Image Inpainting with External-internal Learning and Monochromic Bottleneck

Tengfei Wang*   Hao Ouyang*   Qifeng Chen
The Hong Kong University of Science and Technology

Abstract

Although recent inpainting approaches have demonstrated significant improvement with deep neural networks, they still suffer from artifacts such as blunt structures and abrupt colors when filling in the missing regions. To address these issues, we propose an external-internal inpainting scheme with a monochromic bottleneck that helps image inpainting models remove these artifacts. In the external learning stage, we reconstruct missing structures and details in the monochromic space to reduce the learning dimension. In the internal learning stage, we propose a novel internal color propagation method with progressive learning strategies for consistent color restoration. Extensive experiments demonstrate that our proposed scheme helps image inpainting models produce more structure-preserved and visually compelling results. Our source code is available at https://github.com/Tengfei-Wang/external-internal-inpainting.

1. Introduction

Image inpainting is a task that aims to complete the missing regions of an image with visually realistic and semantically consistent content. Image inpainting can benefit general users in various practical applications, including unwanted object removal from an image, face defect removal, and image editing. While we have witnessed significant progress in image inpainting, inpainting models still suffer from abrupt color artifacts, especially when the missing regions are large. This work will analyze the weaknesses of state-of-the-art inpainting approaches and present a novel framework to improve existing inpainting methods.

State-of-the-art inpainting methods roughly fall into two categories of patch matching by iteratively nearest-neighbor search and deep learning models, with different pros and cons. PatchMatch [3] is a learning-free method that only utilizes internal statistics of a single image. As shown in Fig. 1, it generates smooth patterns and colors that are consistent with the non-missing region, but it fails to fill in

---

*Equal contribution

semantic-aware content. The deep learning based inpainting approaches can learn semantic-aware models by training on large-scale datasets. These approaches have explored coarse-to-fine inpainting models in different fashions. They may first generate edges [20, 17], structural information [24], segmentation maps [29] or blurry images [37, 38, 36], and then use these intermediate outputs as guidance for filling in details. However, their results still suffer from color and texture artifacts. One of the most common artifacts observed is color bleeding, as shown in Fig. 1. These methods trained on a large-scale dataset tend to introduce inconsistent colors that do not conform to the color distribution of the test image. On the other hand, we observe that color bleeding artifacts seldom appear in the internal methods.

Based on the observations above, we propose a robust inpainting method by combining the best of both worlds. We adopt a novel external-internal inpainting scheme with a monochromic bottleneck: first completing the monochromic image via learning externally from large-scale datasets and then colorizing the completed monochrome by learning internally on the single test image. Our proposed method is orthogonal to early inpainting approaches and thus can be easily applied to improve previous learning-based inpainting models for a higher-quality generation. In the external learning stage, by changing the output of the reconstruction network from polychromatic images to monochromic images, we reduce the dimension of the optimization space from $\mathbb{R}^3$ to $\mathbb{R}$, leading to more structure-preserving reconstruction (Section 4.3). Models trained in this way also show stronger generalization ability on cross-dataset evaluation. In the colorization stage, motivated by the recent advancement in deep internal learning, we propose a novel internal color propagation approach guided by the completed monochromic bottleneck. However, similar monochromic values can map to different polychromic outputs even in a single image. We, therefore, adopt a progressive restoration strategy for combining both local and global color statistics. Our external-internal learning scheme not only facilitates structure reconstruction but also ensures color consistency. By focusing on the internal color distribution of a single image, we can eliminate

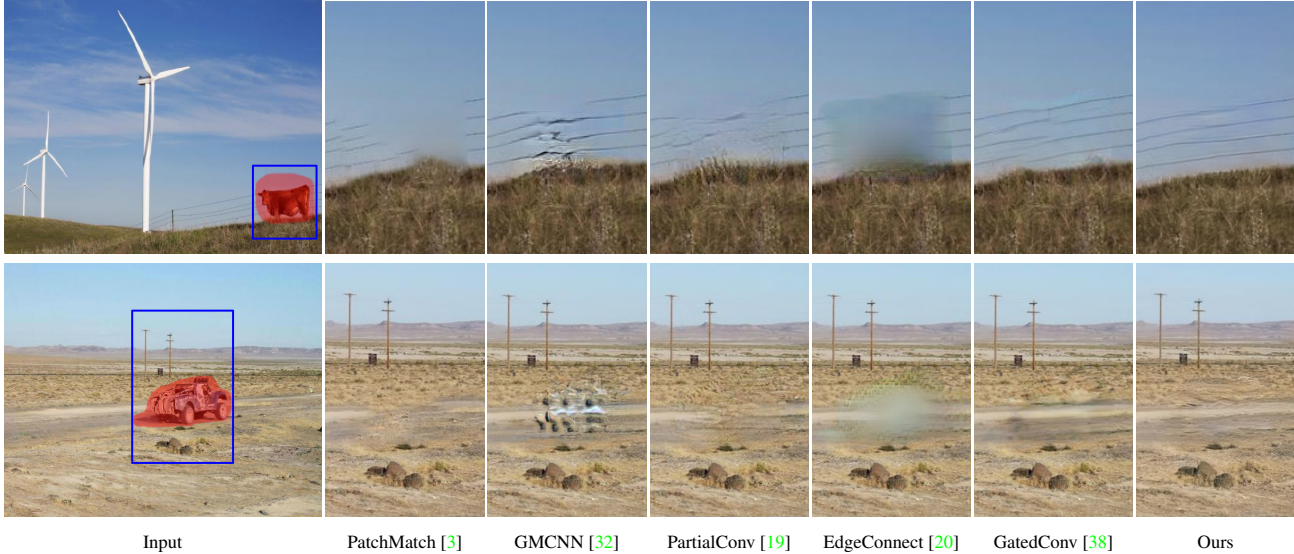| Input | PatchMatch [3] | GMCNN [32] | PartialConv [19] | EdgeConnect [20] | GatedConv [38] | Ours |

Figure 1. Image inpainting results by traditional and deep learning methods. Zoom in for details.

abrupt colors and produce a visually pleasing image (Section 3.1.1).

We conduct extensive experiments to evaluate the performance of our method on four public datasets Places2 [42], Paris StreetView [21], CelebA-HQ [15] and DTD [6]. We apply our method to different baseline networks (GatedConv [38], EdgeConnect [20], HiFill [36] and GMCNN [32]), and observe meaningful improvement in terms of structure preservation and color harmonization. Furthermore, we perform model analysis and ablation studies to verify our hypothesis and modifications. The main contributions of our paper can be summarized as:

- To the best of our knowledge, we are the first to introduce an external-internal learning method to deep image inpainting. It learns semantic knowledge externally by training on large datasets while fully utilizes internal statistics of the single test image.
- We design a progressive internal color restoration network that achieves outstanding colorization performance in our case.
- We generalize our proposed method to several deep inpainting models and observe clear improvement in terms of visual quality and model generalization ability on multiple datasets.

2. Related Work

2.1. Image Inpainting

Traditional learning-free image inpainting methods can be roughly divided into two categories: diffusion-based and patch-based methods. Diffusion-based methods [1, 8, 2, 9] propagate neighboring information using techniques such as isophote direction field. These methods perform well on texture data or images with narrow holes while they will fail when the masked region is large or contains meaningful structures. Patch-based methods such as PatchMatch [3] fill in the missing region by searching the patches outside the hole with a fast nearest neighbor algorithm. However, the pattern in the missing region cannot always be found in the image, and also repetitive patterns tend to appear in the reconstructed image. These methods utilize only the internal information that achieves color consistency but fails in filling in semantic-aware contents.

The recent development of deep learning has greatly improved the performance of image inpainting, especially in image categories like faces and complex natural scenes. The inpainting model benefits from learning and understanding semantic meanings from large-scale datasets [7]. Pathak et al. [21] first proposed context encoders that utilized an encoder-decoder network to extract features and reconstruct the outputs. Iizuka et al. [14] used both global and local discriminator, and Yu et al. [37] proposed the contextual attention for retrieving remote features and achieving global coherency. Liu et al. [19] applied the partial convolution, and Yu et al. [38] proposed the gated convolution to overcome the weakness of the vanilla convolution. Yi et al. [36] proposed the contextual residual aggregation module, and Zeng et al. [39] adopted a guided upsampling network for high-resolution image inpainting.

Most recent methods first predict coarse structure such as edges [20, 17], foreground contours [34], structure shape [24] and semantic maps [29], and then provide additional prior for guiding the completion of images. These methods show that conducting inpainting in a spatially coarse-to-fine way will benefit the training process. Our method also adopts a similar idea while completing images
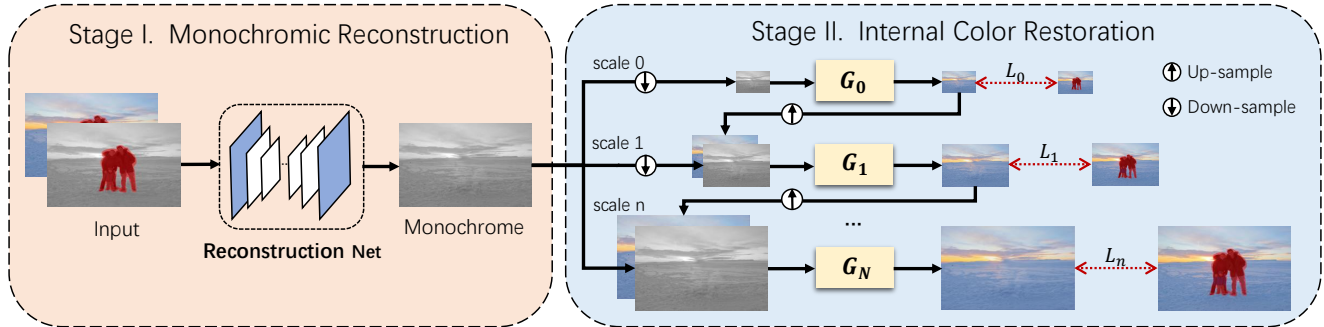
Figure 2. Overview of our external-internal inpainting method. It externally learns to reconstruct structures in the monochromic space via training on large datasets, while implicitly propagates colors within the single test image via internal learning. The colorization loss L_n is only calculated on the unmasked regions.

not only spatially but also in a channel-wise coarse-to-fine way via external-internal learning.

2.2. Guided Colorization

User-guided colorization methods focus on local input, such as user strokes or color points. The color is propagated using low-level similarity metrics based on luminance [16, 13], textures [22], and intrinsic distance [35]. In addition to local hints, Li et al. [18] utilized color theme and Chang et al. [5] used color palette for expressing global color control. Zhang et al. [41] also combined low-level cues along with high-level semantic similarities. Example-based approaches transferred color from a single or multiple reference images to the target image. These approaches, no matter using which techniques (color transfer [11, 23] or image analogies [12]), all focused on finding the correct correspondence between the reference and target images. In our internal colorizarion, we use the monochromic output from the first stage as a conditional input and thus propagate the internal color information from the non-missing region to the missing region. Different from the user-guidance and example-guidance, the guidance in our case is not only extraordinarily dense but also has accurate one-to-one correspondences, which can provide sufficient information both locally and globally. We show in the paper that existing guided colorization methods cannot fully utilize the reliable color information in the non-missing regions.

2.3. Deep Internal Learning

Training a deep convolutional neural network on only a single image has shown effectiveness in various image generation tasks such as super-resolution, texture synthesis, and so on [28, 4, 43, 27, 25]. Ulyanov et al. [31] were the first to utilize deep model as a prior to train image inpainting. They trained a deep model on the non-missing region of a single image from random Gaussian noise and try to propagate similar content information to the missing regions. However, their model fails to generate realistic details in the inpainted area. Shocher et al. [27] introduced the InternalGAN for conditional image generation. However, since our ground truth image is only partially available (the non-missing part), it is difficult to apply the adversarial training. Considering our case, we carefully design a progressive deep network for internal image colorization.

3. Method

In this section, we first analyze the drawbacks of state-of-the-art image inpainting methods and the motivation of our external-internal learning scheme. We then present the details of the two stages: external monochromic reconstruction trained on large-scale datasets for generating semantically correct content, and internal color restoration on a single image for propagating the color from non-missing parts to missing regions. The overview architecture is shown in Fig. 2. The proposed method does not conflict with existing inpainting approaches but instead completes a more coarse-to-fine procedure.

3.1. Motivation

3.1.1 Color Bleeding Removal

Early image inpainting networks trained on large datasets usually suffer from the "color bleeding" artifacts. As shown in Fig. 3, colors in the inpainted area of previous approaches [38, 20] show abrupt discrepancy from non-missing regions. For example, the green and pink color in the first image, and the purple color in the second image are very different from the color distribution of non-missing parts. This distribution gap indicates the possibility of improving inpainting quality by eliminating outliner colors in the missing region. Hence, we are motivated to further improve the color consistency by learning only from the internal color distribution of the non-missing parts.

To show the visual quality gain brought by the internal colorization, we apply our method to re-colorize the results of previous inpainting approaches. As in Fig. 3, by strength-

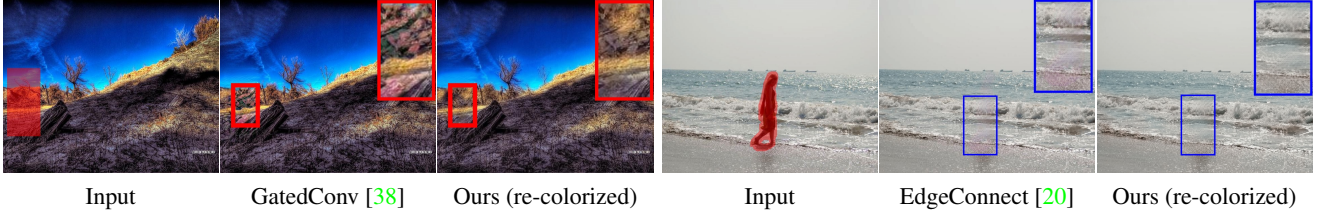| Input | GatedConv [38] | Ours (re-colorized) | Input | EdgeConnect [20] | Ours (re-colorized) |

Figure 3. Re-colorized results of applying our internal colorization method to GatedConv (left) and EdgeConnect (right). Colors of original results are defective and inconsist, while our re-colorized results are visually harmonized.

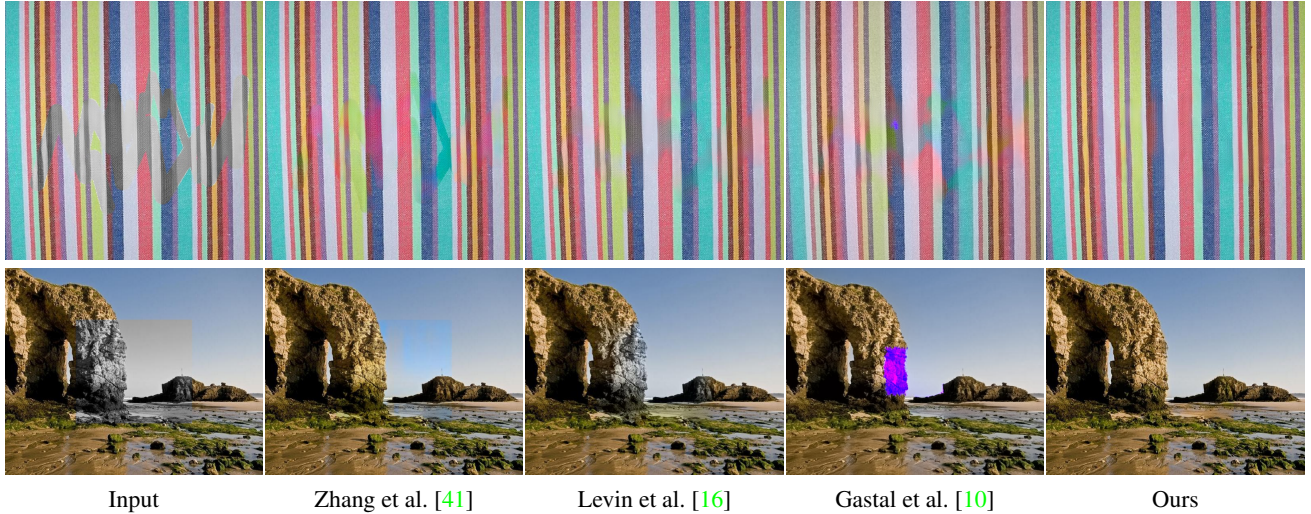| Input | Zhang et al. [41] | Levin et al. [16] | Gastal et al. [10] | Ours |

Figure 4. Visual comparison with different guided colorization methods on the inpainted monochromic bottleneck (top) and natural (w/o inpainting) monochrome (bottom). Zoom in for details.

ening the impact of internal color statistics in the single image, the abrupt colors can be eliminated.

3.1.2 External-internal Learning

However, learning only from internal statistics is inappropriate since external information is significant for content-aware image inpainting. A feasible solution is to set an intermediate bottleneck as a bridge between the external and internal learning. In traditional image reconstruction tasks, many researchers utilize monochromes to learn structures and then directly add color information back [30, 26]. Inspired from these works, we choose monochrome images as the intermediate output. This leads to another advantage that by reducing the output dimension from $\mathbb{R}^3$ to $\mathbb{R}^1$, the complexity of training is alleviated. We expect that models trained with monochromic bottlenecks can reconstruct higher-fidelity structures than original ones.

3.2. External Monochromic Reconstruction

Our method can be easily applied to improve the reconstruction quality of learning-based image inpainting models. Specifically, we concatenate the monochromic input to the original RGB input channel-wisely, and also modify the output from polychromic to monochromic images. We

experiment with representative inpainting baselines as our reconstruction network:

- **GMCNN** [32]: a generative multi-column model, which synthesizes different image components in a parallel manner.
- **HiFill** [36]: a coarse-to-fine network for high-resolution images with light-weight gated convolution.
- **EdgeConnect** [20]: a two-stage adversarial method, which hallucinates missing edges first as guidance for image completion.
- **GatedConv** [38]: a coarse-to-fine network based on gated convolution, which achieves state-of-the-art inpainting performance with free-form masks.

In our implementation, we convert an RGB image to a monochromic image by $0.30R + 0.59G + 0.11B$. For simplicity, we denote our models with different reconstruction networks as ***Ours ("backbone")***.

3.3. Internal Color Restoration

3.3.1 Guided Colorization

In this stage, the input of the colorization network is the completed monochromic bottleneck from the first stage, while the goal is to restore colors consistent with the poly-

chromic distribution of non-missing regions. We first tested with several guided colorization methods including:

- **Zhang et al.** [41]. A deep-learning based guided colorization method that learns semantic similarities from large datasets.
- **Levin et al.** [16]. A quadratic optimization method that restores colors according to similar intensities.
- **Gastal et al.** [10]. A learning-free image processing method that is based on the edge-preserving filtering .

However, as shown in Fig. 4, the external-learning method [41] tends to magnify the inaccuracy in the monochrome and introduces color bleeding artifacts (e.g. red in the first example). On the contrary, utilizing color hints internally from the same image tends to avoid being confused by external color distributions. Previous learning-free methods [16, 10] produce generally color-consistent results but fail in propagation when the mask region is large. We analyze the special features of our cases that are different from most of previous colorization settings as follows:

- Unlike traditional sparse guidance such as color stroke and color palette, the guidance in our case is multiple accurate one-to-one mappings from monochrome to RGB. Since non-missing regions usually consist of an ample amount of pixels, the correspondence is extremely dense and covers most of patterns.

- Structures in the inpainted missing region I_{hole} and the non-missing region I_{nhole} are often highly correlated.

Inspired by recent work [31], rather than exploring similarity explicitly by feature matching, we propose to utilize a deep neural network f to implicitly propagate color information. Specifically, we internally learn the color mapping function f in the non-missing regions I_{nhole} and directly apply it to the missing regions I_{hole} for colorization. However, similar monochromic inputs can map to different polychromic values even in a single image. We, therefore, design a progressive colorization network to combine the local and global color context.

3.3.2 Progressive Color Restoration

Our model consists of a conditional generator pyramid $\{G_0, G_1, ..., G_N\}$. We construct the corresponding grayscale image pyramid$\{I_0^g, I_1^g, ..., I_N^g\}$, color image pyramid $\{I_0^c, I_1^c, ..., I_N^c\}$ and mask pyramid $\{M_0, M_1, ..., M_N\}$ for internal learning. The colorization process begins at the coarsest scale and goes sequentially to the finest scale. In the coarsest scale, the model takes only the downsampled grayscale image:

$$\hat{I}_0 = G_0(I_0^g). \tag{1}$$

In the finer scale, the generator takes both the grayscale image and the upsampled color output from the lower level:

$$\hat{I}_n = G_n(I_n^g \oplus \hat{I}_{n-1} \uparrow), n = 1, ..., N \tag{2}$$

where $\oplus$ indicates concatenation in channel dimension and $\uparrow$ indicates bilinear upsampling. We adopt a ResNet-like architecture with box downsampling and bilinear upsampling for all generators. Since all the generators have the same receptive field, the model gradually captures global to local information as we process from coarse to fine.

As the ground-truth pixels are only available in the non-missing region, we thus adopt a masked reconstruction loss for each generator, formulated as:

$$L_n = ||(\hat{I}_n - I_n^c) \odot (1 - M_n)||_1, \tag{3}$$

where $\odot$ indicates the Hadamard product. We use max-pooling for downsampling when building the mask pyramid to ensure that pixels from missing regions will not be included.

4. Experiments

4.1. Datasets

We evaluate our method on four public datasets:

Places2 Standard [42] contains more than 18 million natural images from 365 scene categories. We conduct experiment on all the categories, and use the original split for training. All images are resized to 512×640 when testing.

Paris StreetView [21] contains 15,000 outdoor building images. We use the original split for training. All images are resized to 256×256 when testing.

CelebA-HQ [15] contains 30,000 face images. We randomly select 3,000 images for testing, and others for training. All images are resized to 256×256 when testing.

DTD [6] contains 5,640 texture images.We randomly select 840 images for testing, and others for training. All images are resized to 512×512 when testing.

Masks We generate dense irregular masks by the algorithm proposed in [38]. In real-use cases, users usually behave like using an eraser or brush to mask out undesired regions for inpainting. This algorithm simulates this behavior by randomly drawing lines and rotating angles, which are fair and suitable for our evaluation.

4.2. Evaluation

Quantitative Comparison As mentioned in previous work [37], there are no suitable objective metrics for inpainting tasks due to the ambiguity of ground truth. Nevertheless, we still report evaluation results in terms of PSNR, SSIM [33], and a learned perceptual metric LPIPS [40]. As shown in Table 1, for different backbone networks, the proposed external-internal scheme consistently improves the quantitative performance on diverse datasets.

Method	Places2			Paris Streetview			CelebA-HQ			DTD		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
GMCNN [32]	22.18	0.849	0.146	25.10	0.856	0.104	26.89	0.931	0.035	27.58	0.932	0.071
Ours (GMCNN)	22.65	0.858	0.133	25.67	0.859	0.097	27.03	0.933	0.030	28.30	0.945	0.057
EdgeConnect [20]	23.61	0.874	0.125	26.05	0.863	0.088	27.24	0.944	0.027	28.35	0.955	0.055
Ours (EdgeConnect)	23.90	0.876	0.117	26.36	0.865	0.084	27.33	**0.947**	0.026	28.97	0.963	0.038
HiFill [36]	24.35	0.867	0.107	26.24	0.866	0.092	27.20	0.936	0.028	29.14	0.950	0.046
Ours (HiFill)	24.52	**0.881**	0.102	26.47	0.866	0.088	27.31	0.940	0.026	**29.38**	0.953	0.039
GatedConv [38]	23.94	0.871	0.112	26.32	0.861	0.090	27.36	0.938	0.028	28.54	0.947	0.052
Ours (GatedConv)	**24.58**	0.880	**0.098**	**26.75**	**0.868**	**0.082**	**27.51**	0.945	**0.025**	29.31	**0.961**	**0.032**

Table 1. Quantitative comparisons on different datasets. The best results are **boldfaced.**
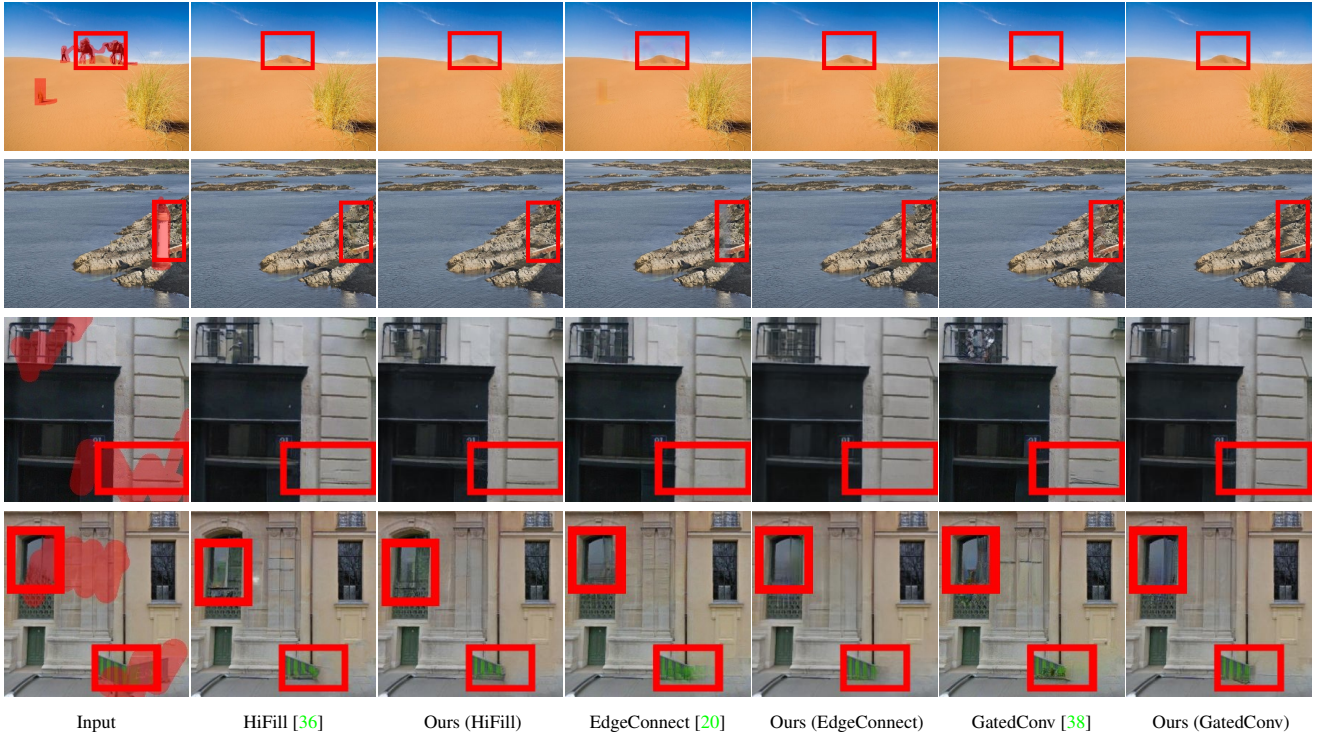

Figure 5. Visual comparisons of different methods. Masked regions are visualized in red. Our method reconstructs coherent structures with fewer color artifacts. Zoom in for details.

Qualitative Comparison As shown in Fig. 5, previous methods can produce semantically-reasonable content with small holes but still show blunt details and abrupt colors. As the hole becomes large, they tend to be unstable. In contrast, we observe that for each baseline network, our method produces compelling results with sharper structures and more consistent colors. This indicates that the proposed approach is not limited to one specific inpainting architecture but can be easily generalized to improve existing inpainting models.

User Study In addition to numerical metrics, we also perform a human perceptual study over the most challenging dataset Places2 on the Amazon Mechanical Turk. Participants are shown a random pair of images (ours and base-line) at once and are asked to select a more realistic image from the two in terms of both color consistency and structure preservation. All images are given at the same resolution in a shuffled order without time limitation. As shown in Fig. 6, models trained with the proposed scheme outperform the corresponding baselines perceptually by a large margin.

4.3. Analysis on Monochromic Bottlenecks

4.3.1 Cross-dataset Analysis

Inpainting models trained on natural datasets usually show huge performance drop on images from other domains (e.g., textures) due to the distribution gaps. Although some frequent patterns (e.g., lines) in texture images are also ubiquitous in natural scenes, the distributions in polychromic
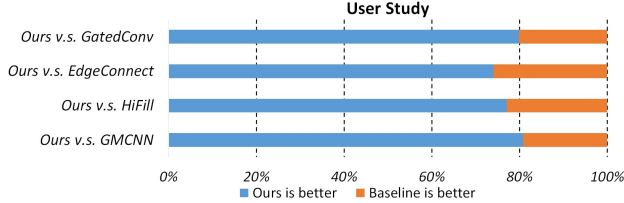
Figure 6. User study results. The reported value indicates the preference rate of Ours ('baseline') against the corresponding baseline.

Train	Test	Baseline	Ours	Train	Test	Baseline	Ours
Places2	DTD	21.85	23.16	DTD	Places2	21.76	22.10
Places2	Pairs	26.24	26.57	DTD	Paris	24.62	24.85
Places2	CelebA-HQ	27.35	27.38	DTD	CelebA-HQ	26.83	26.86

Table 2. Quantitative results of cross-dataset evaluation.

space are still very different since some color patterns seldom appear in natural datasets. While in the monochromic space, this kind of gap is greatly narrowed. We conduct a cross-dataset evaluation to show the generalization ability gain brought by the monochromic bottleneck. As shown in Fig. 7, previous approaches show obvious structure distortion and color discrepancy in missing regions due to the distribution gap, while our model generates sharper lines and more consistent colors with seamless boundary. By learning to reconstruct structures in the monochromic space, the gap between different types of datasets is narrowed. We also find that if the domain gap between two datasets is huge (e.g., CelebA-HQ and others), our model fails to increase the cross-dataset test performance. Otherwise, there is consistent improvement in Table 2.

4.3.2 Effectiveness of Dimension Reduction

As observed in the above experiments, structures and shapes completed by our method are sharper than previous methods. Intuitively, it is easier to learn the reconstruction on monochrome than polychrome because the RGB optimization space $\mathbb{R}^3$ is much larger than the monochromic space of $\mathbb{R}$. To better show the quality gain of reconstruction brought by this dimension reduction, we conduct further analysis on DTD. Since this dataset contains thousands of simple texture images such as line, circle, checkerboard with extremely diverse colors, it is a felicitous example to demonstrate the quality of structure reconstruction.

As shown in Fig. 8, the original baseline model produces curved and blunt details in straight lines. The original model also fails to produce consistent color and seamless boundary when filling in regions with diverse colors. However, the model trained on monochromic space is able to capture the essence of structures and complete correct shapes. It indicates that ignoring color distraction can alleviate the learning complexity and facilitates structure reconstruction.
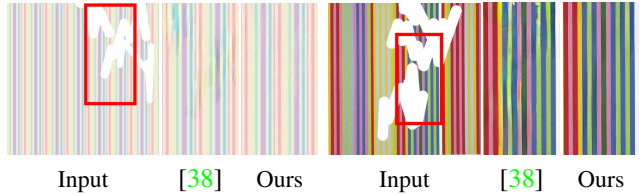

Figure 7. Results of cross-dataset evaluation. We apply the models trained on Places2 to the unseen DTD images.
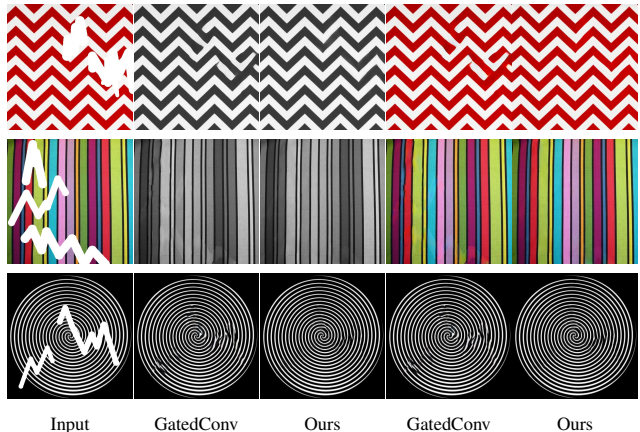

Figure 8. Visual comparisons of reconstruction quality on DTD. We show grayscale images converted from the results of Gated-Conv in the second column for better comparison.

4.4. Analysis on Internal Color Restoration

4.4.1 Ablation Study on Mask Ratios

One key factor that may affect the performance of our internal colorization method is the number of known pixel correspondences. In Fig. 9, we increase the mask ratio of I_{hole} from 22.5% to 73.4% and restore the color of a natural monochrome with our approach. Even in the most challenging case where 73.4% pixels are missing, the model still colorizes I_{hole} in a harmonized style with I_{nhole} without noticeable artifacts. Since in image inpainting, I_{hole} usually accounts for less than 70% of the entire image, and the proposed internal scheme is feasible in most cases.

4.4.2 Ablation Study on Progressive Restoration

We conduct ablation study to figure out how the progressive restoration strategy contributes to the internal colorization. Fig. 10 shows that without the progressive scheme, our model focuses only on local color mappings and generates obvious artifacts and hard boundaries.

4.4.3 Comparison with Other Colorization Methods

As we discussed above, the inpainted monochromes are possibly different from the ground truth due to the ambi-

Figure 9. Feasibility of our internal colorization method. We increase the mask ratio of I_{hole} from 22.5% to 73.4% and colorize the natural monochrome with our method.

Figure 10. Ablation study of the progressive restoration strategy. We colorize the inpainted monochromic bottleneck with our method. Ours (base) is our model without the progressive design.

Mask type	Zhang et al.	Gastal et al.	Levin et al.	Ours (base)	Ours (full)
Rectangular	36.12	29.35	36.68	37.04	**38.45**
Irregular	38.77	39.26	39.24	39.21	**39.50**

Table 3. Results (PSNR) of different guided colorization methods on natural monochromes with different types of masks. Ours (base) is our model without the progressive design.

guity of the inpainting task. It is inappropriate to evaluate the colorization method by comparing colorized inpainted-monochrome with the imperfect ground truth. To avoid the influence of monochrome discrepancy and evaluate the colorization separately, we apply random rectangular and irregular masks to de-colorize ground-truth images from the most challenging dataset Places2. In this way, we simulate the behavior of inpainting masks while having the perfect ground-truth color images for metric calculation. The proposed method achieves a stable performance in Table 3.

4.5. Extensions

In Fig. 11, we show some applications of the proposed method. We demonstrate one user-guided inpainting example, where users can control the color of generated content by giving few color hints interactively. We utilize one extra color point as guidance to inpaint eyes with different colors.

4.6. Failure Cases

Fig. 12 shows failure cases of the proposed inpainting method. In the first example, our method fails to reconstruct a partially-masked bus when the mask is extremely large.

Figure 11. Examples of image editing, extrapolation, and user-guided inpainting. Users can control the style of inpainted content with our approach by giving extra color hints.
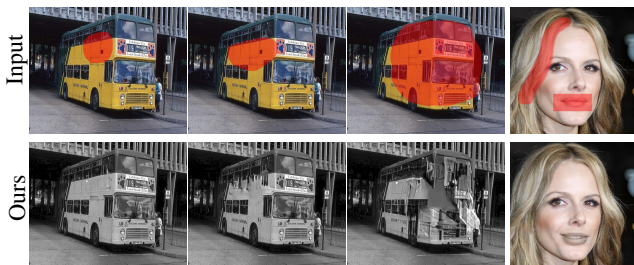

Figure 12. Failure cases. We show failure cases of both reconstruction and colorization.

Similar to previous inpainting approaches, our method has difficulty in completing largely-masked foreground objects, since the structures of these categories are highly complex and diverse. In the second example, our model incorrectly colorizes the mouth due to the lack of colorization hints. In this case, we can consider giving one extra color point as guidance to facilitate the color restoration.

5. Conclusion

In this paper, we propose a general external-internal learning inpainting scheme with monochromic bottlenecks. It first reconstructs the monochrome utilizing semantic knowledge learned externally from large datasets, and then recovers colors internally from a single test image. Our method can produce more coherent structures and more visually harmonized colors compared with previous approaches. Extensive experiments show that our method can lead to stable improvement qualitatively and quantitatively on several backbone models. The major limitation of our method is the inference speed. Since an extra stage is needed for colorization, our method is slower than state-of-the-art approaches. In the future, we plan to accelerate the colorization procedure further and extend the proposed scheme to other low-level vision tasks such as super-resolution.

References

[1] Michael Ashikhmin. Synthesizing natural textures. *SI3D*, 1:217–226, 2001.

[2] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on image processing (TIP)*, 10(8):1200–1211, 2001.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[5] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based photo recoloring. *ACM Transactions on Graphics (TOG)*, 34(4):139–1, 2015.

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 303–312. ACM, 2003.

[9] Selim Esedoglu and Jianhong Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002.

[10] Eduardo SL Gastal and Manuel M Oliveira. Domain transform for edge-aware image and video processing. In *ACM SIGGRAPH 2011 papers*, pages 1–12. 2011.

[11] Mingming He, Jing Liao, Lu Yuan, and Pedro V Sander. Neural color transfer between images. *arXiv preprint arXiv:1710.00756*, 2017.

[12] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Conference on Computer graphics and interactive techniques*, 2001.

[13] Yi-Chin Huang, Yi-Shin Tung, Jun-Cheng Chen, Sung-Wen Wang, and Ja-Ling Wu. An adaptive edge detection based colorization algorithm and its applications. In *Proceedings of ACM international conference on Multimedia (MM)*, 2005.

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH*. 2004.

[17] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[18] Xujie Li, Hanli Zhao, Guizhi Nie, and Hui Huang. Image recoloring using geodesic distance based color harmonization. *Computational Visual Media*, 1(2):143–155, 2015.

[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[20] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. In *Workshop on the IEEE International Conference on Computer Vision (ICCVW)*, 2019.

[21] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM Transactions on Graphics (TOG)*, 25(3):1214–1220, 2006.

[23] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[24] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[26] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang. Fast image/video upsampling. *ACM Transactions on Graphics (TOG)*, 27(5):1–7, 2008.

[27] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and remapping the "dna" of a natural image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[28] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. In *British Machine Vision Conference (BMVC)*, 2018.

[30] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on image processing (TIP)*, 13(4):600–612, 2004.

[34] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE Transactions on image processing (TIP)*, 15(5):1120–1129, 2006.

[36] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7508–7517, 2020.

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[39] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision (ECCV)*, pages 1–17. Springer, 2020.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017.

[42] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[43] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *ACM Transactions on Graphics (ToG)*, 37(4):49:1–49:13, 2018.

Appendix

A. Implementation Details

After experimenting with different network architectures (U-net, ResNet, SkipNet) for our progressive internal colorization network, we adopted a ResNet-like architecture since it is the fastest among all the models that generate high-quality results. The number of feature channels is set to 32, and the filter size is 3. The network is composed of 9 two-layer residual blocks, 1 input Conv layer, and 1 output Conv layer. The input channel number of the first generator is 1 (gray only), and for other generators is 4 (gray+RGB). The output channel for each generator is 3 (RGB). Note that all our experiments are conducted in the RGB color-space instead of Lab color-space, so the luminance of the input image can possibly change slightly. BatchNorm, reflection padding, and LeakyReLU are adopted.

We choose the pyramid height for our progressive color propagation based on the image size and image content. The default pyramid height is set to 3 for Places2 images. The corresponding iteration number and learning rate at each level is [500,1000,1000] and [0.01, 0.005, 0.003]. We empirically found that with a larger learning rate at a lower level, the output contains more diverse colors, and with a lower learning rate at a higher level, the output contains more fine-grained details. In case the default configuration dose not yield a compelling colorization results, we can tune the pyramid height and learning rates for better performance.

B. Feasibility of Internal Colorization

We also provide more examples to validate the feasibility of the proposed internal propagation. These examples are from different categories, including natural scenery, buildings, human faces, and animals. As shown in Fig. 13 and Fig. 14, our scheme achieves stable colorization results on various images with different mask ratios.

C. User-guided Inpainting

Users can control the color of inpainted content with our inpainting method. We provide more details of user-guided inpainting and comparison results in Fig. 15. In this example, our model utilizes only one extra user-guided color point as a hint to generate realistic eyes with different colors. In contrast, other guided colorization methods fail to produce visually pleasing results and show obvious color bleeding artifacts in the eyes.

D. More Results
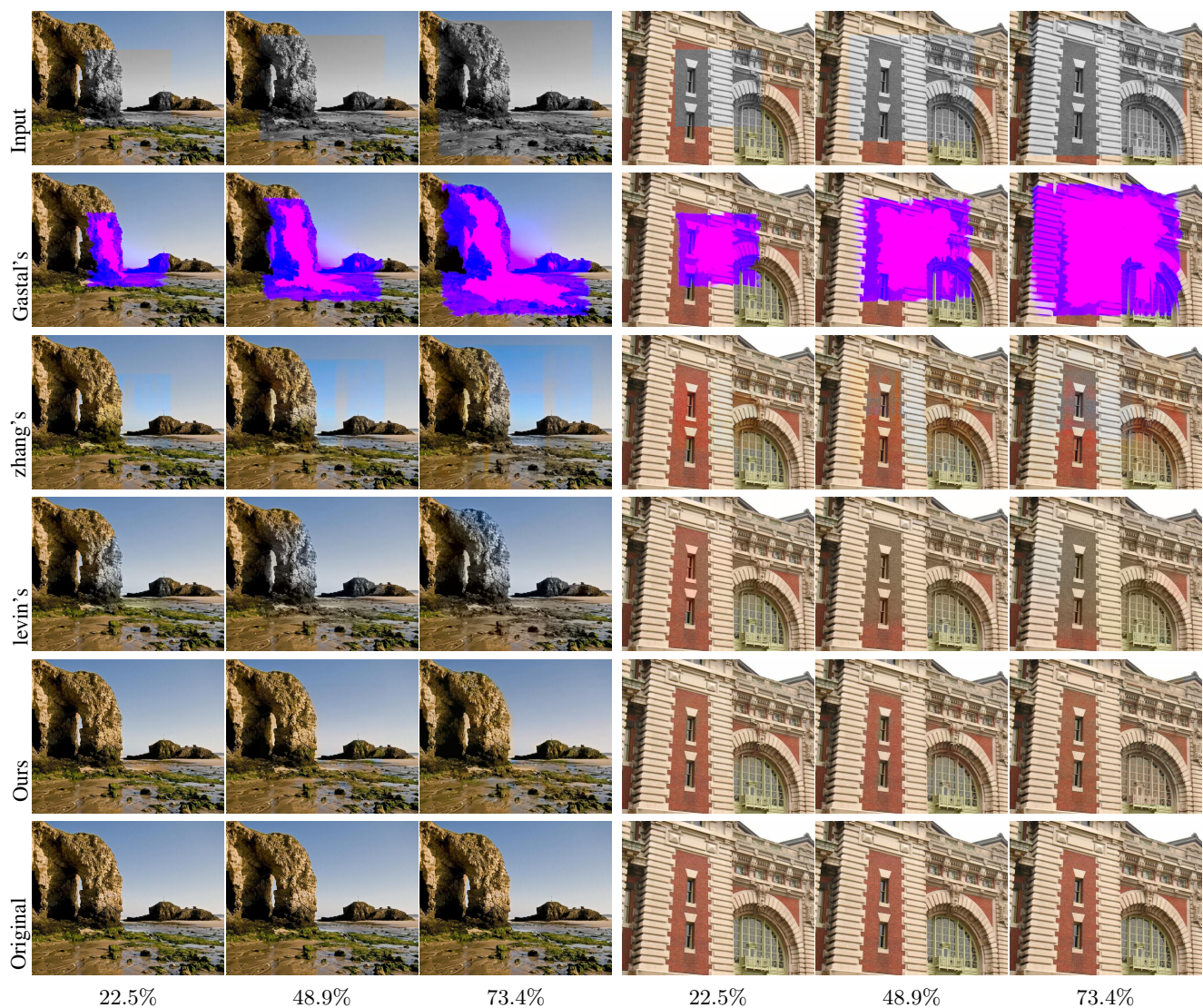
More visual results are shown in Fig. 16, Fig. 17.

Figure 13. Feasibility of our internal colorization method. We increase the mask ratio of I_{hole} from 22.5% to 73.4% and colorize the ground-truth grayscale image with our internal colorization method. We also give colorization results by Zhang et al. [41], Gastal et al. [10] and Levin et al. [16] for comparison.
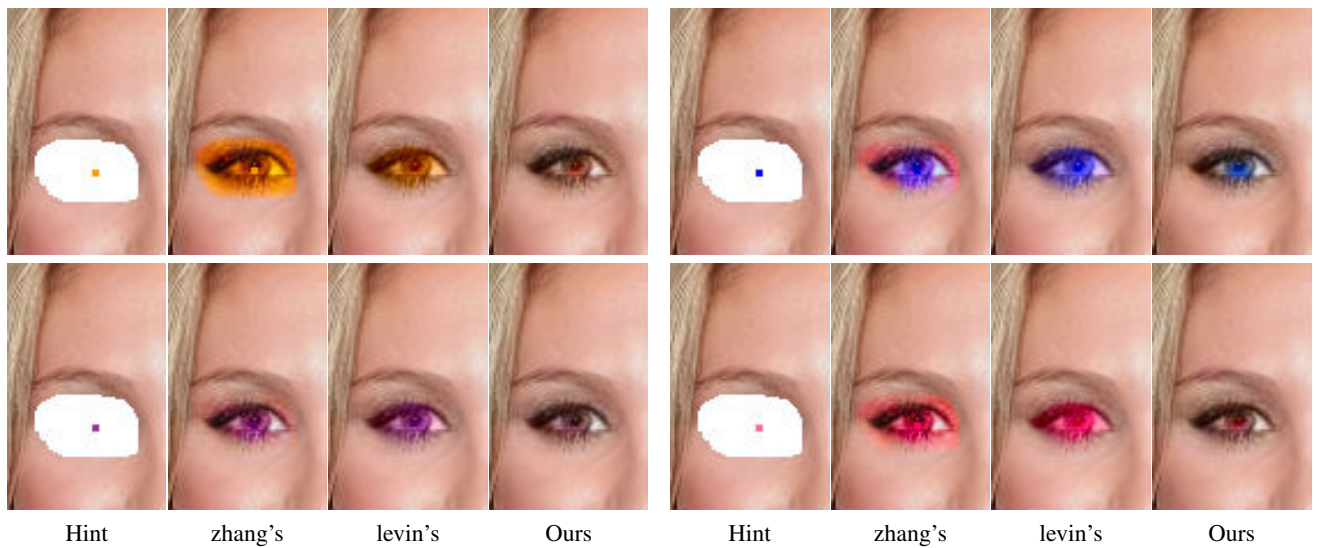
Figure 14. Feasibility of our internal colorization method. We increase the mask ratio of I_{hole} from 22.5% to 73.4% and colorize the ground-truth grayscale image with our internal colorization method. We also give colorization results by Zhang et al. [41], Gastal et al. [10] and Levin et al. [16] for comparison.

| Hint | zhang's | levin's | Ours | Hint | zhang's | levin's | Ours |

Figure 15. Examples of user-guided inpainting by our method. Users can control the color of inpainted content.
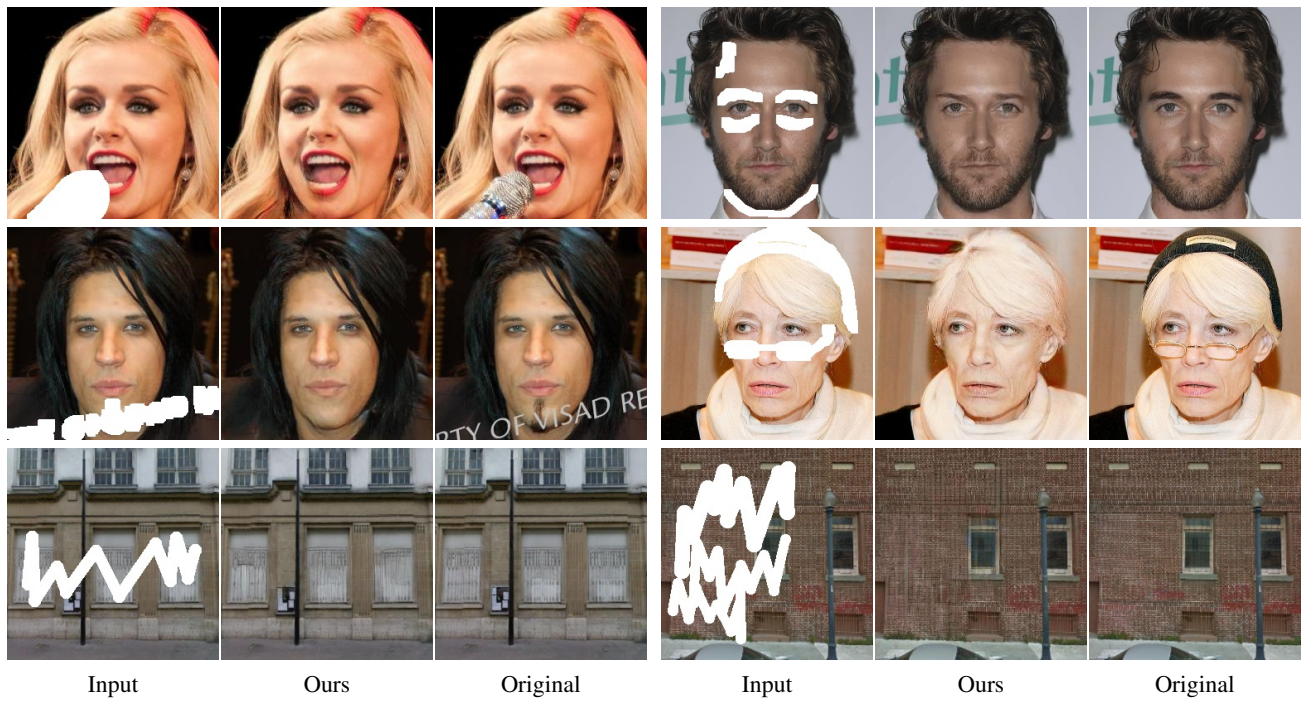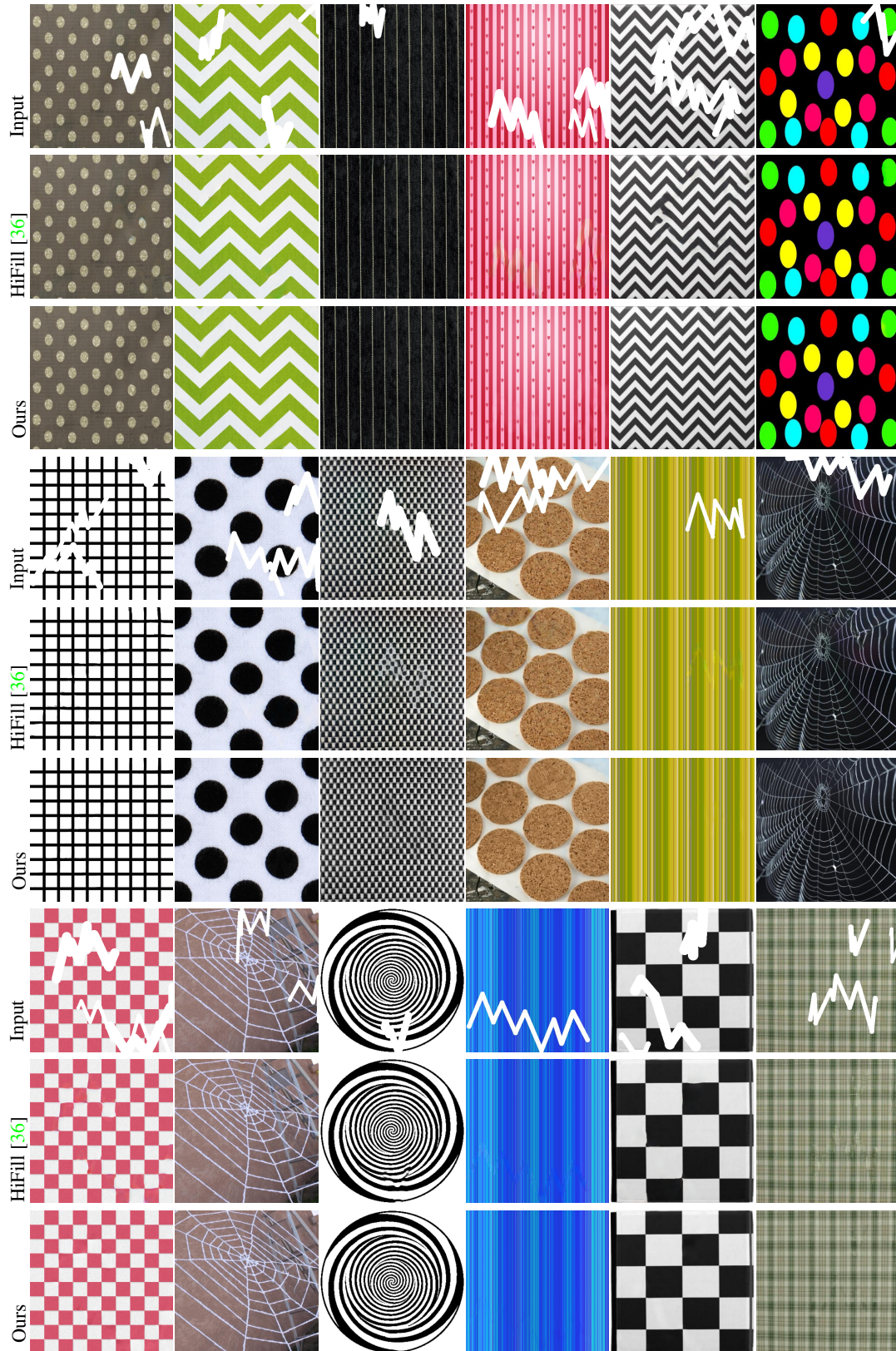

| Input | Ours | Original | Input | Ours | Original |

Figure 16. More results.

Figure 17. Sampled results on DTD.