

# LGA-Net: Learning Local and Global Affinities for Sparse Scribble based Image Colorization (Supplementary Document)

Hongjin Lyu<sup>1</sup>, Bo Li<sup>2\*</sup>, Paul L. Rosin<sup>1</sup>, Yu-Kun Lai<sup>1</sup>

<sup>1</sup> School of Computer Science and Informatics, Cardiff University

<sup>2</sup> School of Mathematics and Information Science, Nanchang Hangkong University

{lyuh2, RosinPL, LaiY4}@cardiff.ac.uk, libo@nchu.edu.cn

We provide extensive experimental results in this supplementary document:

- To present more clearly the performance of LGA-Net and different methods [1, 2, 4, 5] in colorization tasks when using sparse scribbles, this supplementary material additionally shows the colorized results of 60 randomly selected examples displayed in Fig. 1 - Fig. 6 (using both manually drawn and automatically generated sparse scribbles) in Section 1.
- To demonstrate the crucial role of local and global affinities in LGA-Net, we further provide more examples of the ablation study regarding local and global affinity relationships. This includes 20 randomly selected examples displayed in Fig. 7–Fig. 8 in Section 2.1. Furthermore, three ablation experiments involving different loss terms have been quantitatively analyzed in Section 2.2, which further demonstrates the importance of each loss term.
- To verify the robustness of LGA-Net on real and diverse scribble inputs, we compare LGA-Net with different methods [1, 2, 4, 5] under 2 examples, each with 5 different user scribbles (covering difference in density, position, and color selection) in Fig. 9 in Section 3, further illustrating its effectiveness across real application scenarios.
- Section 4 further presents 8 examples shown in published papers (Zhang [5], UniColor [1] and iColoriT [4]) to compare the performance of LGA-Net with other methods, which further demonstrate that LGA-Net can stably generate high-quality colorization result.
- A new automatic scribble generation algorithm described in Section 5 is proposed to simulate user-provided scribbles, which substantially reduces the effort of creating scribble-based test datasets.

## 1. Qualitative Comparison with Other Methods: Expanding Examples and Scenarios

LGA-Net formulates the colorization task as an affinity propagation process, and achieves state-of-the-art performance in colorization when applying sparse scribbles by explicitly learning two different levels of affinities (local and global). Benefiting from the novel and powerful affinity learning and propagation mechanism mentioned above, LGA-Net trained on the training dataset  $\mathcal{D}_t$  containing only 4K images outperforms other methods [1, 4, 5] trained on the ImageNet dataset with 1.3M images.

In this supplementary document, we present extensive experimental results, showing 60 randomly selected test images and comparing four different methods, allowing for a more comprehensive and in-depth understanding of the performance of different methods in various scenarios.

Out of these test cases, 30 test images in Fig. 1 - Fig. 3 are from  $\mathcal{D}_{manual}$ , manually created by authors, while the other 30 test images in Fig. 4 - Fig. 6 are from  $\mathcal{D}_{auto}$ , automatically generated by AutoSS described in Section 5.

The four methods are Levin et al. [2], Zhang et al. [5], UniColor [1] and iColoriT [4]. Among them, Zhang [5] and iColoriT [4] trained under  $\mathcal{D}_t$  are abbreviated as Z4K and iC4K. The official pre-trained models on ImageNet of Zhang [5], UniColor [1], and iColoriT [4] are abbreviated as Z1M, UC1M, and iC1M respectively. Due to the lack of official training instructions, we do not conduct UniColor trained under  $\mathcal{D}_t$ . Thus, the results generated by seven pre-trained models are ultimately showcased: Levin, Z4K, iC4K, Z1M, iC1M, UC1M, and LGA-Net.

Key conclusions are shown below:

- Levin struggles to cope with sparse scribbles inputs due to its lack of capability to learn global affinities.
- Despite their well-designed network architectures, Zhang [5] and iColoriT [4], which heavily rely on large training datasets, produce obvious artifacts when trained

---

\*Corresponding author

on limited datasets (Z4K and iC4k).

- Benefiting from the extensive training data of ImageNet, Z1M, iC1M, and UC1M have acquired better capabilities for scribble-based colorization.
- Without exploiting affinities explicitly as done in our work, Z1M struggles to learn a sufficiently generalizable model, leading to instability when faced with challenging inputs, even with the help of 1.3M ImageNet dataset.
- The unique network design of iC1M enables certain local and global affinities to be captured, but also results in significant artifacts generation when provided with scribble hints.
- Due to the lack of an explicit mechanism for learning global affinities, UC1M still struggles to stably accomplish long-range color propagation, despite leveraging Transformer and VQGAN for higher quality affinities.

## 2. Ablation Study

### 2.1. Ablation Study involving More Examples and Varied Scenarios

LGA-Net relies on learned accurate local and global affinities to effectively accomplish color propagation tasks at both short and long distances, which respectively originate from the introduction of adjacent points and global points.

To better demonstrate the significant role of local and global affinities, this supplementary material further presents the results of 20 randomly selected examples under  $R_{GP}$  (removing global points/affinities),  $R_{AP}$  (removing local points/affinities), and Full LGA-Net. Fig. 7 displays 10 random examples from  $\mathcal{D}_{manual}$ , while Fig. 8 exhibits the remaining 10 random examples from  $\mathcal{D}_{auto}$ . This supplementary material does not continue to present the results of the ablation experiment regarding removing Non-Local blocks ( $R_{NLB}$ ), as the Non-Local block is not the primary contribution of LGA-Net.

Although  $R_{GP}$  can reasonably diffuse the colors provided by users' scribbles in local areas, it fails to accurately transmit colors to distant regions with similar textures. This shortcoming arises from the removal of global affinities in  $R_{GP}$ , which prevents it from effectively learning the crucial global affinities necessary for long-range color propagation.

$R_{AP}$ , which retains only global affinities, exhibits some degree of short-range and long-range color propagation but falls short of stably generating high-quality colorization results. This limitation stems from the inability to accurately learn precise local and global affinities relying solely on global relationships.

During the training process, LGA-Net jointly learns local and global affinities based on adjacent and global points, which can enable a better understanding of image structure, spatial layouts and relationships among different objects in the image. Based on the aforementioned mecha-

nisms, LGA-Net facilitates more accurate affinities, thereby achieving more precise localization and more proper color propagation.

Table 1. Quantitative Analysis on three separate losses

$\mathcal{D}_{manual}$ 200 cases	MSE↓ $\times 10^{-2}$	PSNR↑ dB	MSSSIM↑ $\times 10^{-2}$	LPIPS↓ $\times 10^{-2}$
$R_{\mathcal{L}_1}$	1.72	25.34	91.94	12.23
$R_{\mathcal{L}_{lap}}$	1.50	25.88	92.63	11.07
$R_{\mathcal{L}_{TV}}$	5.57	20.20	88.27	20.75
LGA-Net	<b>1.42</b>	<b>26.07</b>	<b>92.78</b>	<b>10.49</b>
$\mathcal{D}_{auto}$ 3000 cases	MSE↓ $\times 10^{-2}$	PSNR↑ dB	MSSSIM↑ $\times 10^{-2}$	LPIPS↓ $\times 10^{-2}$
$R_{\mathcal{L}_1}$	1.10	29.39	95.45	10.53
$R_{\mathcal{L}_{lap}}$	0.98	29.58	95.55	10.10
$R_{\mathcal{L}_{TV}}$	1.51	27.97	94.76	12.51
LGA-Net	<b>0.90</b>	<b>29.89</b>	<b>95.73</b>	<b>9.62</b>

### 2.2. Ablation study involving Loss Terms

In this section, we further conducted ablation experiments on three different loss terms: removing  $\mathcal{L}_1$  abbreviated as  $R_{\mathcal{L}_1}$ , removing  $\mathcal{L}_{lap}$  abbreviated as  $R_{\mathcal{L}_{lap}}$ , and removing  $\mathcal{L}_{TV}$  abbreviated as  $R_{\mathcal{L}_{TV}}$ .

$R_{\mathcal{L}_1}$  fails to accurately learn the distribution of pixel values due to the lack of pixel-level  $\mathcal{L}_1$  loss, resulting in reconstructed images with relatively large pixel value deviations.  $R_{\mathcal{L}_{lap}}$  approximates full LGA-Net but underperforms in all tests since it is unable to accurately guide LGA-Net at multiple different scales, thus losing the effective learning ability for detailed information. Removing  $\mathcal{L}_{TV}$  in  $R_{\mathcal{L}_{TV}}$  makes LGA-Net unable to effectively remove noise which consequently leads to the worst performance. Full LGA-Net consistently excels in all evaluation scenarios (bold), underscoring the significance of each loss term.

## 3. Usability Study: Evaluating Method Robustness with Real Diverse Scribble Inputs

To further validate the effectiveness and robustness of LGA-Net, we conducted a usability study to evaluate how different colorization approaches perform with real and diverse scribble inputs for the same grayscale image. For identical grayscale images, color scribbles provided by different users vary in many aspects, such as density, position, and color selection. Comparing the colorization results of different methods under these diverse input conditions enables a more convincing assessment of whether our method can stably generate high-quality results across real-world user input patterns, further demonstrating its effectiveness.

As shown in the third column of Fig. 9, Levin et al. [2]'s method fails to handle real sparse scribble inputs due to its lack of long-range color propagation capability.

Results in the fourth and fifth columns of Fig. 9 demonstrate that both Z4K [5] and iC4K [4] cannot stably generate



satisfactory colorization outputs, as they lack support from a large-scale training dataset.

The sixth column of Fig. 9 reveals that Z1M [5], trained on 1.3 million samples, acquires a certain degree of long-range color propagation ability. However, Z1M still struggles to stably perform this function under real sparse scribble inputs (see 3rd, 4th, 5th, 6th, and 8th rows).

The seventh column of Fig. 9 shows that although long-range color propagation is achieved via the self-attention mechanism, the pixel shuffling operation in iC1M [4] added to accelerate data processing introduces obvious artifacts, degrading result quality.

In the eighth column of Fig. 9, UC1M [1], benefiting from the integration of Transformer and VQGAN, effectively handles most real scribble inputs. Nevertheless, its performance is limited under diverse real scribble inputs due to the absence of explicit long-range affinity learning: for instance, it fails to accurately perform long-range color propagation in row 5, and in row 6, it colors the right wall (without color scribbles) based on prior knowledge rather than user-provided information.

Results in the ninth column of Fig. 9 validate LGA-Net’s performance: it stably transfers colors from user-provided scribbles to reasonable regions and effectively adapts to diverse sparse scribble inputs in real scenarios, further confirming the strong robustness of our proposed method.

#### 4. Qualitative Analysis involving Examples from Published Papers

To further compare the performance of LGA-Net with other methods, we selected 8 examples presented in Fig. 10, which are taken from latest published papers (Zhang [5], UniColor [1] and iColoriT [4]). In the first four rows, each case is provided with a small number of point-based hints, representing typical sparse hint scenarios. In contrast, the last four rows illustrate results under dense hint conditions, where more than 50 point-based hints are provided. This division between sparse and dense hint distributions helps readers more intuitively compare the performance of different methods.

Under sparse hint conditions, Levin, Z4K, and iC4K struggle to generate high-quality colorization results consistently. This is primarily due to limitations in their inherent mechanisms: Levin relies solely on pixel intensity, making it incapable of capturing global affinity relationships, while Z4K and iC4K are heavily dependent on the generalization capabilities of large-scale training datasets, lacking robustness when faced with sparse hints. In contrast, Z1M, iC1M, and UC1M benefit from the strong generalization capabilities of large-scale training datasets (ImageNet), which enable them to handle sparse hints to a certain extent.

LGA-Net can properly achieve colorization task when facing sparse hints input as shown in the 9th column in

Fig 10. The reason behind is that LGA-Net innovatively formulates the scribble-based colorization task as a color propagation problem grounded in affinity relationships. By decomposing this task into two explicit processes, image structure understanding using CNNs and color propagation based on both local and global affinity relationships, LGA-Net achieves precise and faithful propagation of user-provided hints to the appropriate regions, which ensures that LGA-Net can stably generate high-quality colorized results and strictly adheres to user intentions when facing sparse hints input.

The 5th row to 8th row of Fig 10 further illustrate the performance of each method under dense hint inputs. Levin, Z4K, and iC4K still fail to achieve stable and reliable high-quality colorization results. In contrast, LGA-Net, trained on a small dataset containing only 4K cases, stably achieves colorization results comparable to Z1M, iC1M and UC1M (official pre-trained models on the 1.3M-case ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]), which further highlights the innovation and effectiveness of LGA-Net. Moreover, LGA-Net consistently produces high-quality colorization results when applied to real-world scenarios involving scribble-based hints. The above observations further validate that LGA-Net can accurately achieve scribble-based colorization task and faithfully propagate user-provided color hints to the proper regions.

#### 5. Automated Scribble Generation (AutoSS)

---

##### Algorithm 1 Automatic Sparse Scribble Algorithm

---

**Require:** Input =  $\mathcal{J}_i$

- 1:  $\text{seg}_{\text{result}} = \text{HillClimbingSegment}(\mathcal{J}_i, \text{iBinNum});$   
iBinNum controls the color sensitivity degree;
  - 2:  $\text{seg}_{\text{result}} = \text{filtering}(\text{seg}_{\text{result}}, [\text{W}_s, \text{W}_s]);$   $\text{W}_s$  controls the considering area;
  - 3:  $\text{ske}_{\text{result}} = \text{Skeletonizing}(\text{logical\_mask}(\text{seg}_{\text{result}}));$
  - 4:  $\text{ske}_{\text{result}} = \text{Clean\_process}(\text{ske}_{\text{result}});$
  - 5: **for** each scribble **do**
  - 6:   Pick the largest scribbles from the top B sub-regions;
  - 7: **end for**
  - 8:  $\text{K} = \text{Similar}(\text{B});$  calculating the K most similar scribbles pairs;
  - 9: Randomly remove R scribbles from K to simulate different impatience degree;
  - 10: Output final sparse scribble mask;
- 

In order to ease the practical challenge of creating scribble-based datasets, this paper proposes a heuristic method for automatically generating scribbles of color images, called AutoSS. Given a color image  $\mathcal{J}_i$ , we

firstly get the segmentation result  $\text{seg}_{\text{result}}$  based on *HillClimbingSegment* [3] and *filtering* which performs mode filtering for the most frequently occurring value within a neighborhood around each pixel.  $W_s$  controls the considering area. Then a logical mask is generated by *logical\_mask*, facilitating the following skeletonization process by *Skeletonizing* in Line 3; detailed cleaning operations are then conducted by *Clean\_process*, which contains eliminating too short lines, thickening skeletons, etc. The biggest scribbles of the top B sub-regions are selected as the preliminary baseline. In Lines 8 and 9, AutoSS calculates the most similar scribble pairs according to the *Similar()* function, and randomly deletes the top R scribbles to simulate users being impatient.

In AutoSS, to mimic varying user patience levels realistically, R selects a value randomly from  $\{0, 1, 2, 3, 4, 5, 6, 7\}$ , with higher values indicating greater impatience. iBinNum as the input parameter of *HillClimbingSegment*, is set to 20, which controls the color sensitivity.  $W_s$  is set to 11, which controls the size of the filtering window. The *Similar()* function determines similarity by averaging the difference of the RGB channels. We commit to releasing the source code of this research upon acceptance of the paper.

## References

- [1] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. [1](#), [3](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [13](#), [14](#)
- [2] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH*, pages 689–694. 2004. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [13](#), [14](#)
- [3] Yiquan. Hill-climbing color image segmentation. <https://www.mathworks.com/matlabcentral/fileexchange/22274-hill-climbing-color-image-segmentation>. Accessed: 2010-09-30. [4](#)
- [4] Jooyeol Yun, Sanghyeon Lee, Minho Park, and Jaegul Choo. iColoriT: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1787–1796, 2023. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [13](#), [14](#)
- [5] 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. *arXiv preprint arXiv:1705.02999*, 2017. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [13](#), [14](#)



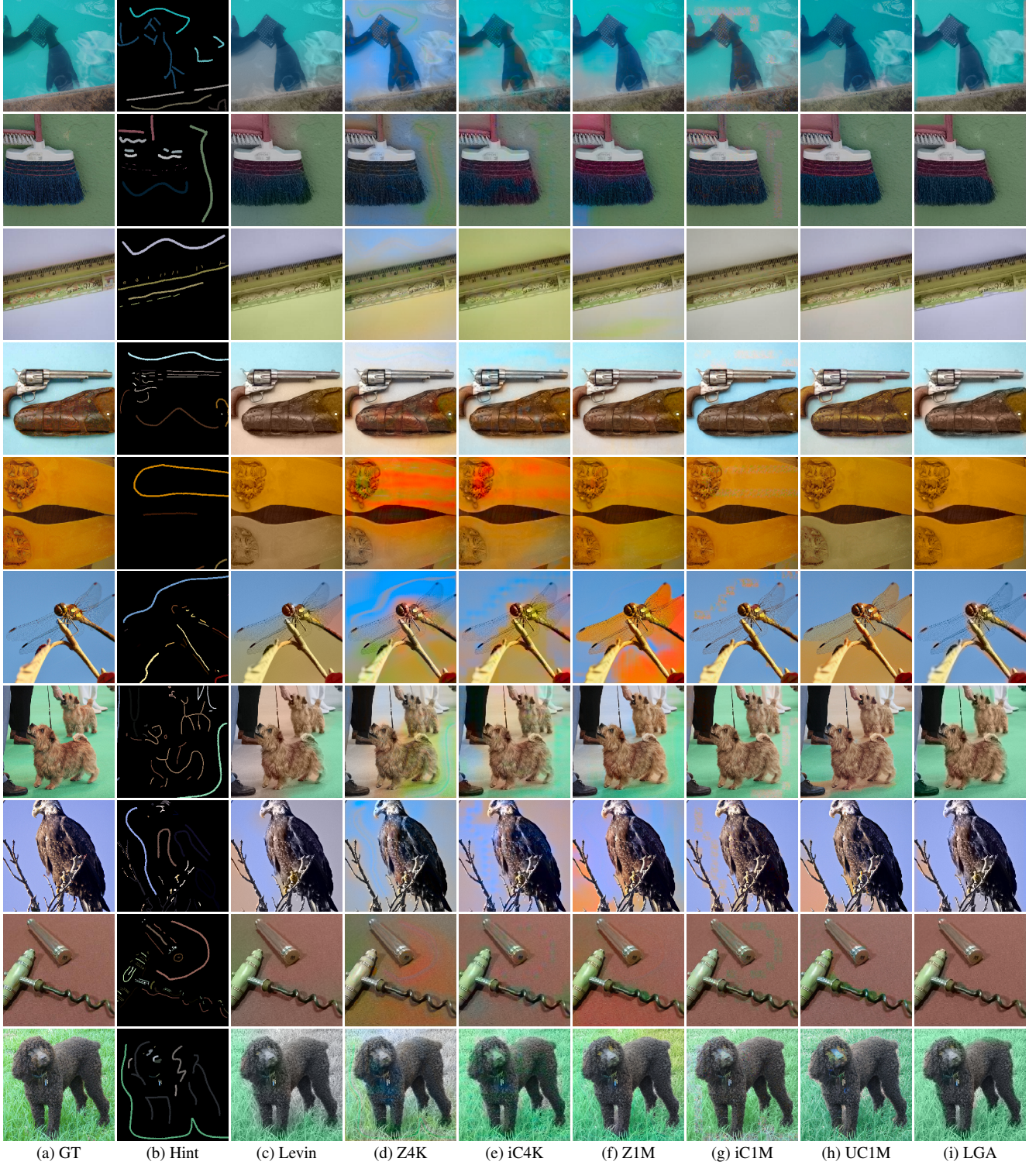


Figure 1. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{manual}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_t$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.





Figure 2. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{manual}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_t$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



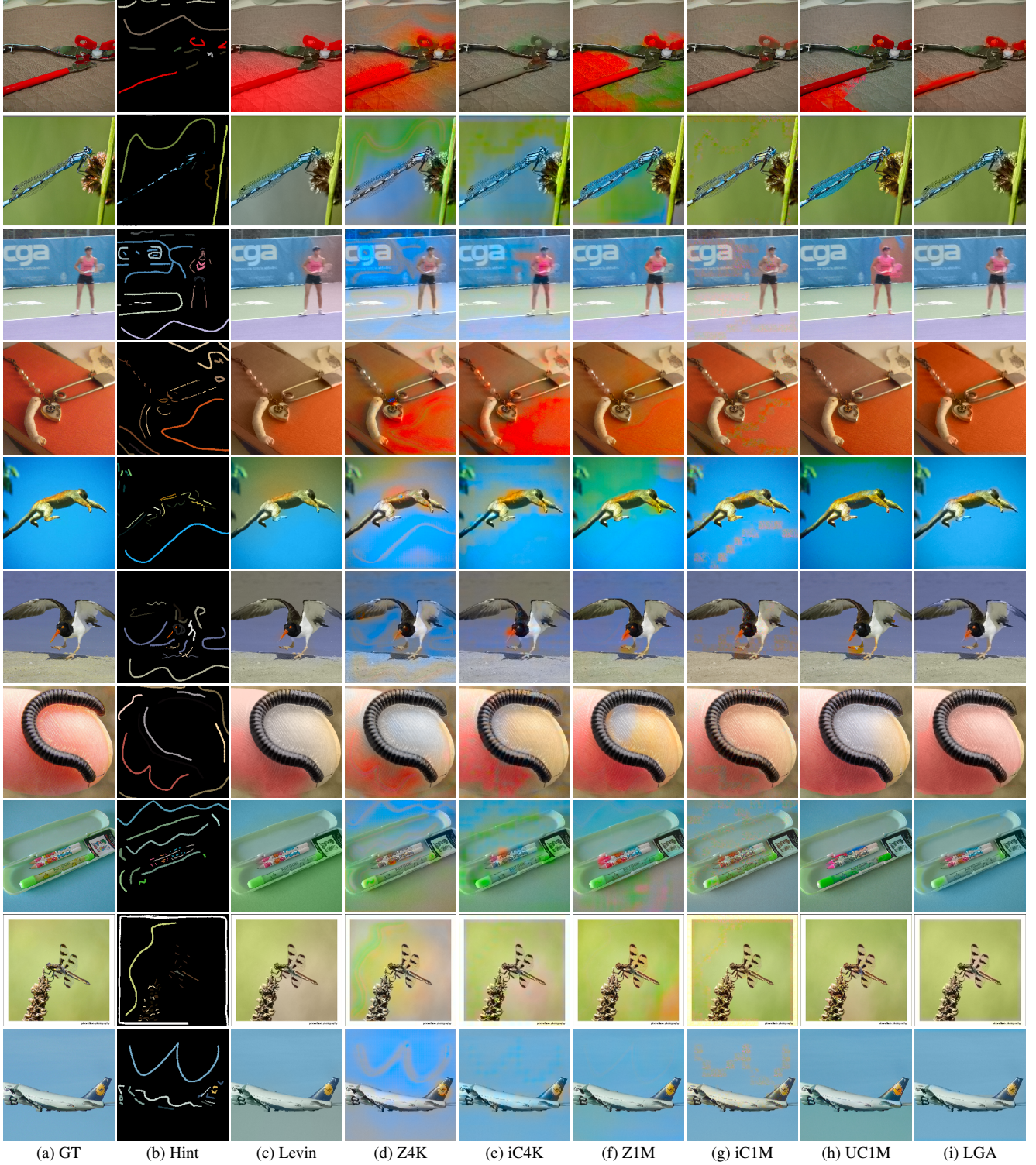


Figure 3. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{manual}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_t$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



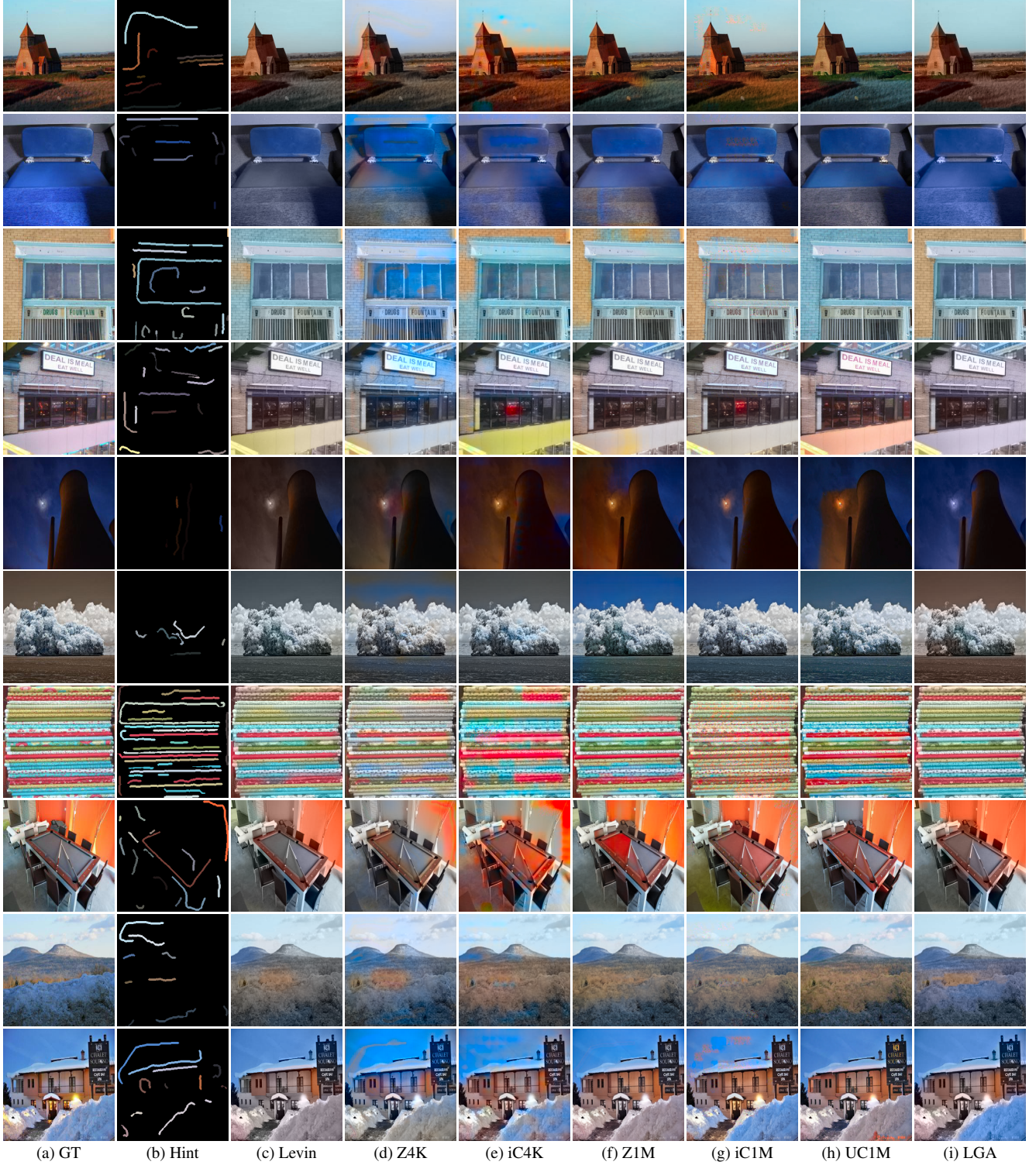


Figure 4. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{auto}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_i$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



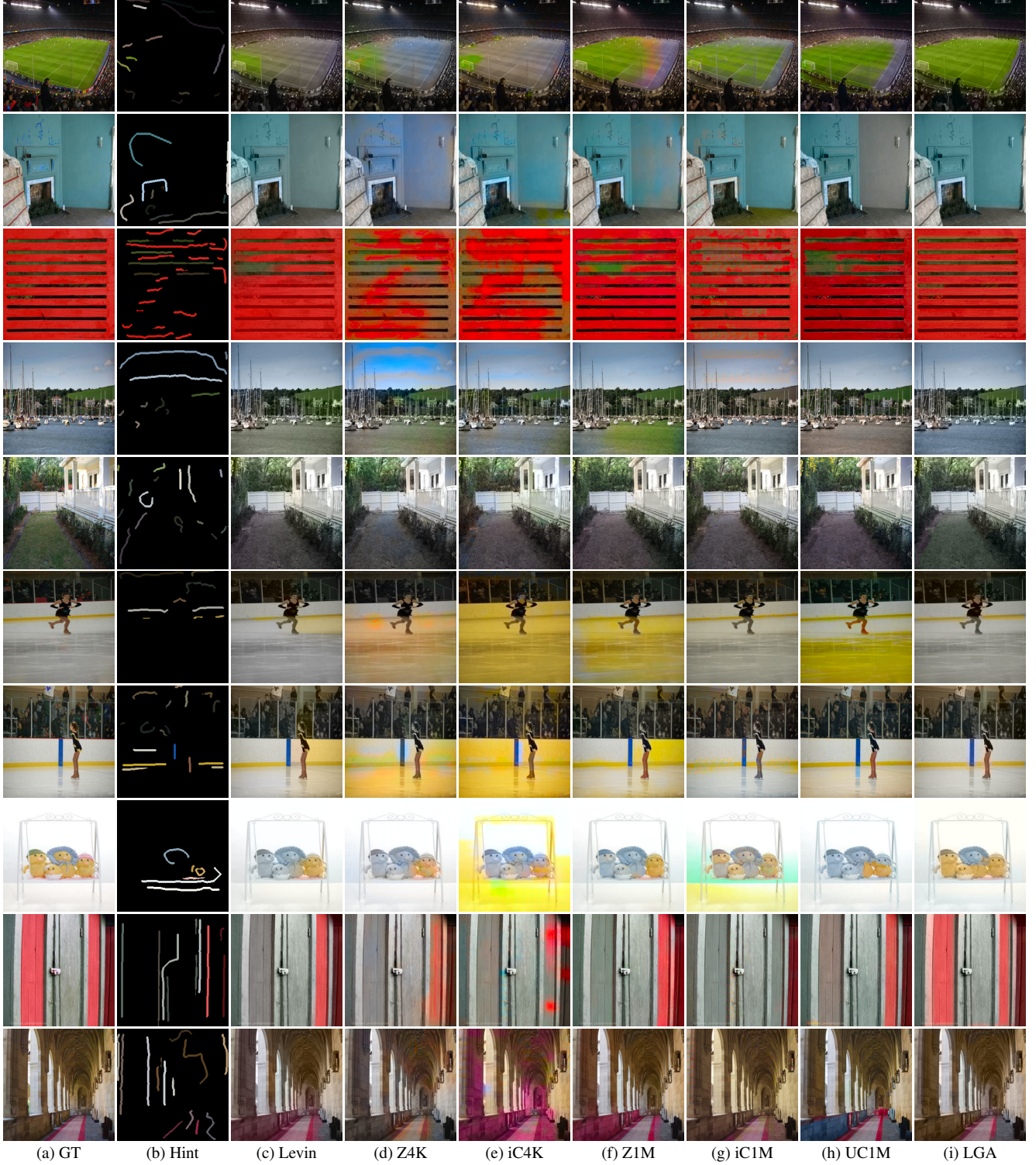


Figure 5. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{auto}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_i$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



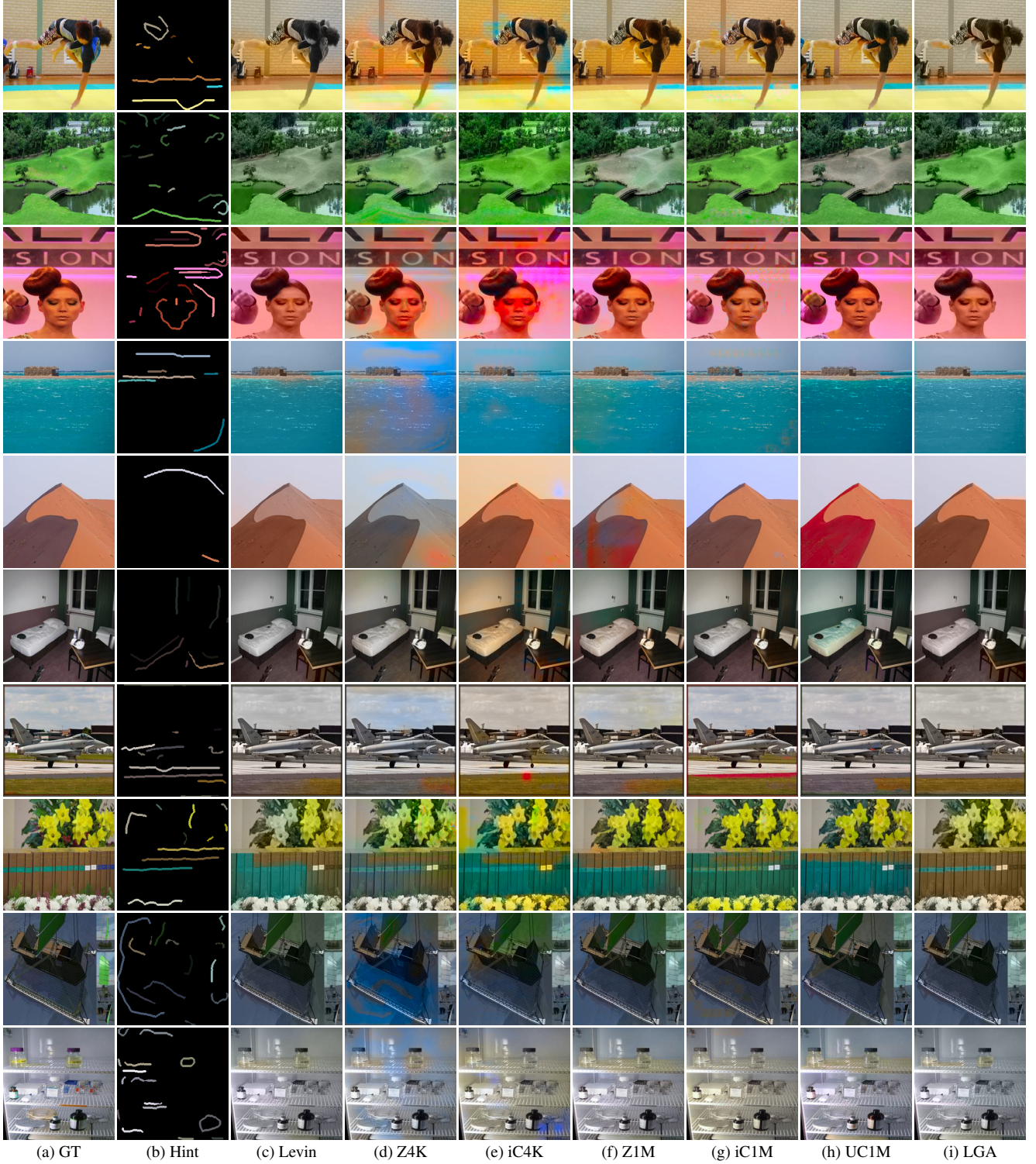


Figure 6. Qualitative comparison with state-of-the-art methods in Section 1. This figure shows 10 randomly selected examples from  $\mathcal{D}_{auto}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_i$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



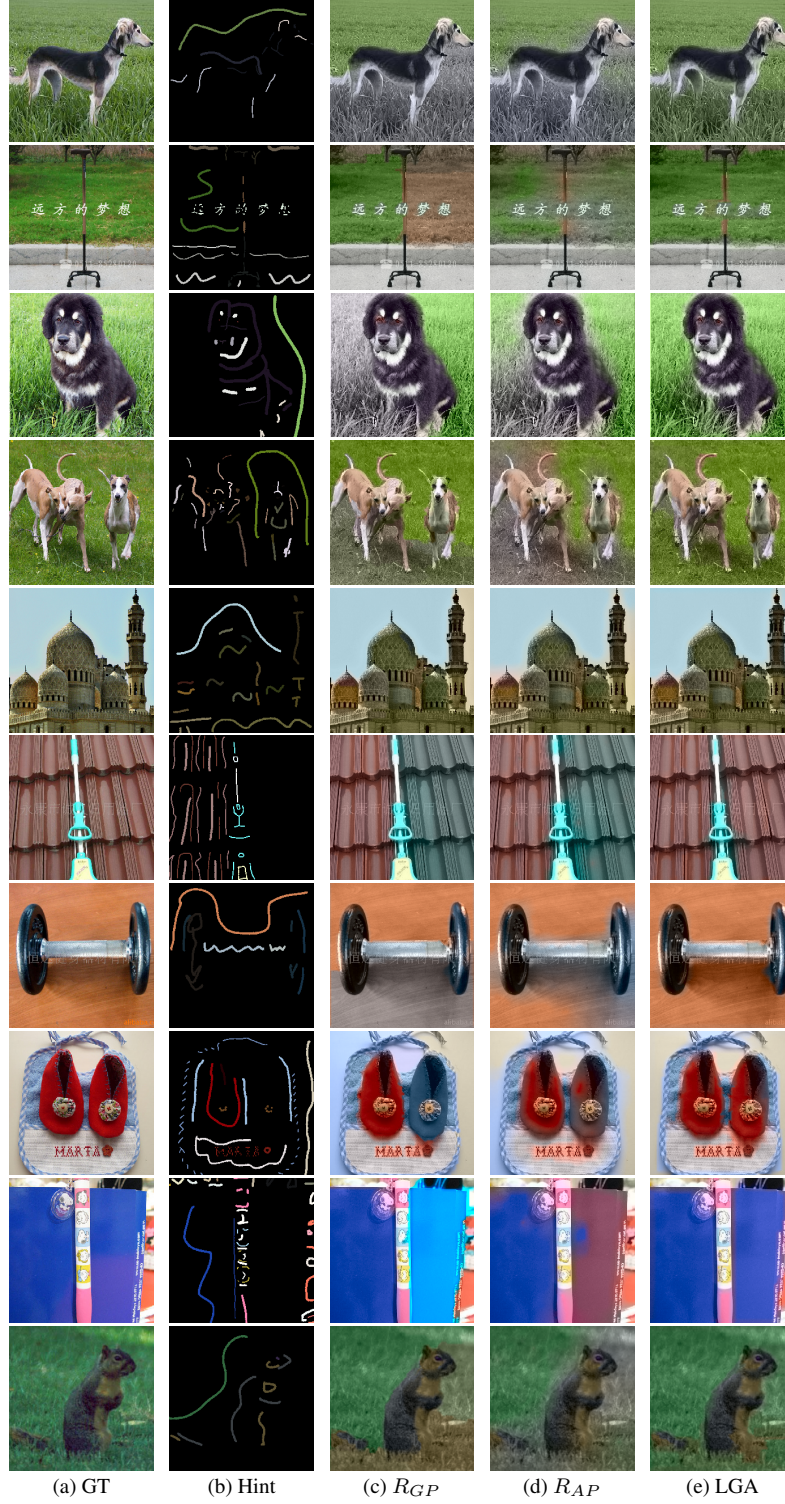


Figure 7. Qualitative comparison of Ablation Study in Section 2. This figure shows 10 randomly selected examples from  $\mathcal{D}_{manual}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display:  $R_{GP}$  removing global points/affinities;  $R_{AP}$  removing adjacent points/affinities; full LGA-Net.

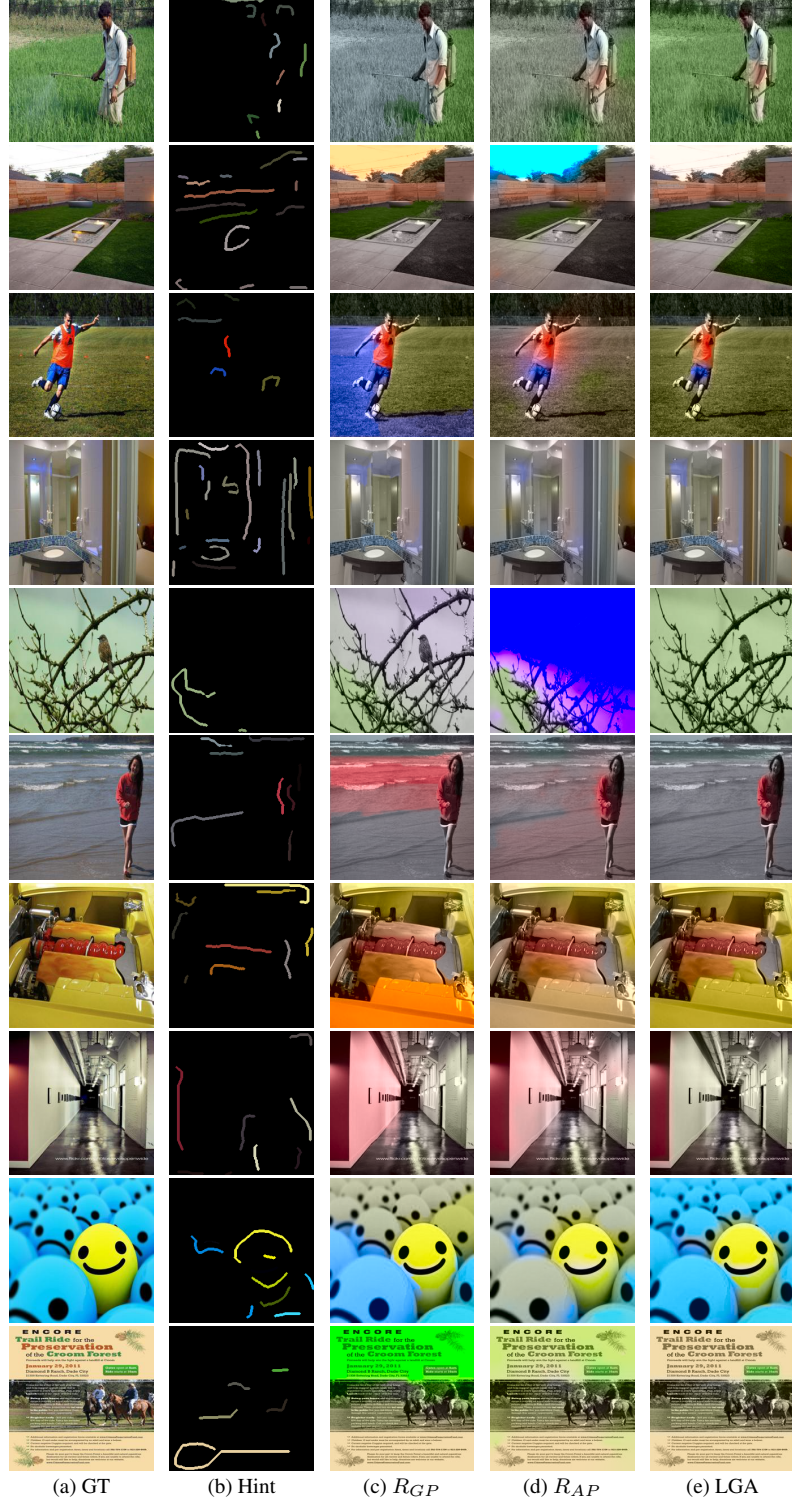


Figure 8. Qualitative comparison of Ablation Study in Section 2. This figure shows 10 randomly selected examples from  $\mathcal{D}_{auto}$ . The first two columns show the ground truth and the applied scribbles, while the following columns separately display:  $R_{GP}$  removing global points/affinities;  $R_{AP}$  removing adjacent points/affinities; full LGA-Net.



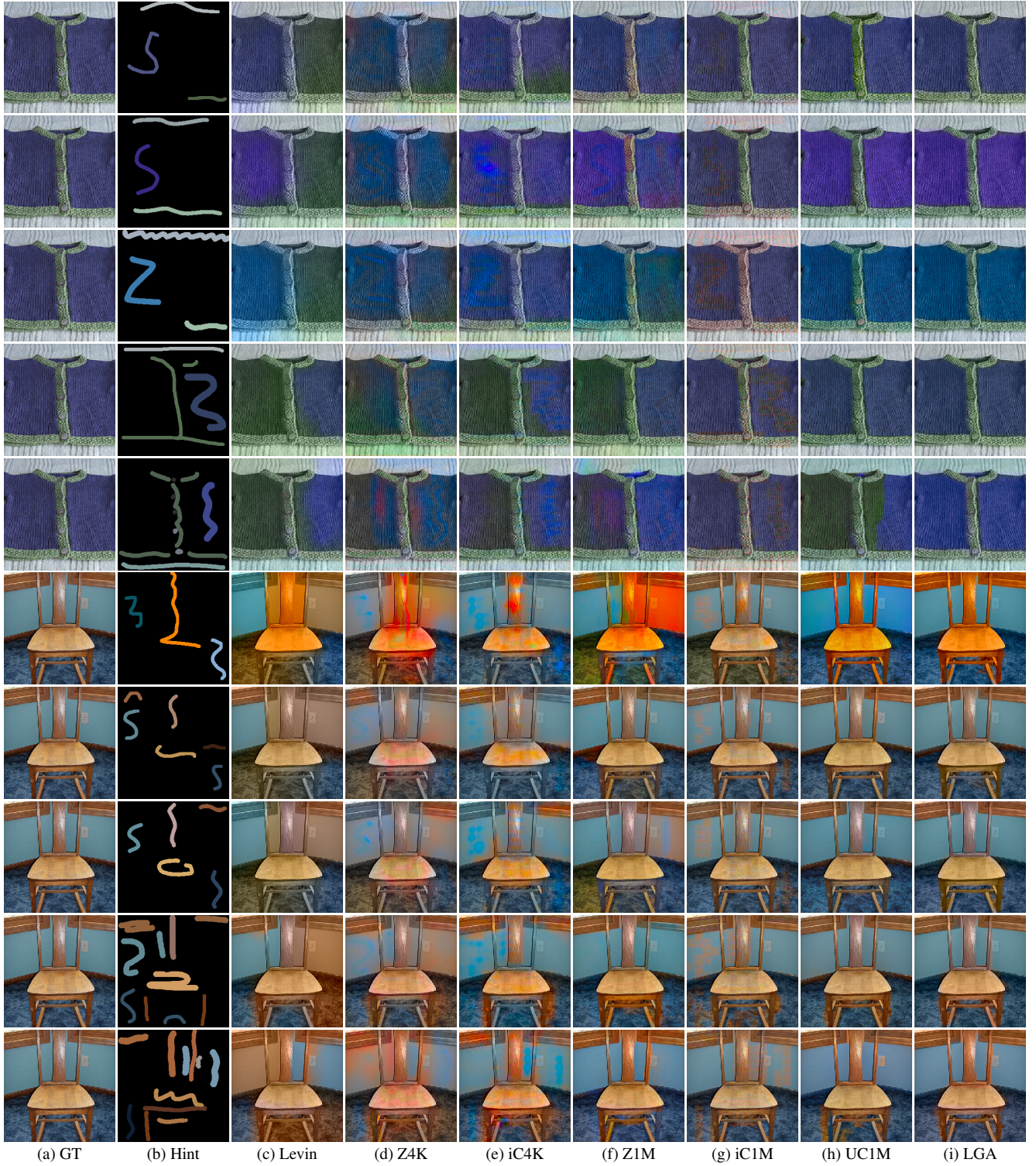


Figure 9. Usability Study: Evaluating Method Robustness Across Real and Diverse Scribble Inputs in Section 3. This figure shows 2 randomly selected examples under five scribbles from different users. The first two columns show the ground truth and five applied scribbles from different users, while the following columns separately display: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_t$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.



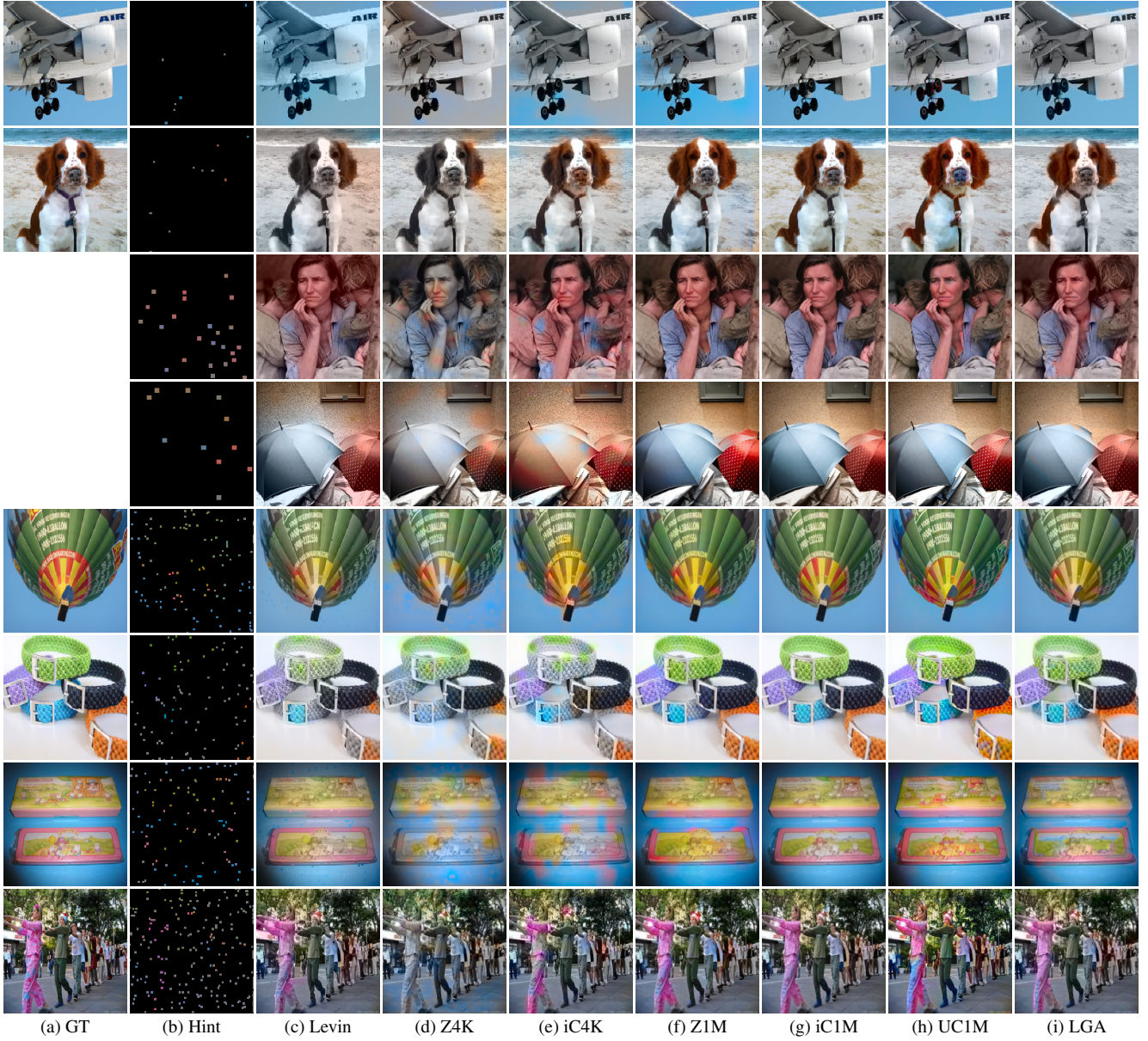


Figure 10. Qualitative comparison with state-of-the-art methods in Section 4. This figure displays 8 shown examples from other published papers (Zhang [5], UniColor [1] and iColoriT [4]). The first two columns show the ground truth and the applied scribbles, where the blank means there is no ground truth released from other papers. The following columns separately display the results of: Levin [2]; Z4K and iC4K (Zhang [5] and iColoriT [4] trained on  $\mathcal{D}_t$ ); Z1M, iC1M and UC1M (official pre-trained models on ImageNet for Zhang et al. [5], iColoriT [4] and UniColor [1]); LGA-Net.