Supplementary Material:
Mask-guided Matting in the Wild

Kwanyong Park[1] Sanghyun Woo[1] Seoung Wug Oh[2] In So Kweon[1] Joon-Young Lee[2]

[1]KAIST [2]Adobe Research

Appendices

This supplementary material contains more details including:

A. Illustrative examples of Composition-Wild,

B. More qualitative comparisons,

C. Further ablation study and analysis,

D. Implementation details.

A. Illustrative Examples of Composition-Wild

In Fig. 1, we visualize examples of the proposed Composition-Wild benchmark, including composited images, guidance, and groundtruth alpha matte. Composition-Wild is a challenging proxy of in-the-wild matting dataset, where the complex interaction between multiple objects occurs.

Figure 1. **Illustrative examples of Composition-Wild.**

B. More Qualitative Comparisons

Here we provide more qualitative comparisons on diverse scenarios: mask-guided matting on the coco dataset, mask-guided video matting, and panoptic matting.

B.1. COCO Dataset

Comparison with trimap-based method. In the main paper, we compare our proposal to a baseline of the mask-guided matting model [8]. In this section, we conduct qualitative comparisons with a trimap-based baseline. To automatically generate a trimap from the given mask, we apply simple heuristics, treating the boundary of the given mask as an uncertain region. Given this trimap, we run the state-of-the-art trimap-based matting model (MatteFormer [6]) using their official pre-trained weights. The results are summarized in Fig. 2. While the baseline also produces reasonable alpha matte for defined uncertain regions (see the first example in Fig. 2), such heuristics cannot detect diverse patterns of uncertain regions in the wild. As a result, the trimap-based baseline fails to correct the large interior in mask (second example) or predict plausible opacity value for transparent regions (third and fourth examples). Thus, for the trimap-based baseline, additional human interaction is necessary to handle such challenging cases. On the contrary, our mask-guided matting model semantically understands the given guidance and produces a pleasuring alpha matte without additional interaction.

Qualitative Results with Different Guidance. We analyze how our model performs with different guidance. To this end, we test two different types of mask guidance, predictions of the off-the-shelf model [3] or hand drawn mask [5]. As shown in Fig. 3, our model predicts fine details of alpha matte regardless of quality or type of guidance.

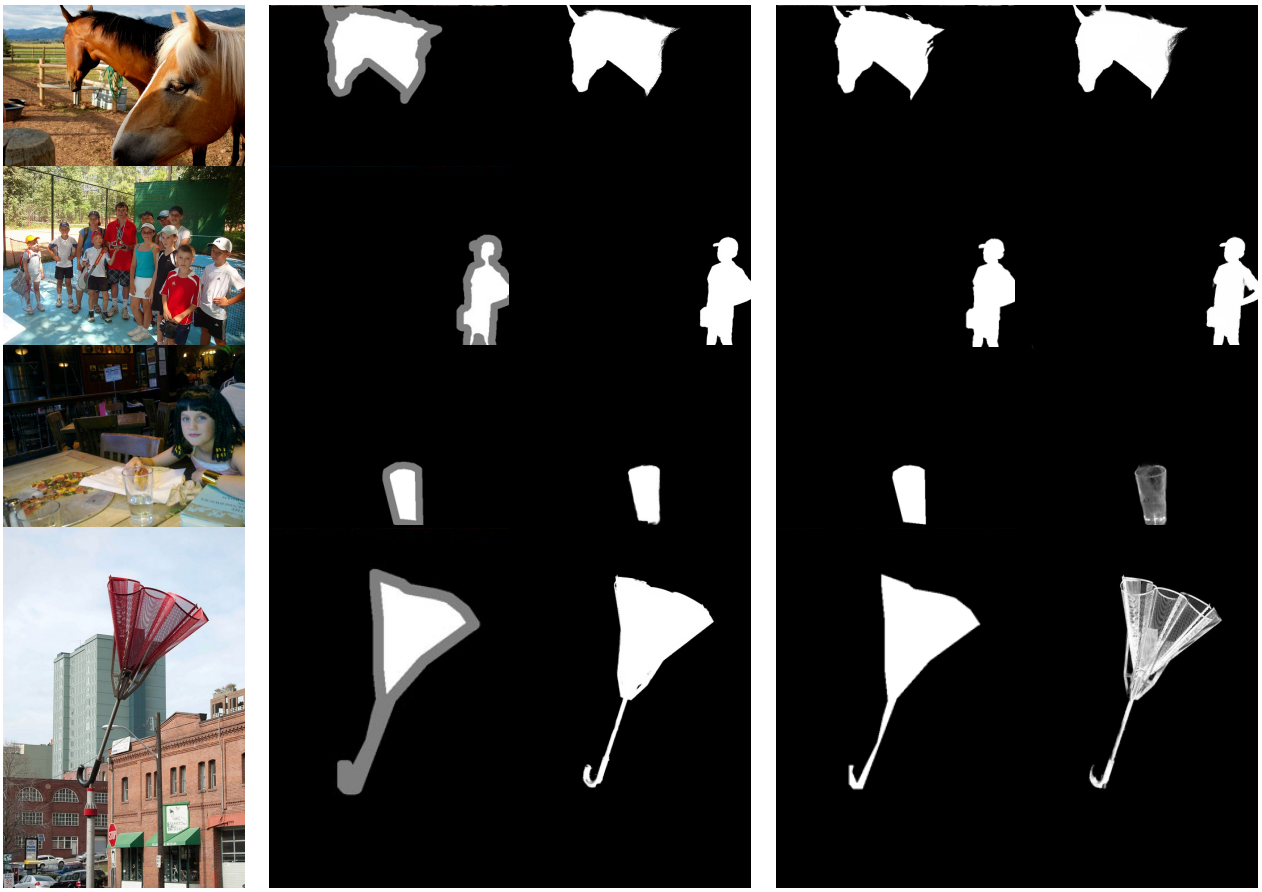

Figure 2. **Qualitative comparisons with trimap-based method.** We indicate the uncertain region in trimap as gray color.

B.2. Mask-guided Video Matting

In our supplementary video, we provide qualitative comparisons on the task of mask-guided video matting. In this setting, only a single instance mask is given in the first frame. We leverage a video object segmentation network [1, 7] to propagate the given mask to the rest of the video frames. Then the propagated masks are utilized as guidance to the mask-guided matting model.

Despite the noise in propagated masks, our model is able to differentiate the target objects from the surroundings clearly. Besides, it produces reasonable and temporally consistent opacity in challenging scenarios such as motion blur. We could seamlessly replace the background with other videos with the predicted alpha matte. On the contrary, the editing results derived from the baselines show severe artifacts: 1) the boundary of the composited target object is awkward when the propagated mask is naively utilized, and 2) non-target objects or background wrongly appears in the results of MGMatting [8].

B.3. Panoptic Matting

In this section, we show the additional qualitative comparisons on panoptic matting. Mask-guided matting model considers each segment of a given panoptic mask separately and produces the corresponding alpha matte. Panoptic matting results could be obtained via aggregation of these predictions. To visualize the soft transition between different segments, we mix the original color of each segment with the ratio of opacity value. We include a large volume of qualitative comparisons in Fig. 9 $\sim$ Fig. 38.

Strength. From the comparison between our model and MGMatting [8], we see the clear strength of our proposal.

- *Strong Instance Discrimination Ability*: Our model has a strong semantic understanding so that it captures the target object from the given mask and discriminates well from the nearby similar objects. On the contrary, the baseline model [8] mostly relies on low-level features and struggles to differentiate between objects, resulting in inferior matting quality. Representative examples are shown in Fig. 5.

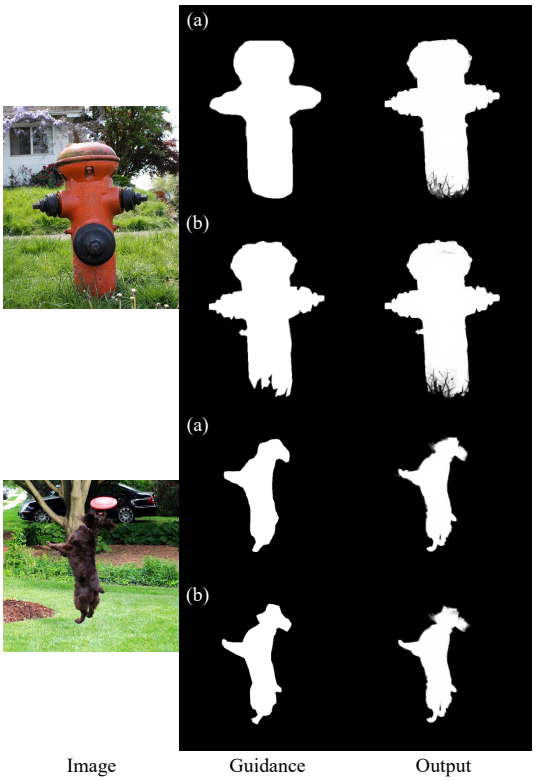

Figure 3. **Qualitative results with different guidance.** Guidance (a) and (b) denote model predicted masks [3] and manually drawn ones [5], respectively.

- *Great Generalization to Unseen Objects*: Under the proposed learning framework, the model learns generalized knowledge. Therefore it also well perform on unseen targets such as stuff region in Fig. 6. On the other hand, MGMatting [8] fails to handle such new targets and tends to produce zero opacity value for them.

Limitation. Although our model produces improved results in most cases, the model still shows limitations in several challenging scenarios. These are key aspects in order to build a more competitive mask-guided matting model.

- *Extremely Crowd Scene with Similar Objects*: As shown in Fig. 7, our model has trouble with handling extremely crowded scenes. The model can not capture the clear boundaries of the target objects.
- *Small Objects*: We observe that the model shows results of inferior fine details for small target objects (see Fig. 8).

C. Further Ablation Study and Analysis

Effects of Guidance Perturbations. In the main paper, we show that both image and guidance perturbation are key components to the success of the self-training framework.

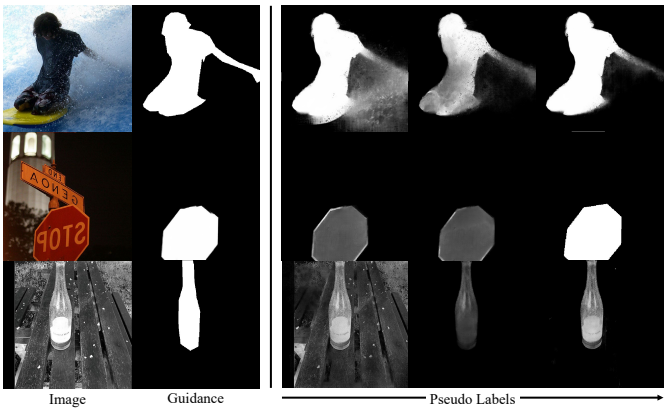

Figure 4. **Visualization of Pseudo Labels.** The quality of pseudo labels is improved through the self-training.

We further explore the effects of guidance perturbations. Table 1 summarizes the results of the ablation study under the different qualities of mask guidance. In particular, we dilate the mask guidance with different kernel sizes from 15 to 35 (guidance with smaller kernel sizes contains more precise information). Leveraging guidance perturbations in self-training brings clear improvements for all the settings. In addition, we empirically confirm that the guidance perturbations boost the robustness to noise in mask as more gains for noisy guidance are observed.

	Dilation Size (D)				
Method	D=15	D=20	D=25	D=30	D=35
No Guide.	16.46	17.28	18.27	20.21	22.08
Ours	15.11	15.97	16.72	18.18	19.35
Gain	1.36	1.30	1.56	2.03	2.73

Table 1. **Ablation study on Guidance Perturbation.** We report the SAD metric on AIM-500 [4] dataset under the different quality of mask-guidance.

Effects of Teacher-Student Framework. To hallucinate the high-quality pseudo labels, we introduce the teacher-student framework, where the teacher network is an exponential moving average (EMA) of the student model. We leverage stable predictions of the teacher network as pseudo labels and use them to guide the student network.

In this section, we analyze the effects of the teacher-student framework. We first visualize how the pseudo labels evolve as the training proceeds in the Fig. 4. At the beginning of the self-training, the pseudo labels are likely to have the wrong opacity value for in-the-wild objects. As the training proceeds, pseudo labels get calibrated and give valid supervision signals to the student networks. This results in more strong teacher network in turn, and the networks self-evolve during the training.

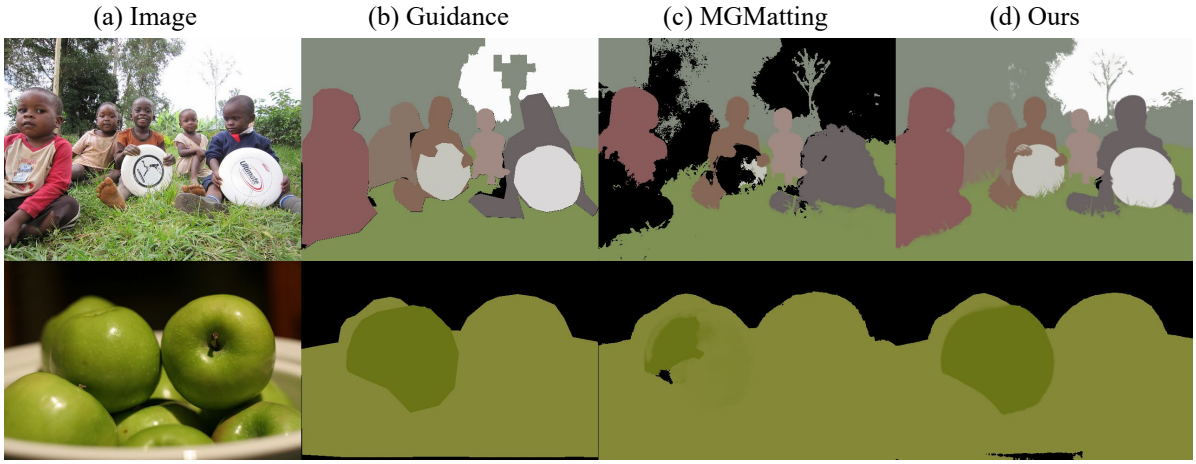

Figure 5. **Qualitative comparisons on images with similar objects.** Our model shows strong instance discrimination ability. Best viewed zoomed in.

Slowly advancing the teacher network is crucial to boost performance. When we replace the weights of the teacher network with that of the student network at every iteration (*i.e.,* fast update), the self-supervision becomes unstable, and the final performance drastically drops (See Table 2).

Method	Composition-Wild				AIM-500	
	SAD	MSE	SAD_{FG}	SAD_{BG}	SAD	MSE
Ours	58.16	0.0046	47.32	10.84	16.72	0.0030
Fast Update	61.48	0.0049	49.08	12.40	17.96	0.0034

Table 2. **Ablation Study on Update Strategy of Teacher Network.** By default, we slowly update the teacher network via exponential moving average (EMA) of the student network.

D. Implementation Details

Training details. For the pre-training stage on composited images, we follow the same training parameter as MGMatting [8]. For the fine-tuning stage, we fine-tune the network for 50,000 iterations with warm-up at the first 5,000 iterations. To form the strongly augmented version of the input image, we randomly apply linear contrast adjustment, brightness adjustment, channel shuffling, and additive gaussian noise as pixel-level augmentations. For the ablation study in Sec.6 of the main paper, we use gaussian blur and jpeg compression as region-level augmentation.

Editing details. We use [2] to extract the foreground color based on the image and predicted alpha matte. Then, for the background replacement, the extracted foreground objects are composited on new backgrounds following the composition formula.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[2] Thomas Germer, Tobias Uelwer, Stefan Conrad, and Stefan Harmeling. Fast multi-level foreground estimation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1104–1111. IEEE, 2021. 4

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 3

[4] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 3

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3

[6] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11696–11706, 2022. 1

[7] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1352–1361, 2022. 2

[8] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021. 1, 2, 3, 4

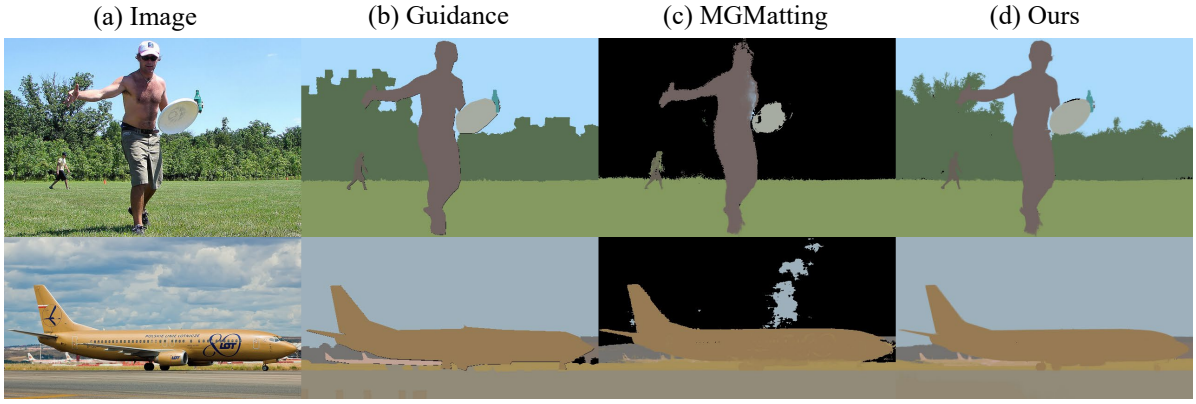

Figure 6. **Qualitative comparisons on images with unseen classes.** Our model successfully deals with unseen targets during training. Best viewed zoomed in.

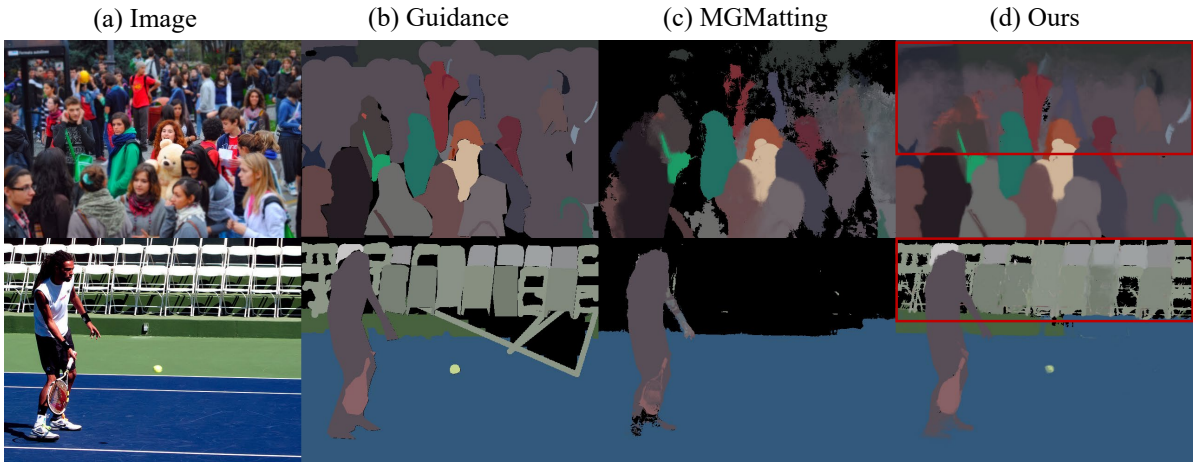

Figure 7. **Qualitative comparisons on extremely crowded images.** We indicate the red box for failure cases. Best viewed zoomed in.

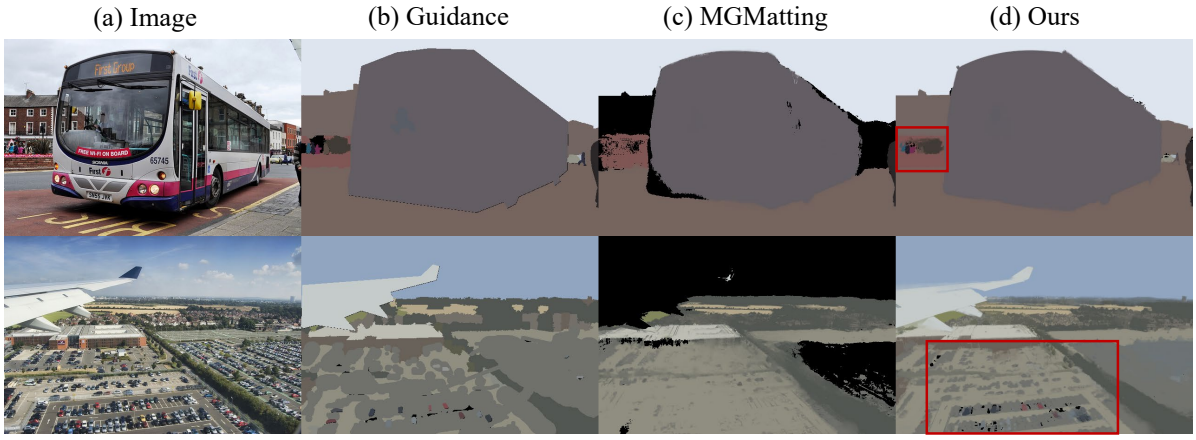

Figure 8. **Qualitative comparisons on images with small objects.** We indicate the red box for failure cases. Best viewed zoomed in.

Figure 9. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 10. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

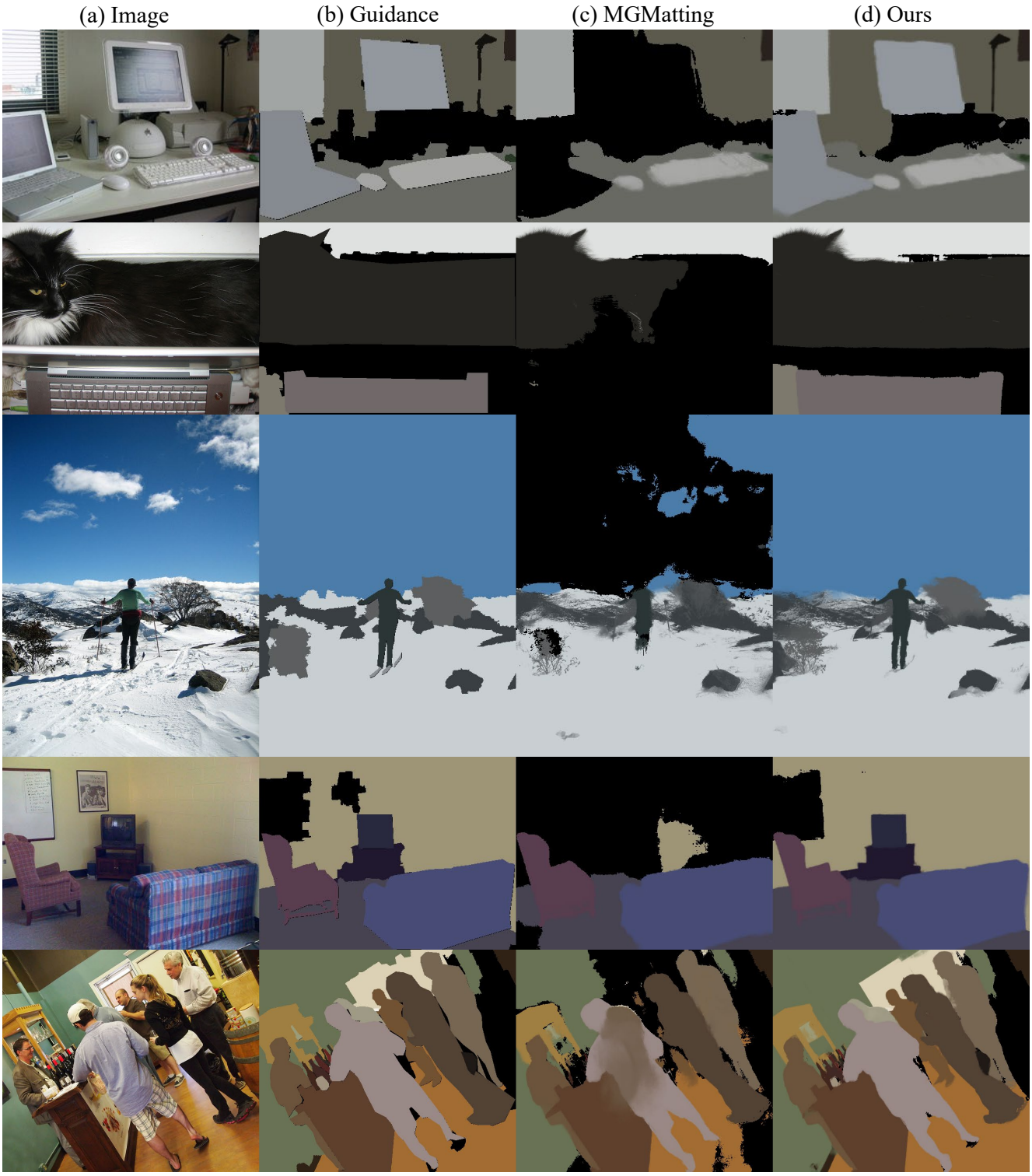

Figure 11. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

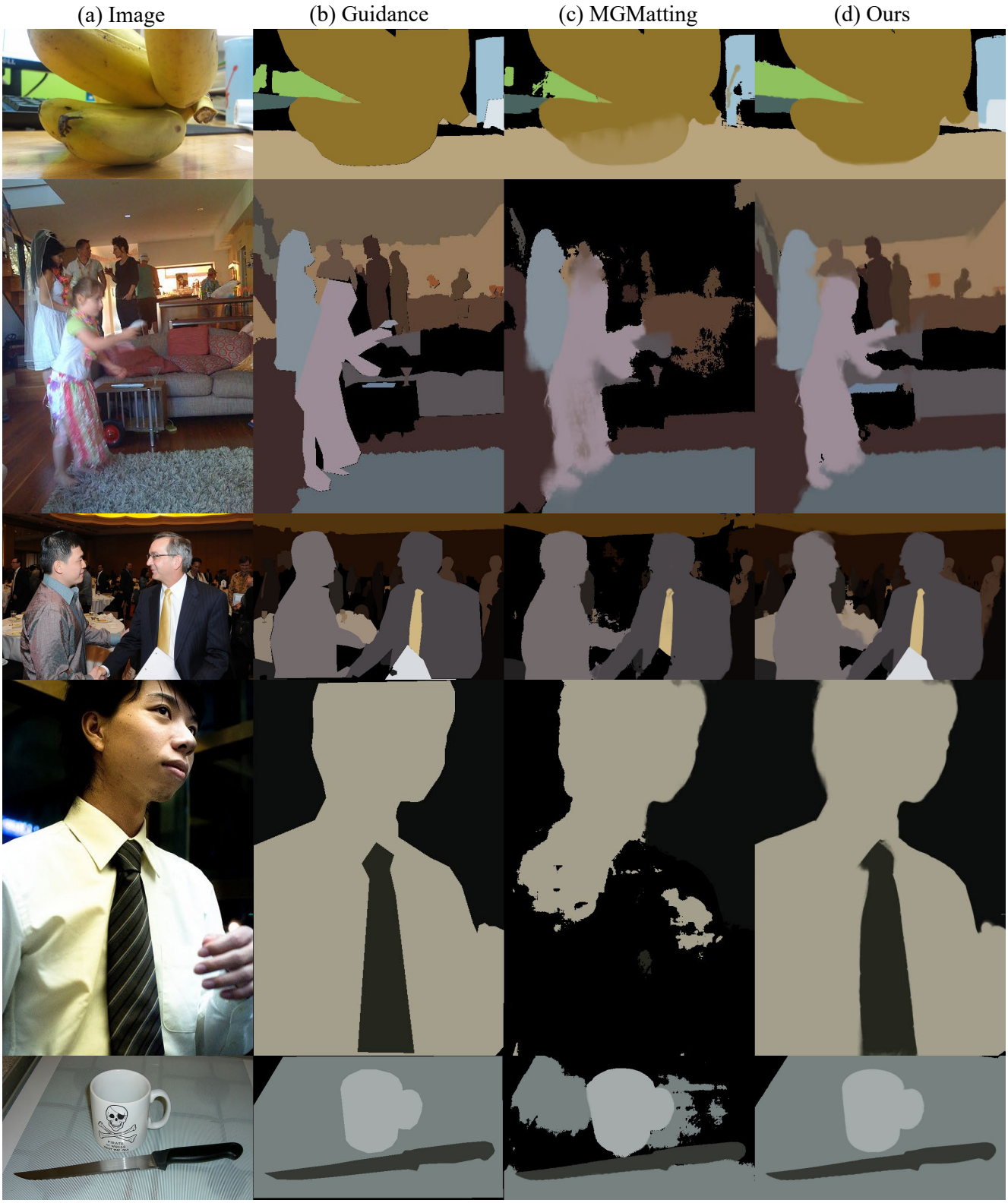

Figure 12. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 13. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 14. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

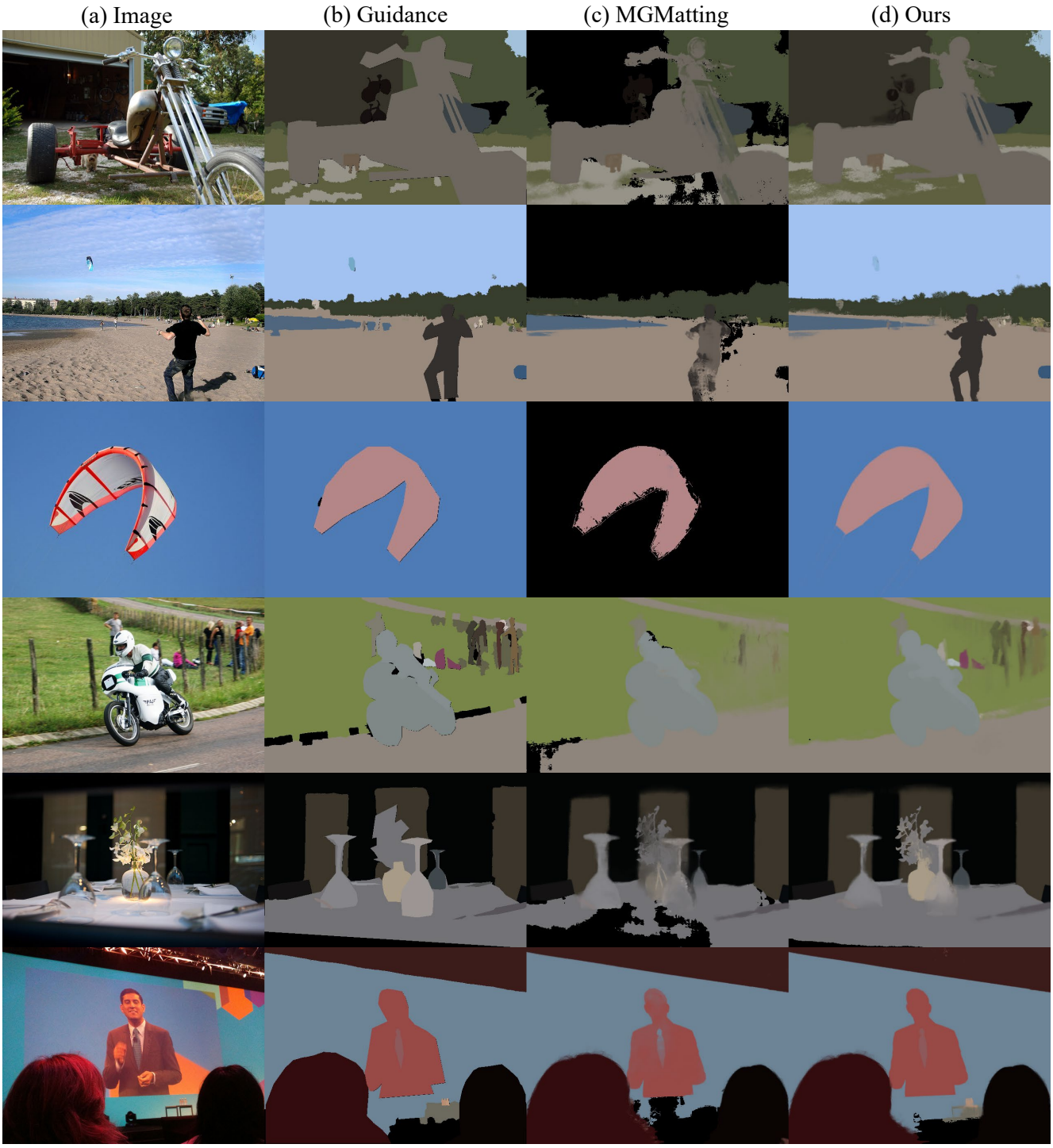

Figure 15. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

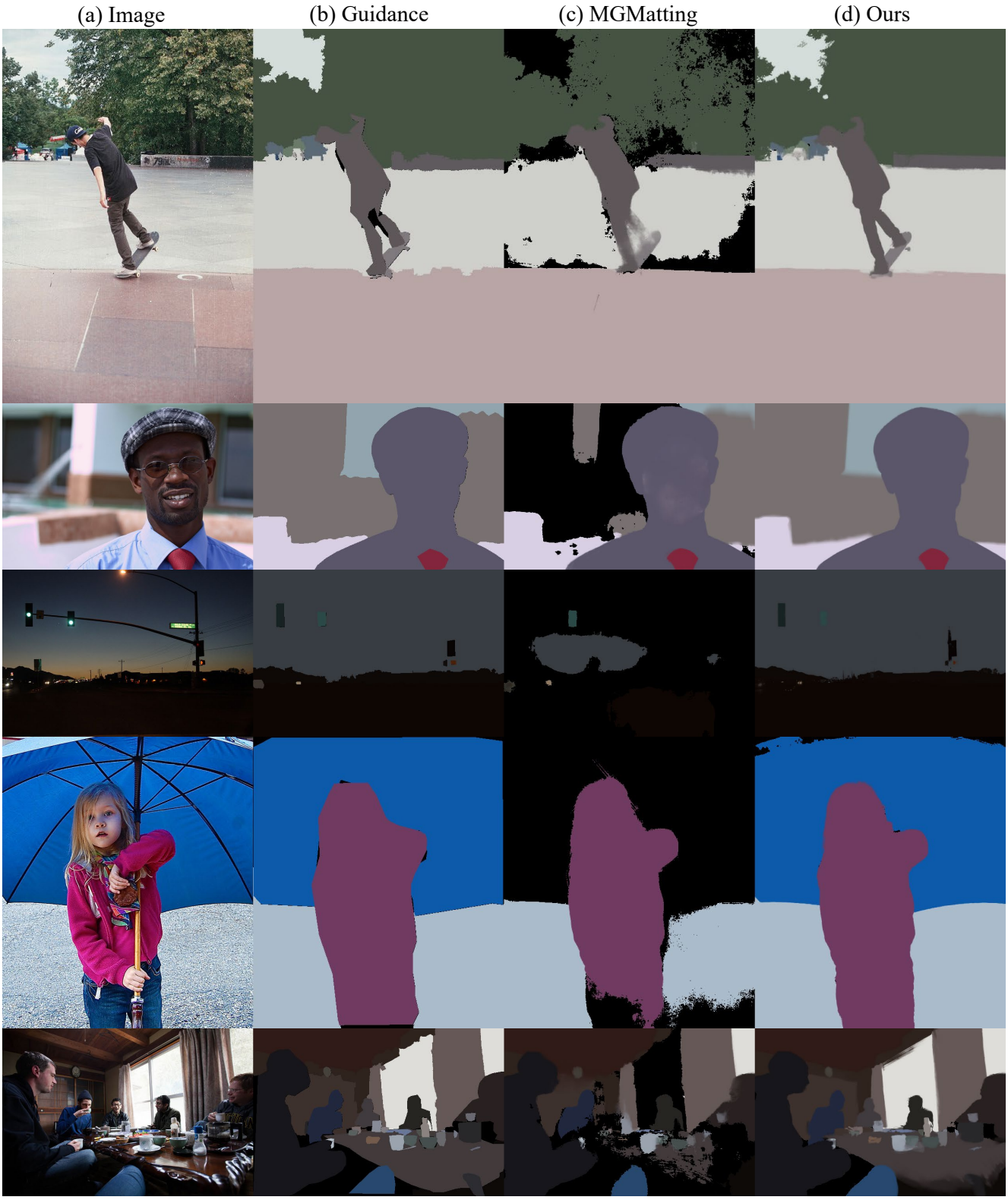

Figure 16. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 17. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 18. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

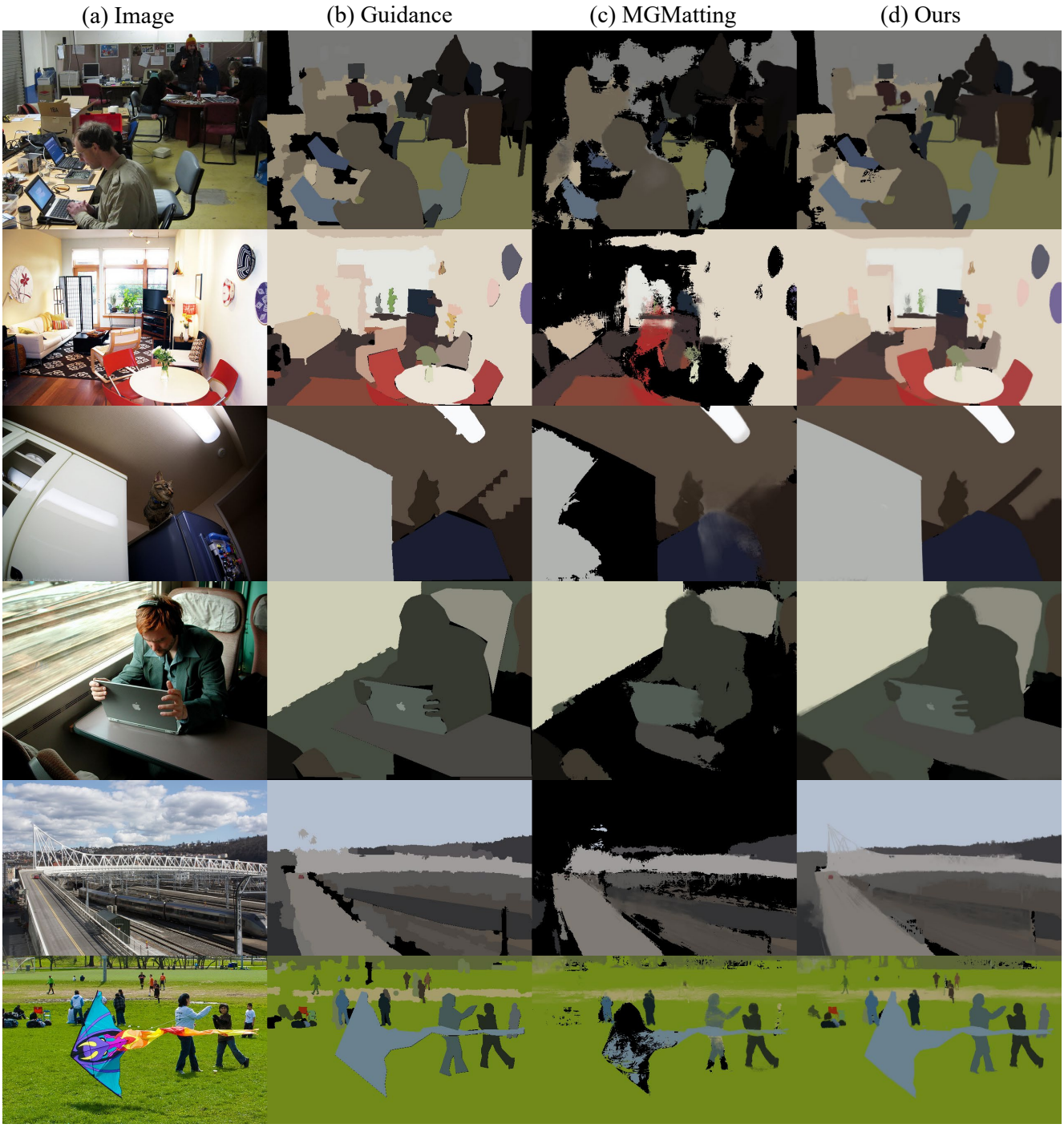

Figure 19. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

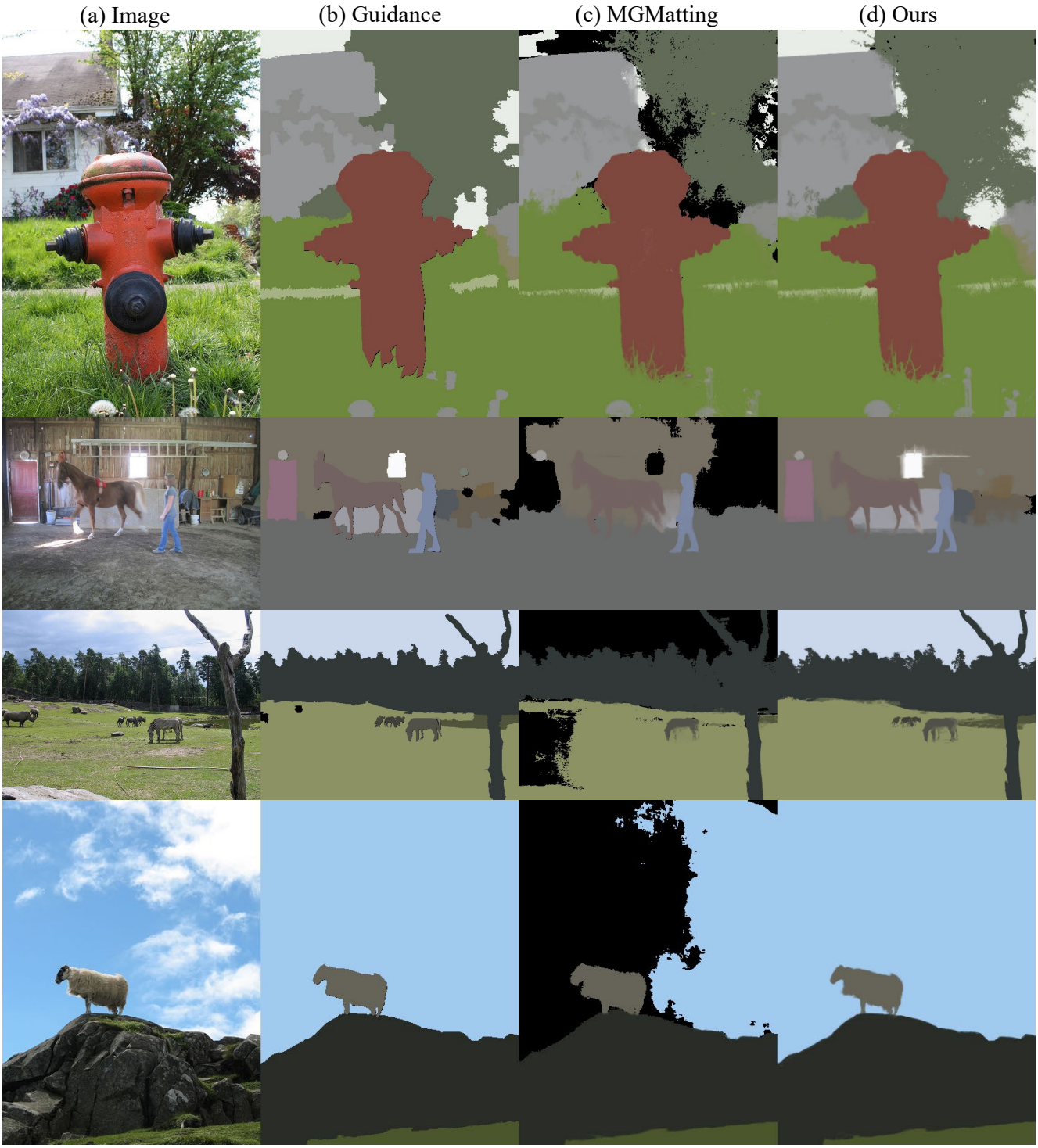

Figure 20. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

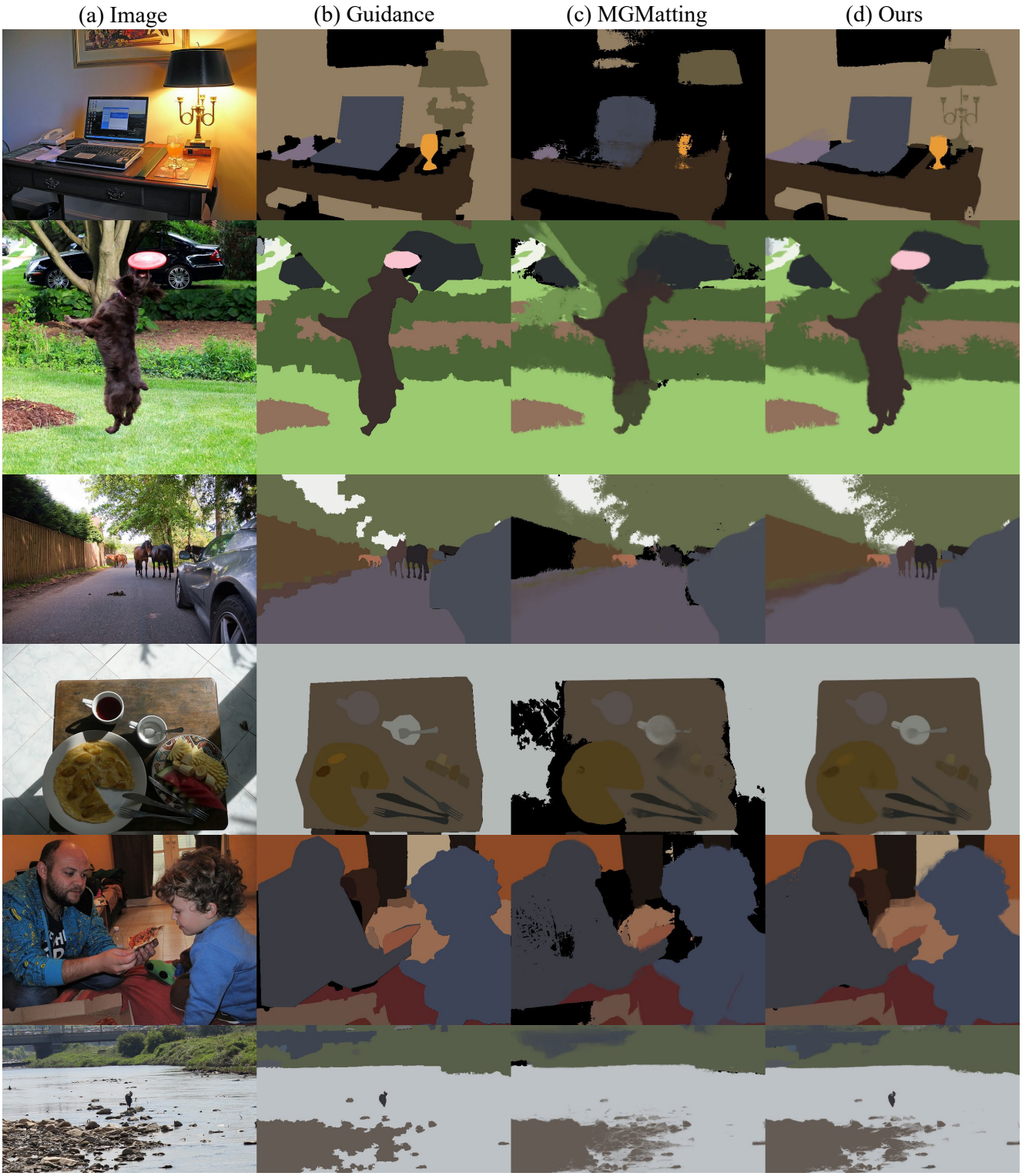

Figure 21. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

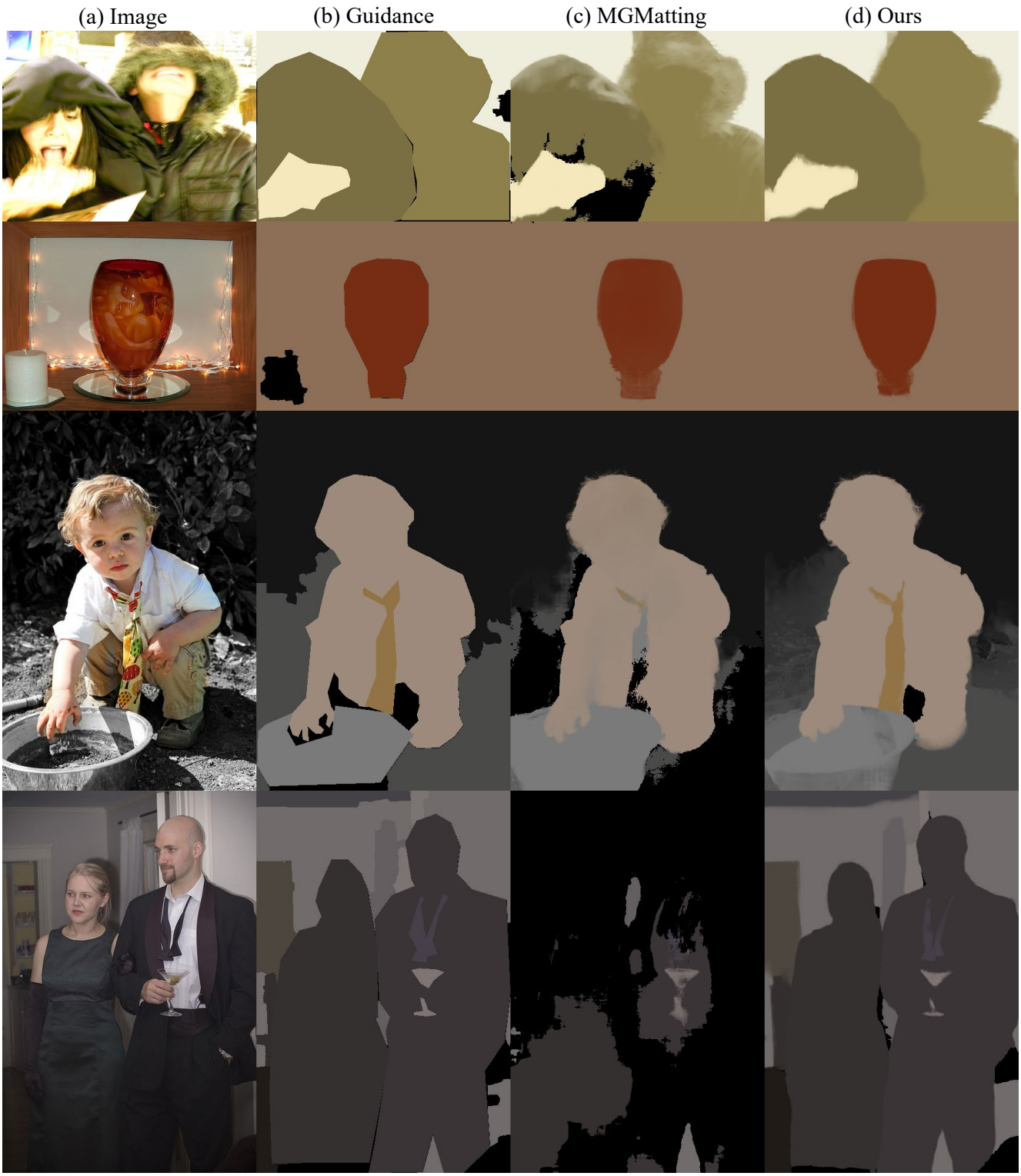

Figure 22. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 23. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

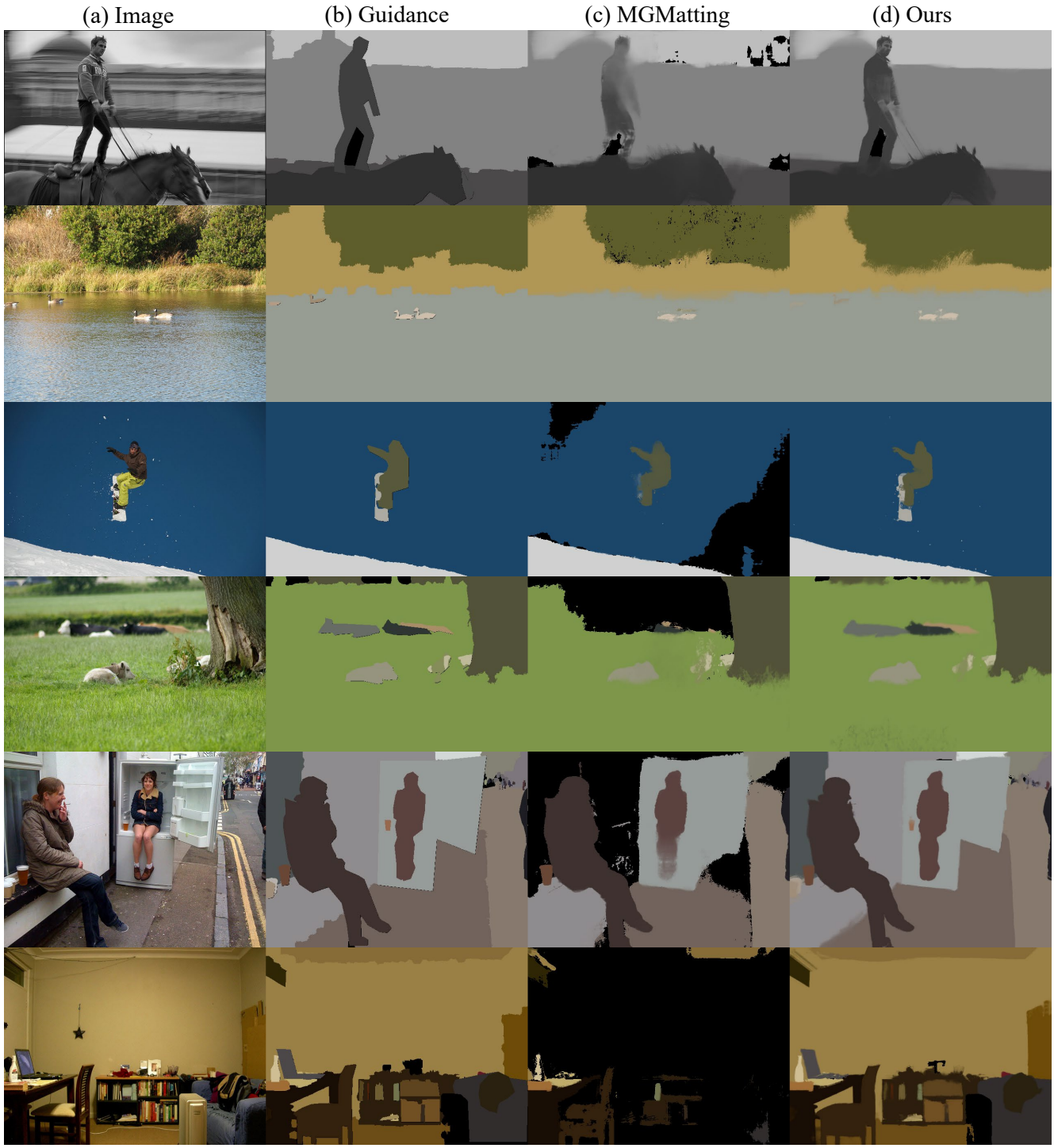

Figure 24. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 25. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 26. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

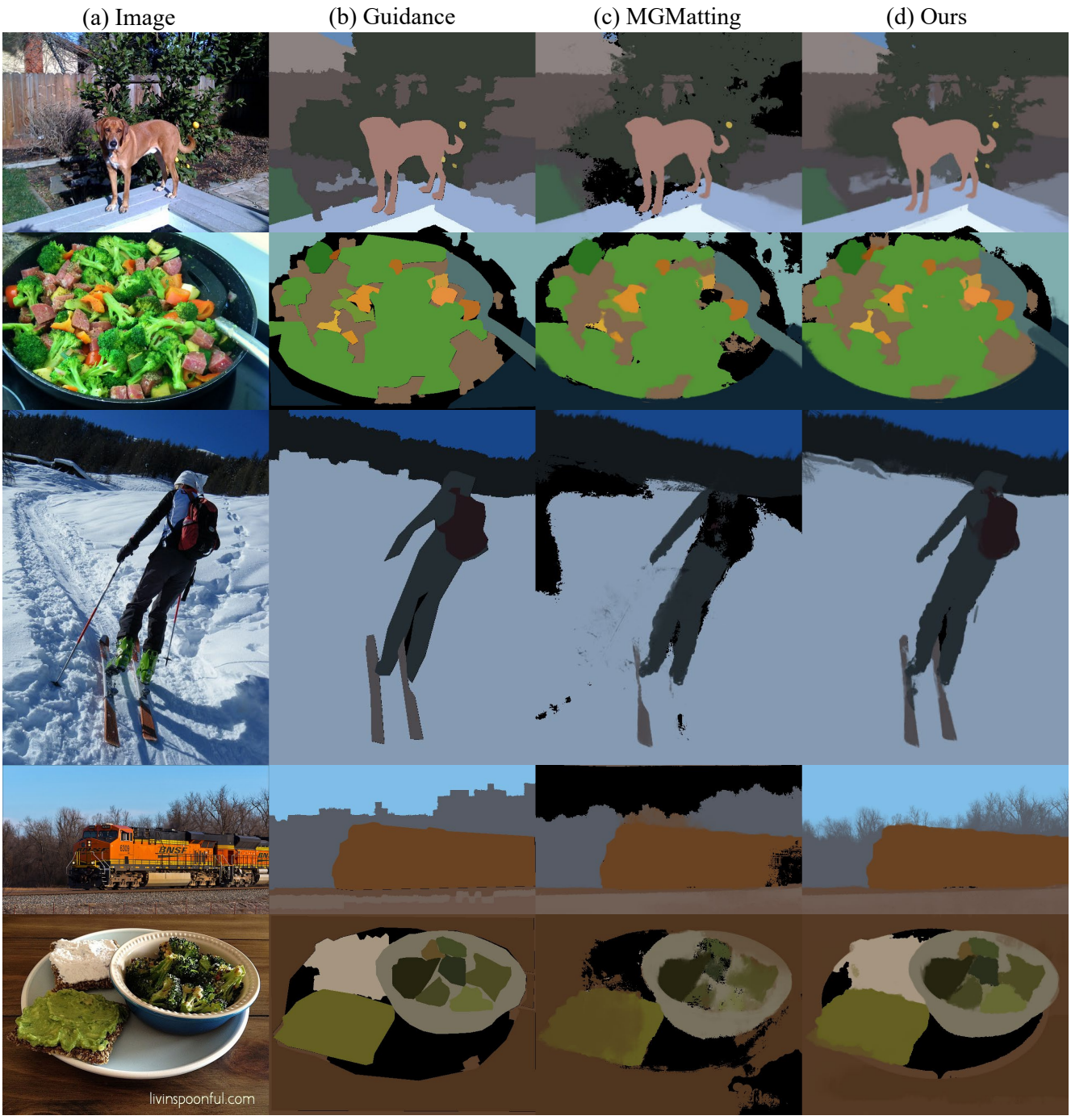

Figure 27. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 28. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

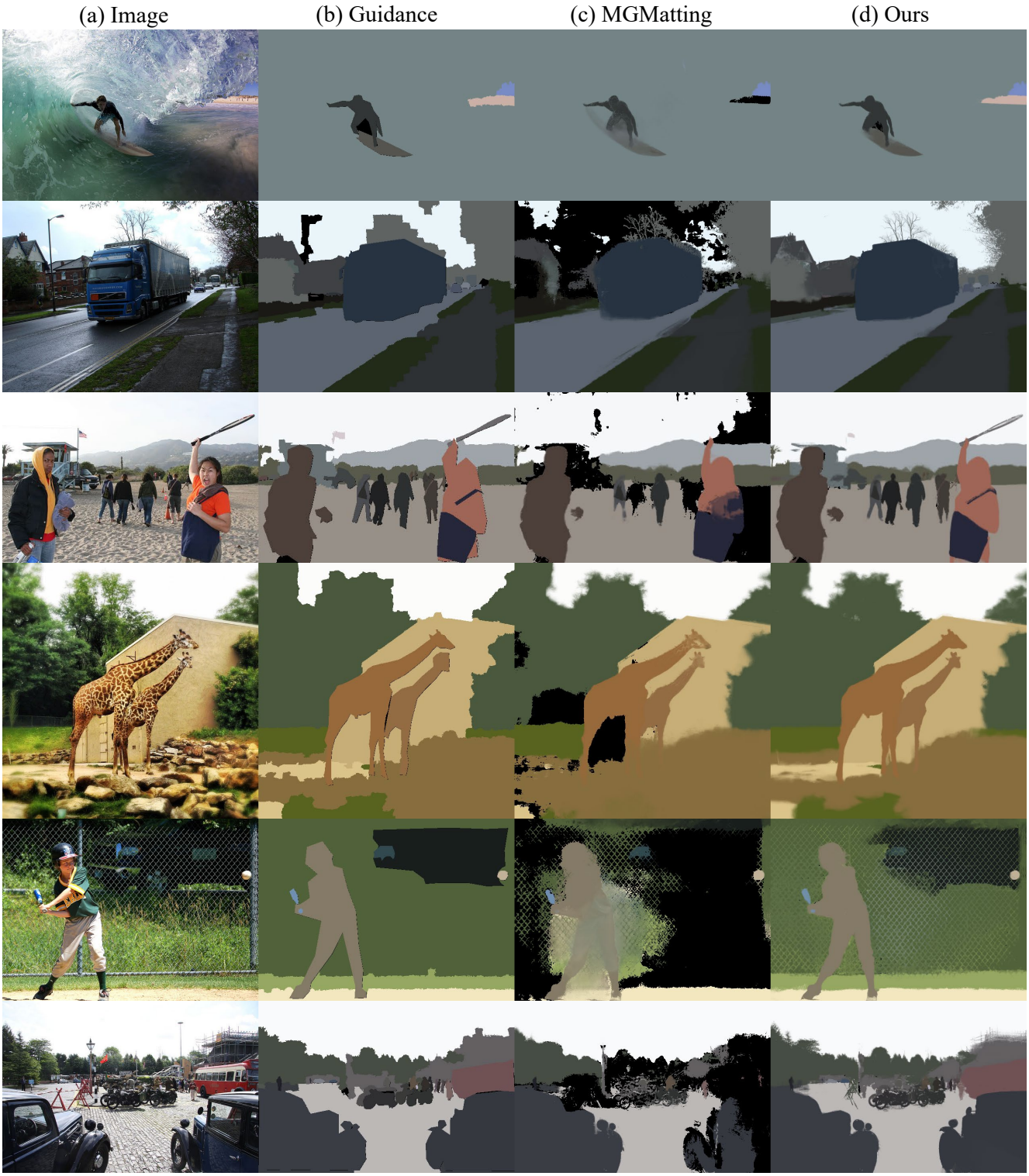

Figure 29. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

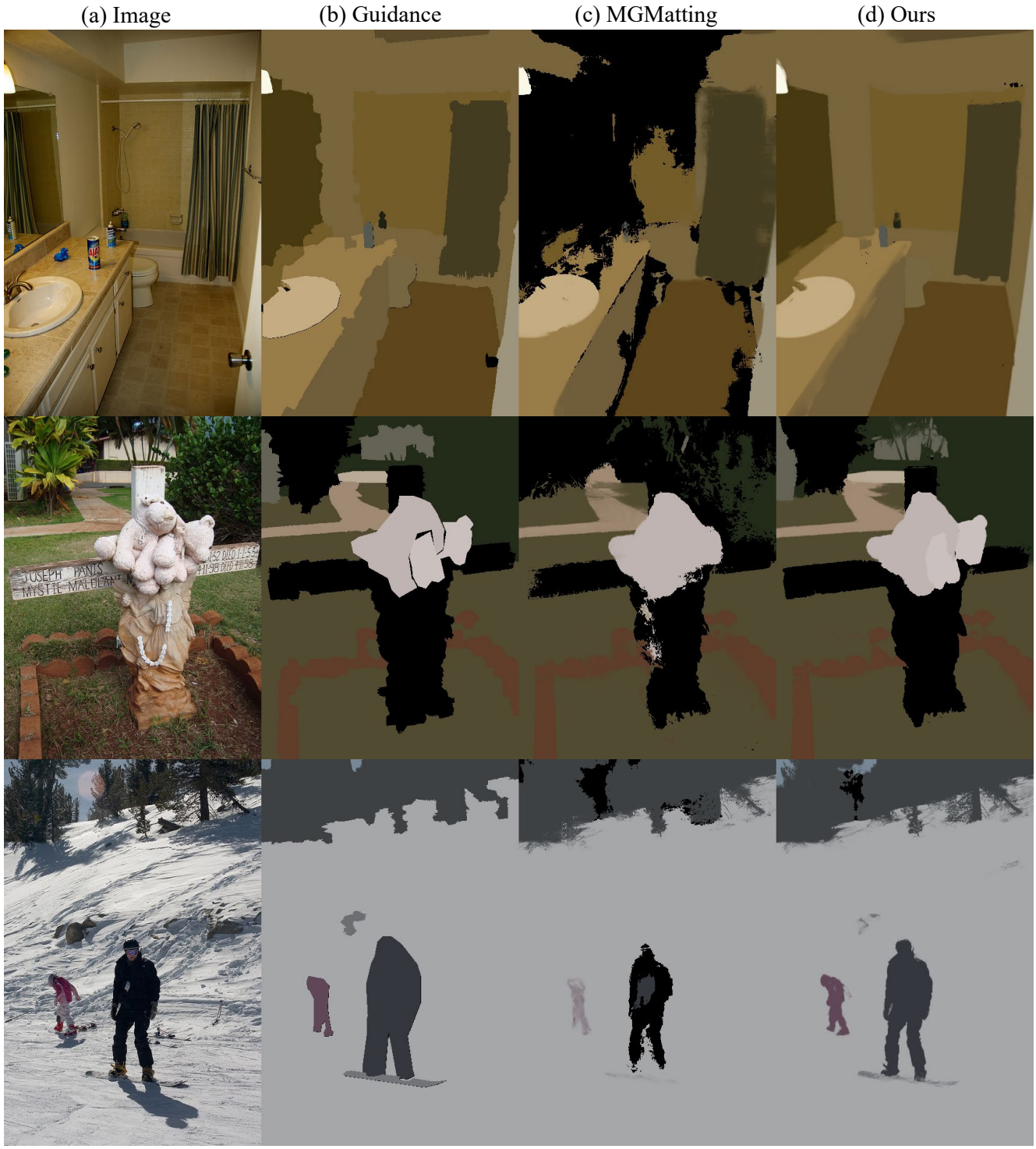

Figure 30. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 31. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 32. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 33. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

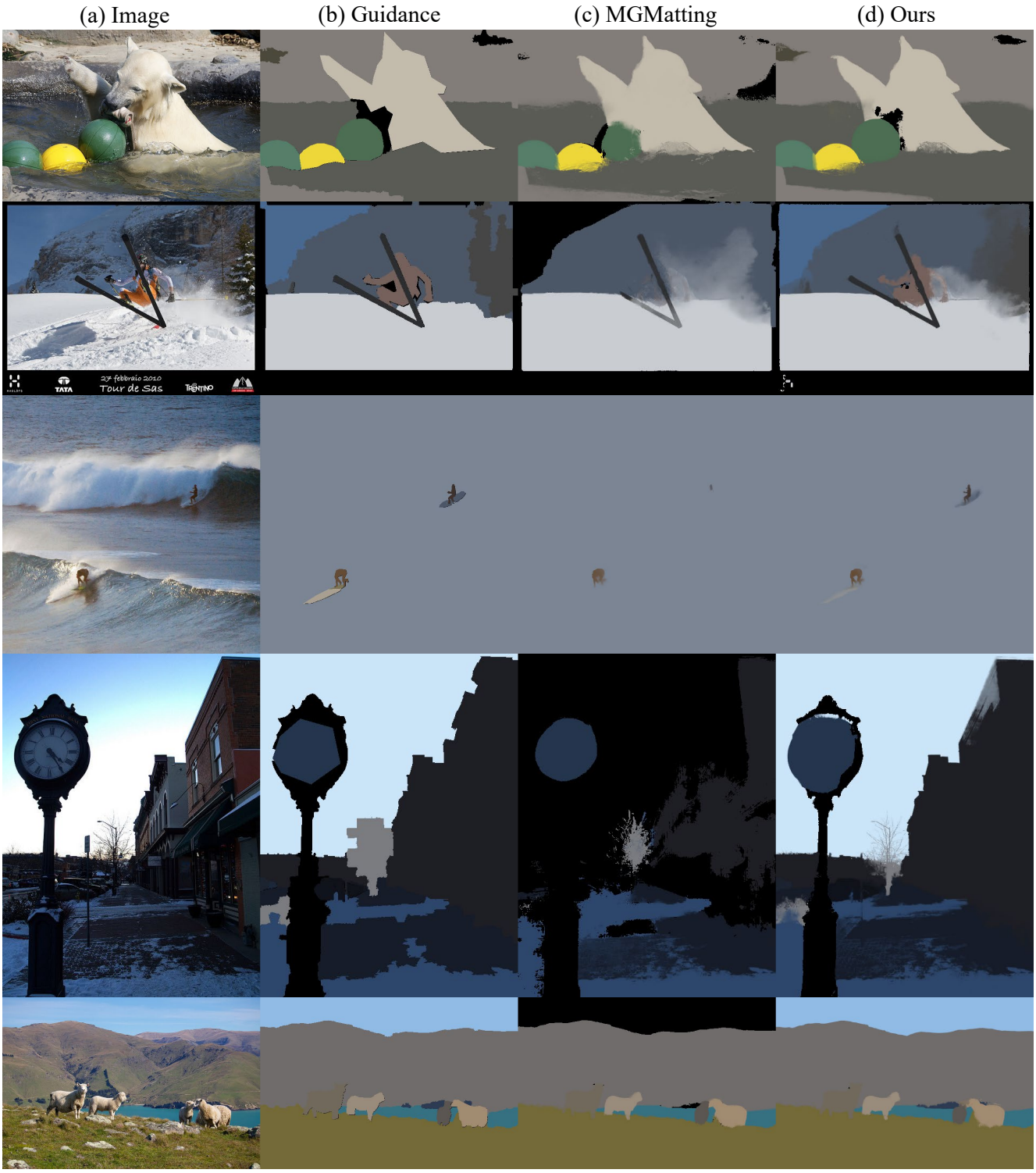

Figure 34. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 35. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 36. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

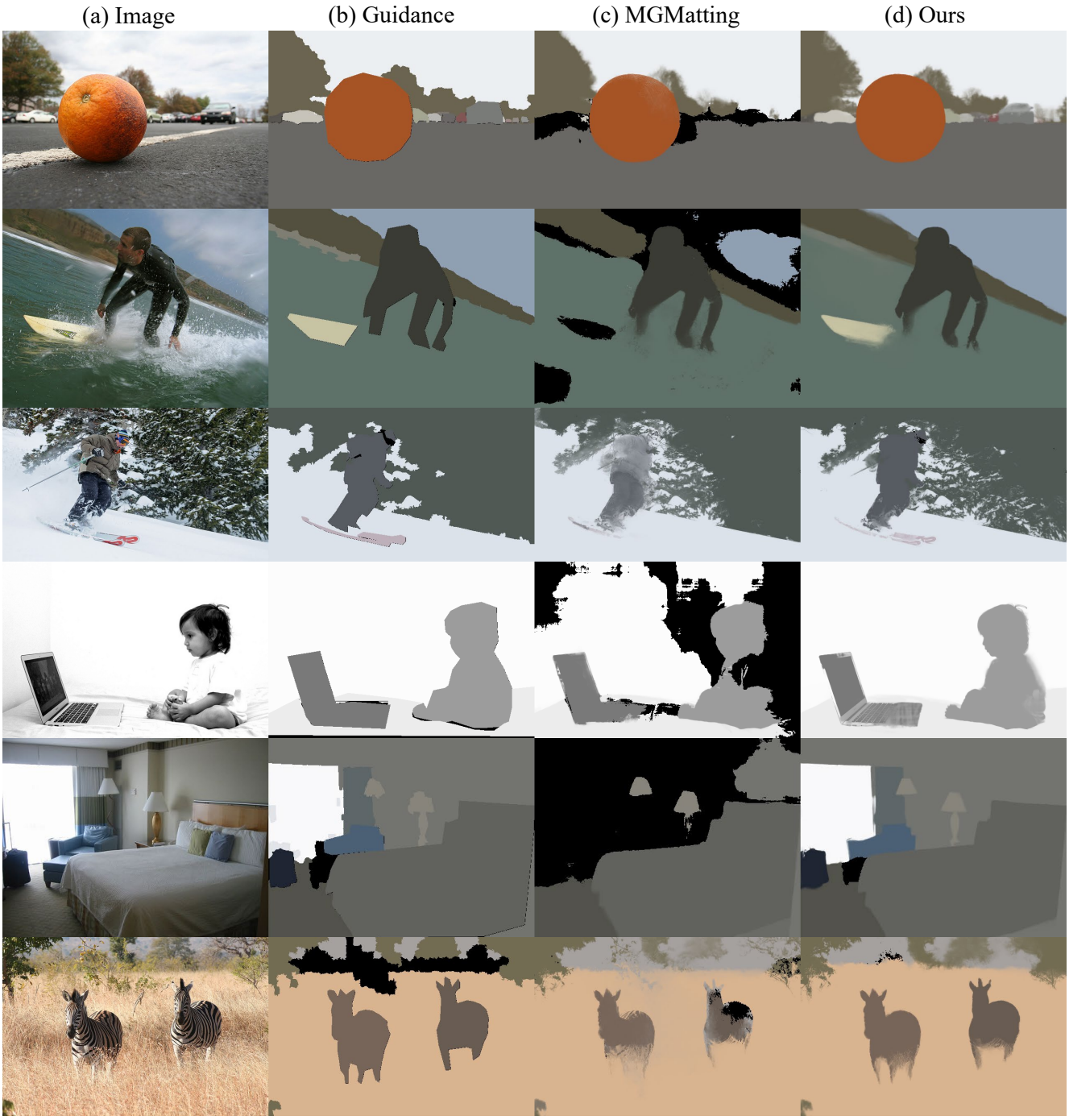

Figure 37. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.

Figure 38. **Qualitative Comparisons on Panoptic Matting.** Best viewed zoomed in.