer. Input and output layers of PRDB receive r_0 feature maps. Densely connected layers receive $r_0 + (l-1) \times r$ feature maps, where r denotes growth rate [HLWvdM17]. We increase r to 256 gaining wider layers by only paying negligible computational complexity, as PRDB is merely applied in the small size of features of each level. In Section 4.4, we conduct an experiment to compare the performance of chained RDB and PRDB.

3.4. Hierarchical decoder–encoder filter

Similar to the encoder–decoder structures [BKC17, MSY16, CZK*17], our pyramidal network is a global encoder–decoder. The encoder injects low-level information into high-level features, while the decoder restores the chrominance of final results from the classified value. However, the error from the lower level will be accumulated, and thus increasing the redundant information. To avoid error accumulation between adjacent levels, we define the HDEF, an asymmetric decoder–encoder, as the connection of two adjacent levels to pass on information from $\mathcal{F}^{i-1}$ to $\mathcal{F}^i$ $(i > 0)$ effectively. Meanwhile, the encoder is applied to transfer the colour values into the colour distributions. The decoder is able to produce the result Y^i of each level.

3.4.1. *Encoder*

Similar to Zhang *et al.* [ZIE16], we encode the dense colour information to colour distributions by applying soft-encoding. It allows the network to distill knowledge more quickly than one-hot encoding in the classification problem [HVD15]. The encoder maps an image to colour distributions by searching the K-nearest neighbours of bins weighted by a radial basis function (RBF) kernel. Noted that the method [ZIE16] considers only single level soft-encoding and it is just used for calculating the loss function. In our network, we adopt the hierarchical soft-encoder to support our pyramid structure. For example, we want to encode the colour distribution Z^i from reference image Y^i at ith level, the encoder is defined as follows:

$$Enc^i\left(Y^i_{h,w}\right) = \left\{ \begin{array}{c} \frac{RBF\left(Y^i_{h,w}, Z^i_{h,w,k}\right)}{\sum_k^{|K^i_{h,w}|} RBF\left(Y^i_{h,w}, Z^i_{h,w,k}\right)}, Z^i_{h,w,k} \in K^i \\ 0, else \end{array} \right\}, \quad (4)$$

where K^i is the total number of nearest neighbours to Y^i at the ith level. The RBF is defined as follows:

$$RBF\left(Y^i_{h,w}, Z^i_{h,w,k}\right) = \exp\left(-\frac{\|Y^i_{h,w} - Z^i_{h,w,k}\|^2}{2\sigma^2}\right). \quad (5)$$

In practice, we let $\sigma = 5$.

Inspired by the work [CK17], we reduce $|K^i|$ gradually from $\mathcal{F}^1$ to $\mathcal{F}^l$ instead of keeping it constantly. It is because while decreasing the i, the resolution of I^i will also decrease. The image I^i will be rougher with fewer details. Thus, we increase $|K^i|$ in the lower resolution to encourage a larger searching field in order to increase the variation of colour distributions. At higher resolutions, images contain more details which means the colour distributions should be more accurate, and therefore we set $|K^i|$ smaller. We find that the hierarchical encoder not only performs well in the supervision of multi-scale and extraction of reference features (Section 3.5), but also encourages progressive aggregation while combining with the decoder.

3.4.2. *Decoder*

Considering each level $\mathcal{F}^i$, decoder converts the classification result Z^i back to ab channels. We utilize softmax function to produce the probability $P(Z^i_{h,w})$ at each pixel as follows:

$$P\left(Z^i_{h,w}\right) = \frac{\exp\left(Z^i_{h,w}\right)}{\sum_{q=1}^{|Q|} \exp\left(Z^i_{h,w,q}\right)}. \quad (6)$$

The mean of the distribution in Q is calculated by the inner product between the classified probability $Z^i_{h,w}$ of each pixel and Q:

$$Dec^i\left(Z^i_{h,w}\right) = P\left(Z^i_{h,w}\right) \odot Q. \quad (7)$$

3.4.3. *Encoder–decoder assembling*

The encoder maps the colour image to colour distributions, and the decoder is the opposite. However, these two processes are not reversible. That is because the hierarchical encoder utilizes soft-encoding scheme, forming a sparse vector while sacrificing the dense colour information. We assemble the hierarchical decoder and encoder together as a filter in the progressive aggregation as follows:

$$\Theta(Z^i) = Enc^i(Dec^i(Z^i)). \quad (8)$$

Instead of directly pass the colour distribution into the next network branch, our HDEF first decodes colour distributions to colour information then encodes it to dense features. This process is capable to avoid the error accumulation from all the preceding levels. On the other hand, the dense output from Enc^i is helpful to exploit the deep information as guidance for the next network branch.

3.5. Reference image features extraction

We extract the colour distributions from the reference image Y without luminance channel L, and only ab channels are used to guide the colourization. Due to the multi-scale nature of our network, the reference image is also required to be downsampled several times for each specific level. We transform the reference image into the colour histogram by the encoder in Section 3.4.1. After that, deep features of the statistical results are extracted by the PRDBs. Colour distributions of the reference image at a specific resolution will be passed into each network branch to guide the colourization process.

During the training phase, we use the ground truth colour image as the reference image to give the supervision of the network. While in the testing phase, there is no limit for the reference image.

3.6. Implementation and training details

At each level $\mathcal{F}^i$, the first four convolution blocks extract shallow features with a 3×3 size kernel, each of which keeps two normal convolution layers followed by ReLU with the same feature channels. In the DEPN-trunk, the number of feature maps is doubled while the resolution of the block is reduced. The convolution blocks are followed by three PRDBs to explore local–global features. After PRDBs, a single deconvolution layer is adopted to enlarge the resolution with a 3×3 kernel, two stride and one padding side. The specific number of feature maps in any block can be found in Figure 2.

From $\mathcal{F}^{i-1}$ to $\mathcal{F}^i$ ($i > 1$), the resolution of input image is doubled. We set 64×64 as the input resolution of $\mathcal{F}^1$ and resize the resolution progressively. In order to compare with the method [ZIE16], the number of classification in Q is set as 313. For the dynamic $|K^i|$ of the hierarchical encoder, we use 20 for $\mathcal{F}^1$, 15 for $\mathcal{F}^2$ and 10 while $i > 2$.

We train our network on the ADE20K [ZZP*17] data set using the Adam solver [KB14] with a batch size of 5. The network was trained with the learning rate of 3e−5. At the beginning, DEPN is trained in $\mathcal{F}^1$ without aggregation from the preceding level. We then fine-tuning $\mathcal{F}^2$ with the training parameters θ^1 from $\mathcal{F}^1$ and treat the result $\hat{X}^1$ as another input except for G and X. Well trained parameters θ^2 are capable to be applied in the larger scale level while testing.

4. Experimental Results

We conducted extensive experiments to quantitatively and qualitatively evaluate the performance of the proposed DEPN. We collect hundreds of images from the Internet with various resolutions and types for evaluation. Since our network has no restriction on the input image and reference image, we prepare a set of real colour images as well as randomly generated colour palettes as our reference images. We first compare our method with state-of-the-art colourization methods (Section 4.1). Then, we assess the robustness of the proposed DEPN using various reference images, including randomly generated colour palettes (Section 4.2). We further perform three user studies to measure colourized results in terms of naturalness, faithfulness to the reference and fidelity to the real image (Section 4.3). A set of ablation studies are also performed to evaluate the effectiveness of key components of our DEPN, including the proposed pyramid structure, PRDB and HDEF (Section 4.4).

4.1. Qualitative comparisons

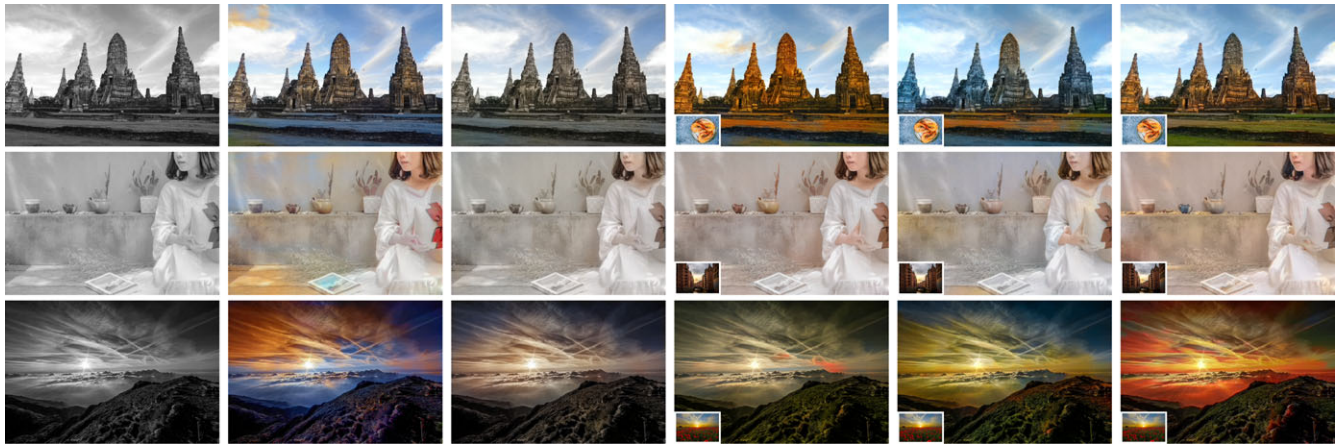
First, we compare our results with four state-of-the-art deep learning-based colourization methods, including two fully automatic methods [ZIE16, LMS16] and two example-based methods [ZZI*17, HCL*18]. Since Zhang *et al.* [ZIE16] and Larsson *et al.* [LMS16] colourize image without any hint, only one greyscale image is fed into their networks. Zhang *et al.* [ZZI*17], He *et al.* [HCL*18] and our method have exactly the same input. The objective is to obtain results with both good naturalness and reference faithfulness.

Figure 1 shows the results generated by five different methods. We can find some obvious colour drifting artefacts appear in results generated by Zhang *et al.* [ZIE16] in Figure 1(b). As can be seen, results from Larsson *et al.* [LMS16] (Figure 1c) lack colour variegation. In Figure 1(d), although the reference image provides rich enough colours, blue colour in the sky still leak onto the haystack in the first row. The result in the second row of Figure 1(e) is dull, this is mainly because He *et al.* [HCL*18] relies heavily on semantic matching (which fails for the colour palette). Results show that our method generates the colourized images (Figure 1f) with more vivid colours and less colour drifting artefacts. The results reveal that pyramidal hierarchy features extracted by our proposed DEPN are capable of producing more semantically correct colours.

Figure 4 shows more results. Note that the networks [ZZI*17, HCL*18] and ours are guided by real images while Zhang *et al.* [ZIE16] and Larsson *et al.* [LMS16] have no reference image. As can be seen, Zhang *et al.* [ZIE16] still cannot provide stable and natural enough results. Such as the first row in Figure 4(b), the grass is colourized in blue incorrectly and the colour drifting appears in the sky. Larsson *et al.* [LMS16] in Figure 4(c) tends to generate images with less vivid colours. The results from these two fully automatic methods are unsatisfactory because they lack additional hints. In Figure 4(d), Zhang *et al.* [ZZI*17] cannot suppress the colour drifting artefacts even with a real image as their global hint due to the lack of local hints in their method. He *et al.* [HCL*18] fails to predict plausible colours in some local areas when there is no semantic correspondence between the greyscale image and the reference image. Especially in the first row of Figure 4(e), the colour of the buildings and the grass looks not natural. On the contrary, our network follows the colour distributions of the reference image, and consider semantic correctness in the colourized image. Thus we can handle semantically different reference and greyscale images. Obviously, Figure 4(f) demonstrates that our colourization results are much more natural and vivid than the results of the other four methods. In addition, our results have the least artefacts.

We further compare with two traditional example-based methods, Gupta *et al.* [GCR*12] (Figure 5c) and Welsh *et al.* [WAM02] (Figure 5d), respectively. Figure 5(a) shows the reference images. The major drawback of traditional example-based methods is that the reference image has to be as similar as possible with the greyscale image. Otherwise, colour drifting may appear, such as Figures 5(c) and (d). Same issue also occurs in He *et al.* [HCL*18] in Figure 5(e) since they also require the scene of the reference image and the greyscale image are more or less the same. Our method (Figure 5f) is able to generate faithful colour image even the reference and greyscale image are not similar.

Figure 6 shows the results guided by randomly generated colour palettes. Since there is no semantic correspondence between the colour palette and the greyscale image, the results in Figure 6(c) from He *et al.* [HCL*18] are unsatisfactory. We can find that even with such sparse colour information, our method is still able to



(a) Greyscale (b) [ZIE16] (c) [LMS16] (d) [ZZI*17] (e) [HCL*18] (f) Ours

Figure 4: *Results with real images as references. Zhang et al. [ZIE16] and Larsson et al. [LMS16] do not require any additional input. We feed the same reference images into the networks [ZZI*17, HCL*18] and our network.*

(a) Reference (b) [ZZI*17] (c) [GCR*12] (d) [WAM02] (e) [HCL*18] (f) Ours

Figure 5: *Comparisons with example-based colourization methods. These are also the example images of our faithfulness user study.*

(a) Greyscale (b) [ZZI*17] (c) [HCL*18] (f) Ours

Figure 6: *Results with colour palettes as references.*

generate more reasonable and vivid results than the networks [ZZI*17, HCL*18].

4.2. Results with various references

Figure 7 further shows some results of one greyscale with different reference images. Figures 7(a) are the input greyscale images. Figures 7(b–e) are our results guided by various images. We can observe that our method can provide diverse and visually pleasant results while keeping semantic correctness, especially for the arts. In the last row of Figure 7, we can find that our method can generate diverse styles of the trees from different colour schemes, just like the scenes in different seasons. It reveals that our method can be applied in colour transfer and even does not require the input image to keep original chrominance values, which is simpler than existing colour transfer methods [CFL*15, WZL*17].

4.3. User study

We further conduct three user studies using Amazon Mechanical Turk, to evaluate the performance of our method with respect to the naturalness, the faithfulness of reference and the fidelity of our results against existing colourization methods [LMS16, ZIE16, WAM02, GCR*12, ZZI*17, HCL*18]. The order of images in each question is random. There were 20 participants joined our user studies.

In the first user study, we evaluated the naturalness against the state-of-the-art fully automatic colourization methods [ZIE16, LMS16] and two example-based methods [ZZI*17, HCL*18]. For the methods [ZIE16, LMS16], we randomly picked a greyscale

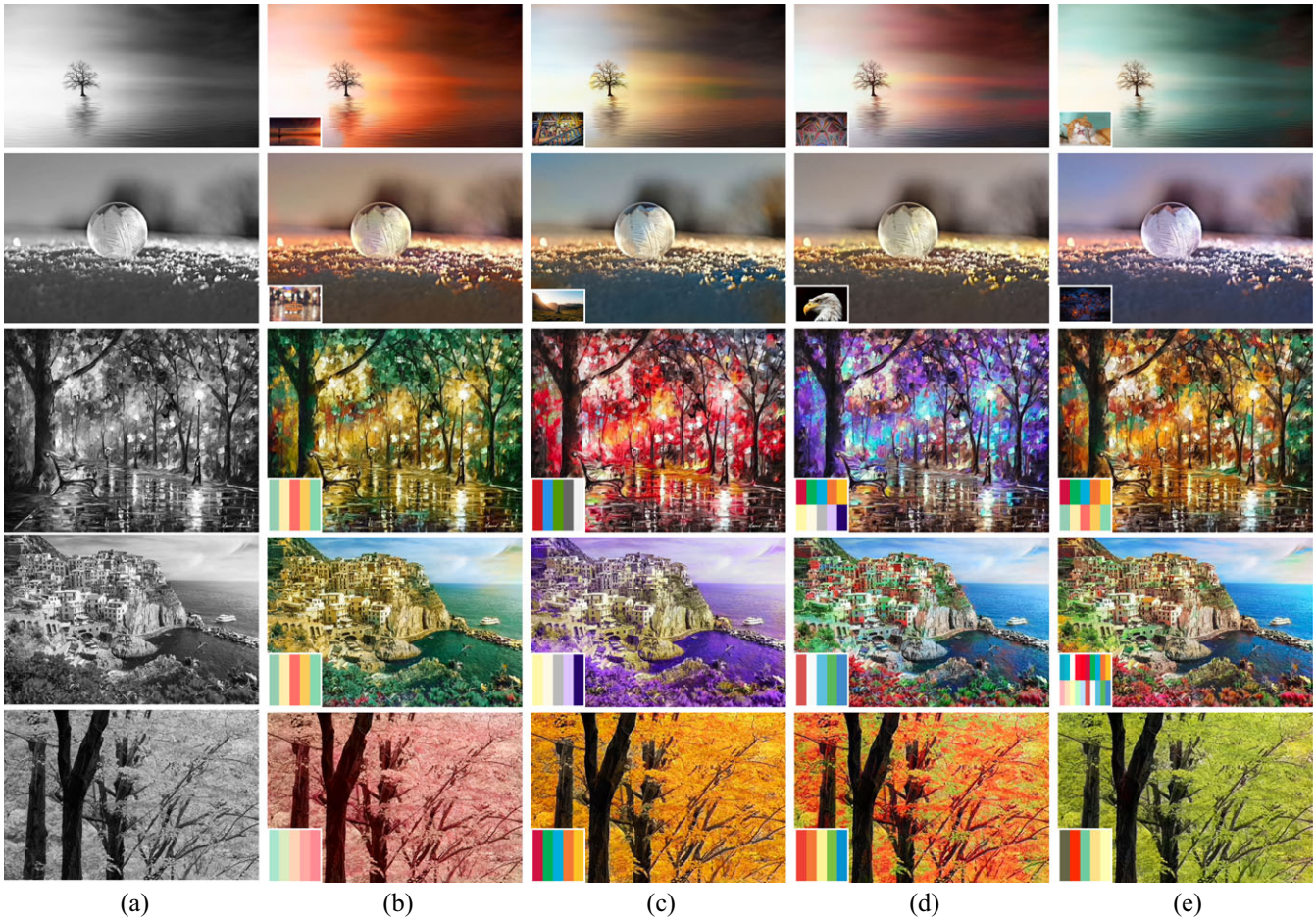


Figure 7: *The same greyscale image with various reference images. (a) Greyscale image. From (b) to (e) are the results generated guided by different reference images. Reference images in first two rows are real images while last two rows are colour palettes.*

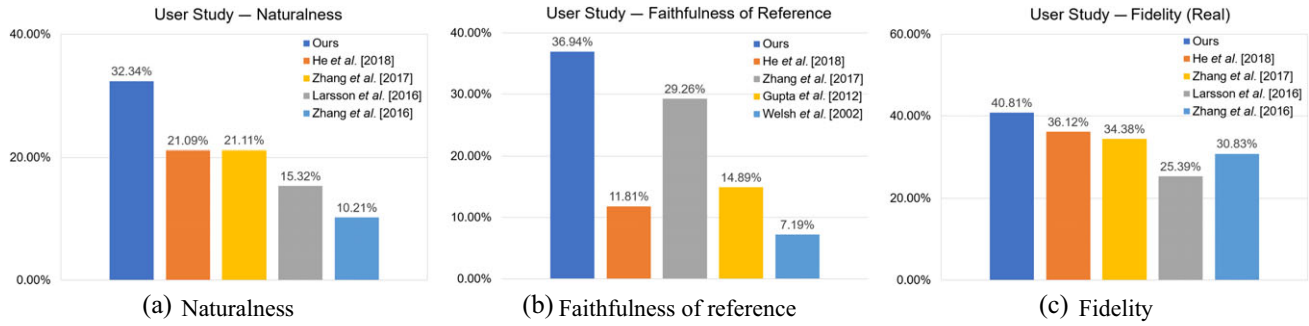
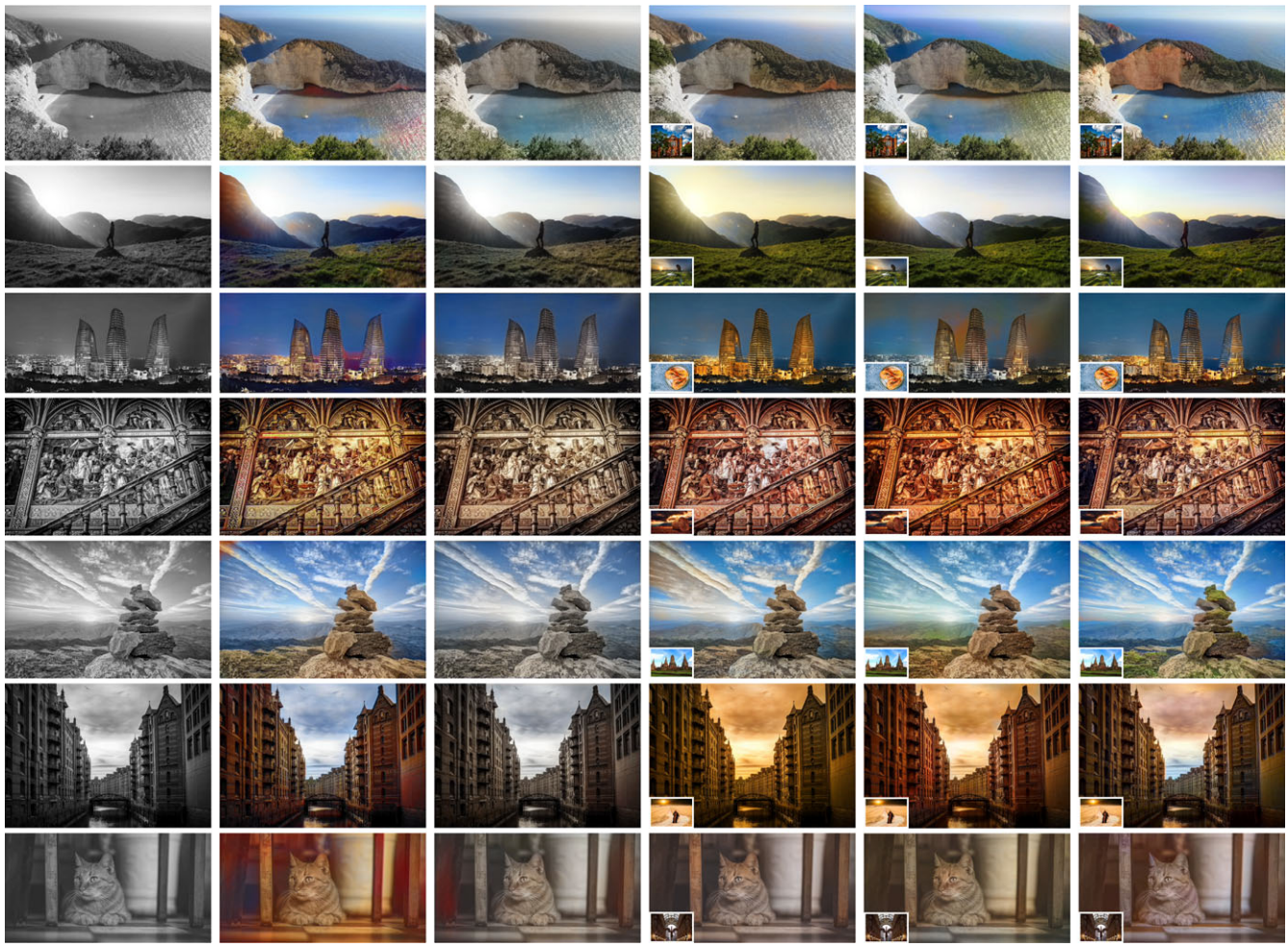


Figure 8: *Three user studies results. Each bar indicates the percentage of participants that voted for this method.*

image from our collected data. With the same greyscale image, we randomly assigned a reference image, which can be a real image or colour palette, to the methods [ZZI*17, HCL*18] and our method. This test contains 15 groups of results. Each group consists of results from the approaches [ZIE16, LMS16, ZZI*17, HCL*18] and ours, respectively. Results are in random sequence without any text hint. For each group of results, participants were asked one multiple choice question: *Which image is more natural?* Figure 8(a) shows the first user study result. Each part indicates the average number of participants that voted for this method. The result reveals that the colour images generated by our method look more natural than the existing state-of-the-art methods. Figure 9 demonstrates some selected results from user study. Note that, we did not provide a greyscale image and reference image to the



(a) Greyscale (b) [ZIE16] (c) [LMS16] (d) [ZZI*17] (e) [HCL*18] (f) Ours

Figure 9: *Selected results of our naturalness user study. The results index that most participants voted from top to bottom is: (f), (f), (f), (d), (f), (e), (e).*

participants. The results that most participants voted are shown in Figure 9.

The second one measures the results faithfulness to the reference images. We compared our results with the methods [GCR*12, WAM02, ZZI*17, HCL*18]. Similar to the first user study, we randomly picked a greyscale image and a reference image as input. The whole test contained 16 groups of results. Each group consisted of a reference image and four colourization results. Participants were asked a multiple choice question for each group: *Which image is more faithful to the reference image?* As can be seen in Figure 8(b), our method shows outstanding performance against the other three example-based colourization methods. Figure 5 shows a group of the test in this user study. The result reveals that our proposed method is able to precisely map the colour distribution while with more semantically correct colourization results.

Lastly, we conducted a *real versus fake* test in the third user study to evaluate the fidelity of recolourized images. In this experiment, we compared our method with the methods [ZIE16, LMS16, ZZI*17, HCL*18]. There are a total of four sessions in these experiments, each session for one method. Each session consisted of 16 pairs of images, a ground truth colour image and a recolourized image. Users were requested to pick the *real* image they believed in a pair of images. Participants were only allowed to finish at most one session to avoid the impression of repeated ground truth images. Figure 8(c) shows the user study of fidelity. As can be seen, among these four methods, our method gets the most votes which mean that our results are the most non-differentiable 'fake' images.

4.4. Ablation study

We set up a set of ablation studies to assess the effectiveness of key components of the proposed DEPN. In these experiments, we directly use the ground truth as the reference image to quantitatively evaluate the PSNR and SSIM [WBSS04] performances of networks with different configurations. Note that, for example-based

Table 1: *Quantitative evaluations on different resolutions with the ground truth as the reference image. We compare our network against the work [ZZI*17, HCL*18] and four variants of our network. All the networks run on 1000 held-out test images, in the ADE20K [ZZP*17] validation data set. Ours with PRDB (boldfaced values) achieves the best performance.*

Method	Resolution	PSNR (dB)	SSIM
[ZZI*17]	256*256	28.47	0.90
	512*512	27.61	0.86
	1024*1024	27.06	0.80
[HCL*18]	256*256	34.99	0.98
	512*512	34.93	0.97
	1024*1024	34.89	0.96
DEPN with conv	256*256	26.15	0.84
	512*512	26.10	0.81
	1024*1024	25.94	0.77
DEPN without HDEF	256*256	26.89	0.86
	512*512	26.70	0.83
	1024*1024	26.31	0.78
DEPN with Chained RDB	256*256	27.05	0.86
	512*512	27.07	0.83
	1024*1024	27.02	0.80
DEPN with static $K = 10$	256*256	27.39	0.87
	512*512	27.30	0.85
	1024*1024	25.44	0.72
Ours (with PRDB)	256*256	**28.61**	**0.90**
	512*512	**27.89**	**0.87**
	1024*1024	**27.47**	**0.83**

colourization methods, like ours and Zhang *et al.* [ZZI*17], the ground truth itself is the ideal image to be compared with. Thus, PSNR and SSIM results can be regarded as the faithfulness of results comparing with their ground truth.

We train four variants of our network: (1) using common convolution blocks to replace our proposed PRDBs, (2) our network without HDEF, (3) using chained RDBs to replace our proposed PRDBs and (4) using static $K = 10$ defined in Section 3.4.1. We also compare our results with those of the methods [ZZI*17, HCL*18]. Note that the network [ZZI*17] is able to perform on the image with any resolution due to the fully convolutional network design. All the networks run on 1000 held-out test images, in the ADE20K [ZZP*17] validation data set.

Table 1 presents the peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) results. Note that we only train the first two levels of the network and then share the parameters to the higher levels. As can be seen, we achieve higher performance than Zhang *et al.* [ZZI*17] while keep increasing the resolution of input image. The reason why He *et al.* [HCL*18] gets much higher PSNR and SSIM than ours and Zhang *et al.* [ZZI*17] is that when letting the ground truth be the reference image, the network of [HCL*18] can easily finds out completely the same semantic correspondence between the greyscale and reference images. Different from He *et al.* [HCL*18], Zhang *et al.* [ZZI*17] and our method only extract the colour distribution of the reference image without any semantic information.

Comparing with four variants of our network, DEPN with PRDBs outperforms DEPN with common convolution block and DEPN with the chained RDB. It proves that our proposed PRDBs are able to embed richer local–global context information than the other two existing blocks, enhancing the semantic understanding of the colourization process. The PSNR and SSIM results reveal that the HDEF is indispensable to our network. Dynamic K (defined in Section 3.4.1) is also proved more effective on our hierarchical network.

Table 2: *Quantitative evaluation on the proposed DEPN with different levels. We fix the input resolution to 512*512 for a better comparison. The reference image is the ground truth itself. Four levels (boldfaced values) achieve the best performance.*

Number of levels	Resolution	PSNR (dB)	SSIM
DEPN with single level	512*512	24.49	0.75
DEPN with two levels	512*512	27.01	0.84
DEPN with three levels	512*512	27.36	0.85
DEPN with four levels (Ours)	512*512	**27.89**	**0.87**
DEPN with five levels	512*512	27.87	0.86
DEPN with six levels	512*512	27.88	0.86

We conduct another experiment to further evaluate the effectiveness of our pyramid structure. The training configurations are all the same as previous. We train the network with various levels, single level, two-level, three-level, four-level, five-level and six-level, respectively. Table 2 demonstrates the PSNR and SSIM values under different levels of our network. As can be seen, without pyramid structure (DEPN with a single level in Table 2), PSNR and SSIM value drastically decrease. While increasing the number of levels, both PSNR and SSIM reveal better results. Note that the network parameters are shared across levels 2 to 6 during both training and testing, which reveals that the improvement is not because of the increased number of parameters. Thus this experiment evidences that the pyramid structure allows our network to learn more semantically correct contextual information than a simple end-to-end network. As shown in Table 2, there is no obvious breakthrough of performance while increasing the level to 5 or even 6. To balance the trade-off between performance and computational time, we choose four levels for our network in practice.

Qualitative evaluation of our ablation study is shown in Figure 10. The results without the proposed PRDBs are shown in Figure 10(a). The problem of using convolution only is that it cannot capture sufficient semantic information to understand the greyscale image, and therefore leading to colour drifting artefacts (see the tree in the first example) and murky colours. The proposed pyramid structure is demonstrated in 10(b). Using a single level of image structure makes the network concentrates to local regions. This is why the generated results contain inconsistent and incontinuous colours. The proposed HDEF bridges different scales of information better, and thus avoid inaccurate information transmission across different scales of networks (see the building the first example, and the sky in the second example of 10c). Finally, with all these components, the proposed method achieves comparable results to the ground truth (10d and e).

4.5. Performance

We construct our network using the deep learning framework Caffe on Ubuntu 16.04. The whole training process took around 96 h on

(a) DEPN with conv (b) DEPN (single level) (c) DEPN w/o HDEF (d) DEPN (all) (e) GT

Figure 10: *Qualitative comparisons of our ablation study. Here, we use ground truth as the reference image.*

(a) Reference image (b) Our result

Figure 11: *One limitation is that our method cannot guarantee the semantically colour correctness of same object between reference image (a) and our result (b), like the first row. The last two rows are unnatural results. They cannot be considered as good colourization results because they contradict with human cognition.*

a single NVIDIA GeForce GTX 1070 with an Intel Core i7-7700K CPU at 4.20 GHz. It takes around 1.4 s for colourizing one image (512*512) in average.

4.6. Limitations

Since our network only consider the colour distribution of the reference image. The network lacks semantic correspondence between the reference image and the greyscale image. That is why we cannot guarantee the semantically colour correctness of the same objects from the reference image to the greyscale image, like the result in the first row of Figure 11(b).

Similar to the other learning-based methods, another limitation of our method is that the unusual colourization results when given inappropriate references. It is hard to balance naturalness and reference faithfulness for example-based methods. When the colour distribution of the reference image is not suitable for the specific scene, our network may generate some results that contradict with human cognition. For example, in the last two rows of Figure 11(b), it is obvious that the skin of the woman and the feather of the eagle are unusual.

5. Conclusion

In this paper, we propose a dense encoding pyramid network for image colourization. We leverage large-scale data to encode the latent colour information into our network. A reference image is utilized to guide the colourization by analysing the colour distribution of it. Due to the flexibility of our network, the reference image can be any colour image even a simple colour palette. Our proposed PRDB can effectively extract local–global colour context. The HDEF is proposed to pass the colour distribution from lower level to higher level. The network provides ab channels as the final output. Combining with the given L channel, we can get the colourized image. Experimental results show that our method is comparable and superior to state-of-the-art methods both in visual pleasant and colour varieties.

Acknowledgements

This project is supported by the National Natural Science Foundation of China (No. 61472145 and No. 61702194), Shenzhen Science and Technology Program (No. JCYJ20160429190300857), RGC General Research Fund No. CUHK14201017, the Innovation and Technology Fund of Hong Kong (Project No. ITS/319/17), the Special Fund of Science and Technology Research and Development

on Application From Guangdong Province (SF-STRDA-GD) (No. 2016B010127003), the Guangzhou Key Industrial Technology Research fund (No. 201802010036), and the Guangdong Natural Science Foundation (No. 2017A030312008).

References

[BKC17] Badrinarayanan V., Kendall A., Cipolla R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 12 (2017), 2481–2495.

[BTP14] Bugeau A., Ta V.-T., Papadakis N.: Variational exemplar-based image colorization. *IEEE Transactions on Image Processing 23*, 1 (2014), 298–307.

[CFL*15] Chang H., Fried O., Liu Y., DiVerdi S., Finkelstein A.: Palette-based photo recoloring. *ACM Transactions on Graphics 34*, 4 (2015), 139:1–139:11.

[CK17] Chen Q., Koltun V.: Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy, 2017), vol. 1.

[CPK*18] Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A. L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 4 (2018), 834–848.

[CYS15] Cheng Z., Yang Q., Sheng B.: Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision* (Las Condes, Chile, 2015), pp. 415–423.

[CZK*17] Chen H., Zhang Y., Kalra M. K., Lin F., Chen Y., Liao P., Zhou J., Wang G.: Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging 36*, 12 (2017), 2524–2535.

[GBB11] Glorot X., Bordes A., Bengio Y.: Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Ft. Lauderdale, FL, USA, 2011), pp. 315–323.

[GCR*12] Gupta R. K., Chia A. Y.-S., Rajan D., Ng E. S., Zhiyong H.: Image colorization using similar images. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan, 2012), ACM, pp. 369–378.

[GD05] Grauman K., Darrell T.: The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of Tenth IEEE International Conference on Computer Vision, ICCV 2005*. (Beijing, China, 2005), vol. 2, IEEE, pp. 1458–1465.

[GDB*17] Guadarrama S., Dahl R., Bieber D., Norouzi M., Shlens J., Murphy K.: Pixcolor: Pixel recursive colorization. arXiv preprint arXiv:1705.07208, 2017.

[HCL*18] He M., Chen D., Liao J., Sander P. V., Yuan L.: Deep exemplar-based colorization. *ACM Transactions on Graphics 37*, 4 (2018), 47:1–47:16.

[HLWvdM17] Huang G., Liu Z., Weinberger K. Q., van der Maaten L.: Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, USA, 2017), vol. 1, p. 3.

[HVD15] Hinton G., Vinyals O., Dean J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[HZRS14] He K., Zhang X., Ren S., Sun J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of European Conference on Computer Vision* (Zurich, Switzerland, 2014), Springer, pp. 346–361.

[HZRS16] He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, USA, 2016), pp. 770–778.

[ICOL05] Ironi R., Cohen-Or D., Lischinski D.: Colorization by example. In *Proceedings of Eurographics Conference on Rendering Techniques* (Konstanz, Germany, 2005), Citeseer, pp. 201–210.

[ISSI16] 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 35*, 4 (2016), 110:1–110:11.

[KB14] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[LLW04] Levin A., Lischinski D., Weiss Y.: Colorization using optimization. *ACM Transactions on Graphics 23* (2004), 689–694.

[LMS16] Larsson G., Maire M., Shakhnarovich G.: Learning representations for automatic colorization. In *Proceedings of European Conference on Computer Vision* (Amsterdam, The Netherlands, 2016), Springer, pp. 577–593.

[LSP06] Lazebnik S., Schmid C., Ponce J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (New York, NY, USA, 2006), IEEE, pp. 2169–2178.

[LWCO*07] Luan Q., Wen F., Cohen-Or D., Liang L., Xu Y.-Q., Shum H.-Y.: Natural image colorization. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Grenoble, France, 2007), Eurographics Association, pp. 309–320.

[MSY16] Mao X., Shen C., Yang Y.-B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Proceedings of International Conference on Neural Information Processing Systems* (Barcelona, Spain, 2016), pp. 2802–2810.

[QWH06] Qu Y., Wong T.-T., Heng P.-A.: Manga colorization. *ACM Transactions on Graphics 25* (2006), 1214–1220.

[RAGS01] Reinhard E., Adhikhmin M., Gooch B., Shirley P.: Color transfer between images. *IEEE Computer Graphics and Applications 21*, 5 (2001), 34–41.

[RDS*15] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei L.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252.

[RKB04] Rother C., Kolmogorov V., Blake A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics 23* (2004), 309–314.

[RKL17] Royer A., Kolesnikov A., Lampert C. H.: Probabilistic image colorization. arXiv preprint arXiv:1705.04258, 2017.

[SBv04] Sýkora D., Buriánek J., Žára J.: Unsupervised colorization of black-and-white cartoons. In *NPAR '04: Proceedings of the 3rd International Symposium on Non-Photorealistic Animation and Rendering* (Annecy, France, 2004), ACM, pp. 121–127.

[TJT05] Tai Y.-W., Jia J., Tang C.-K.: Local color transfer via probabilistic segmentation by expectation-maximization. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*. (San Diego, CA, USA, 2005), vol. 1, IEEE, pp. 747–754.

[WAM02] Welsh T., Ashikhmin M., Mueller K.: Transferring color to greyscale images. *ACM Transactions on Graphics 21* (2002), 277–280.

[WBSS04] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612.

[WDK*13] Wu F., Dong W., Kong Y., Mei X., Paul J.-C., Zhang X.: Content-based colour transfer. *Computer Graphics Forum 32* (2013), 190–203.

[WZL*17] Wang D., Zou C., Li G., Gao C., Su Z., Tan P.: L0 gradient-preserving color transfer. *Computer Graphics Forum 36* (2017), 93–103.

[XYJ13] Xu L., Yan Q., Jia J.: A sparse control model for image and video editing. *ACM Transactions on Graphics 32*, 6 (2013), 197:1–197:10.

[YYGH09] Yang J., Yu K., Gong Y., Huang T.: Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009* (Miami, FL, USA, 2009), IEEE, pp. 1794–1801.

[ZIE16] Zhang R., Isola P., Efros A. A.: Colorful image colorization. In *Proceedings of European Conference on Computer Vision* (Amsterdam, The Netherlands, 2016), Springer, pp. 649–666.

[ZLL*18] Zhang Y., Li K., Li K., Wang L., Zhong B., Fu Y.: Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany, 2018), pp. 286–301.

[ZLX*14] Zhou B., Lapedriza A., Xiao J., Torralba A., Oliva A.: Learning deep features for scene recognition using places database. In *Proceedings of International Conference on Neural Information Processing Systems* (Montreal, Canada, 2014), pp. 487–495.

[ZSQ*17] Zhao H., Shi J., Qi X., Wang X., Jia J.: Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, 2017), pp. 2881–2890.

[ZTK*18] Zhang Y., Tian Y., Kong Y., Zhong B., Fu Y.: Residual dense network for image super-resolution. arXiv preprint arXiv:1802.08797, 2018.

[ZZI*17] Zhang R., Zhu J.-Y., Isola P., Geng X., Lin A. S., Yu T., Efros A. A.: Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics 36*, 4 (July 2017), 119:1–119:11.

[ZZP*17] Zhou B., Zhao H., Puig X., Fidler S., Barriuso A., Torralba A.: Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, USA, 2017).