SCS-Co: Self-Consistent Style Contrastive Learning for Image Harmonization Supplementary Material

Yucheng Hang[1,*], Bin Xia[1,*], Wenming Yang[1,2,†], Qingmin Liao[1,2]
[1] Shenzhen International Graduate School, Tsinghua University, China
[2] Department of Electronic Engineering, Tsinghua University, China
{hangyc20, xiab20}@mails.tsinghua.edu.cn, {yang.wenming, liaoqm}@sz.tsinghua.edu.cn

A. Overview

In this supplementary, we provide the detailed structure of our image harmonization network G in Section B to ensure better understanding and reproducibility. Besides, we conduct more ablation studies about our BAIN in Section C. More comparison results on real composite images are shown in Section D. Finally, we discuss the limitation of our method in Section E.

B. Network Structure

The detailed structure of our image harmonization network G is shown in Table 1. Symbols of the operators are listed as follows:

- Conv(c_{in}, c_{out}, k, s, p): a convolution operation with c_{in} input channels, c_{out} output channels, kernel size of k, stride size of s, and padding p.
- ConvTrans(c_{in}, c_{out}, k, s, p): a transposed convolution operation with c_{in} input channels, c_{out} output channels, kernel size of k, stride size of s, and padding p.
- IN(n): instance normalization with n dimensions.
- RAIN(n): region-aware adaptive instance normalization [3] with n dimensions.
- BAIN(n): the proposed background-attentional adaptive instance normalization with n dimensions.

Following [1,3], we also use attention blocks to improve the performance of the simple U-Net architecture. Specifically, we add four attention blocks in the decoder part. The structure of our attention block is the same as [3].

C. More Ablation Studies

In the main text, we have conducted many ablation studies for our self-consistent style contrastive learning scheme (SCS-Co). To prove the effectiveness of our BAIN, we conduct many ablation studies on BAIN in this section.

Table 1. The structure of our image harmonization network.

#	**Layer name(s)**
0	Conv(3, 64, 3, 1, 1)
1	LReLU + Conv(64, 128, 4, 2, 1) + IN(128)
2	LReLU + Conv(128, 256, 3, 1, 1) + IN(256)
3	LReLU + Conv(256, 512, 4, 2, 1) + IN(512)
4	LReLU + Conv(512, 512, 3, 1, 1) + IN(512)
5	LReLU + Conv(512, 512, 4, 2, 1) + IN(512)
6	LReLU + Conv(512, 512, 3, 1, 1) + IN(512)
7	LReLU + Conv(512, 512, 4, 2, 1) + IN(512)
8	ReLU + ConvTrans(512, 512, 4, 2, 1) + BAIN(512)
9	Concat[#6, #8]
10	ReLU + ConvTrans(1024, 512, 3, 1, 1) + RAIN(512)
11	Concat[#5, #10]
12	ReLU + ConvTrans(1024, 512, 4, 2, 1) + RAIN(512)
13	Concat[#4, #12]
14	ReLU + ConvTrans(1024, 512, 3, 1, 1) + RAIN(512)
15	Concat[#3, #14] + Attention Block
16	ReLU + ConvTrans(1024, 256, 4, 2, 1) + RAIN(256)
17	Concat[#2, #16] + Attention Block
18	ReLU + ConvTrans(512, 128, 3, 1, 1) + RAIN(128)
19	Concat[#1, #18] + Attention Block
20	ReLU + ConvTrans(256, 64, 4, 2, 1) + RAIN(64)
21	Concat[#0, #20] + Attention Block
22	ReLU + ConvTrans(128, 3, 3, 1, 1) + Tanh

In Table 2, we construct three models: (1) A baseline model, which means we replace BAIN of layer #9 (see Table 1) with RAIN. (2) Our model with BAIN, which is consistent with Table 1. (3) For our BAIN, the process of obtaining an attention map based on the foreground-background feature similarity is similar to self-attention [6], and as we all know, self-attention is a powerful tool to improve model performance. Therefore, in order to prove that the gain brought by BAIN is not just because of this, we construct a model that replaces BAIN with a combination

of self-attention and RAIN.

Table 2. Ablation studies on our BAIN.

Method	PSNR↑	MSE↓	fMSE↓
w/ RAIN	37.55	27.81	294.64
w/ BAIN	37.84	25.23	269.05
w/ self-attention + RAIN	37.66	26.29	281.15

As shown in Table 2, we can find that model with our BAIN obtains the best performance. Besides, compared with model with the combination of self-attention and RAIN, model with our BAIN obtains a huge performance gain. These comparisons indicate that properly normalizing the foreground feature by the per-point attention-weighted background feature statistics according to the foreground-background feature similarity contributes to image harmonization greatly.

D. Results on Real Composite Images

In this section, we present more results of 99 real composite images released by [5] and compare our method with other state-of-the-art methods in Figure 1 to 12. As can be found, thanks to the proposed SCS-Co and BAIN, our method achieves more photorealistic visual results than other methods in most cases.

E. Limitation

Our method is a supervised method, so it needs high-quality paired harmonization data. However, collecting these paired data is time-consuming and laborious, which requires an accurate mask of the foreground object in each image. Recently, the first self-supervised image harmonization method is proposed [2] and achieves good performance. Since contrastive learning is a powerful tool for self-supervised learning, in the future we will explore how to build a self-supervised style contrastive learning scheme suitable for image harmonization.

References

[1] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[2] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, pages 4832–4841, October 2021. 2

[3] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, pages 9361–9370, 2021. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[4] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[5] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1

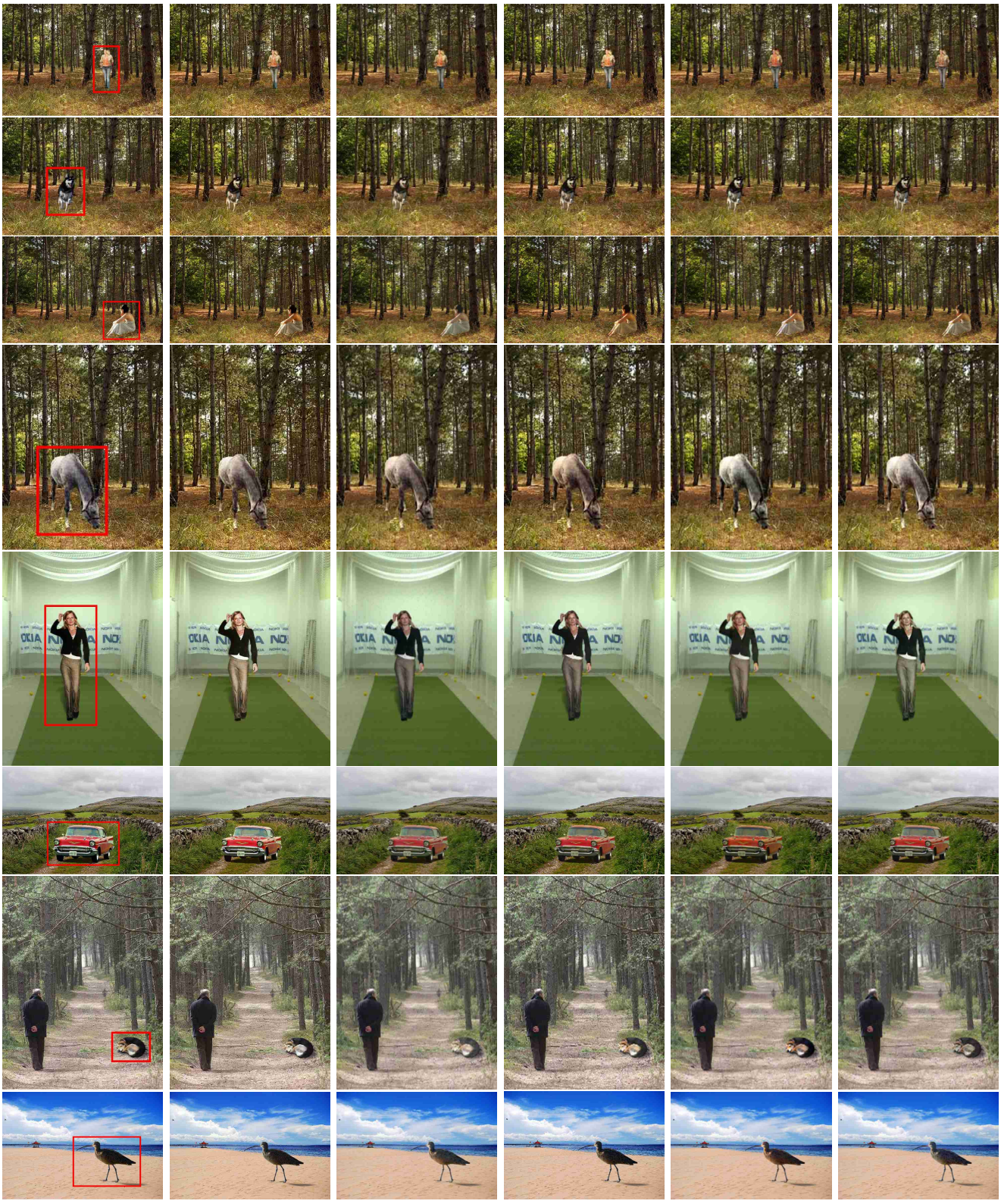

Figure 1. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

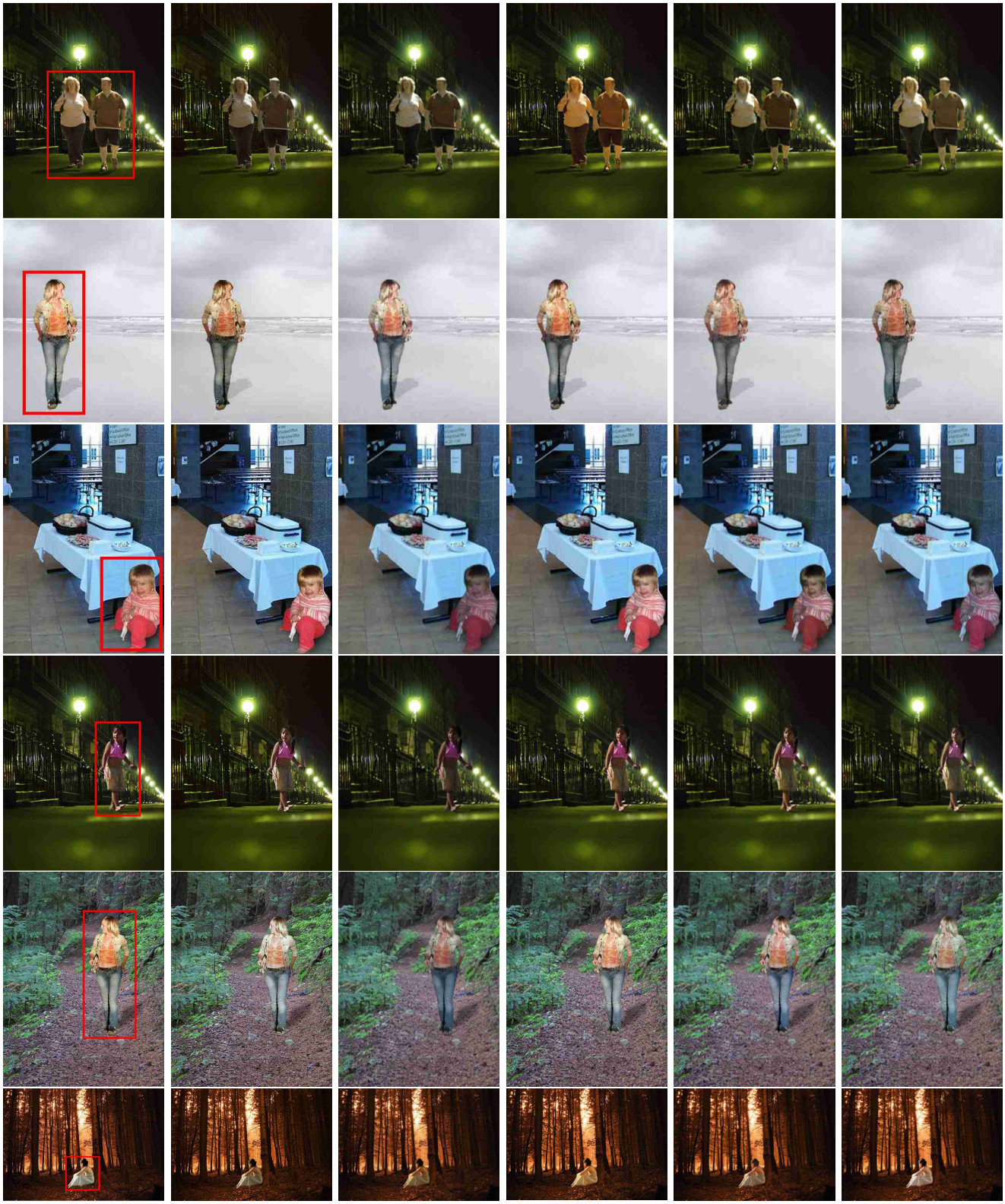

Figure 2. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

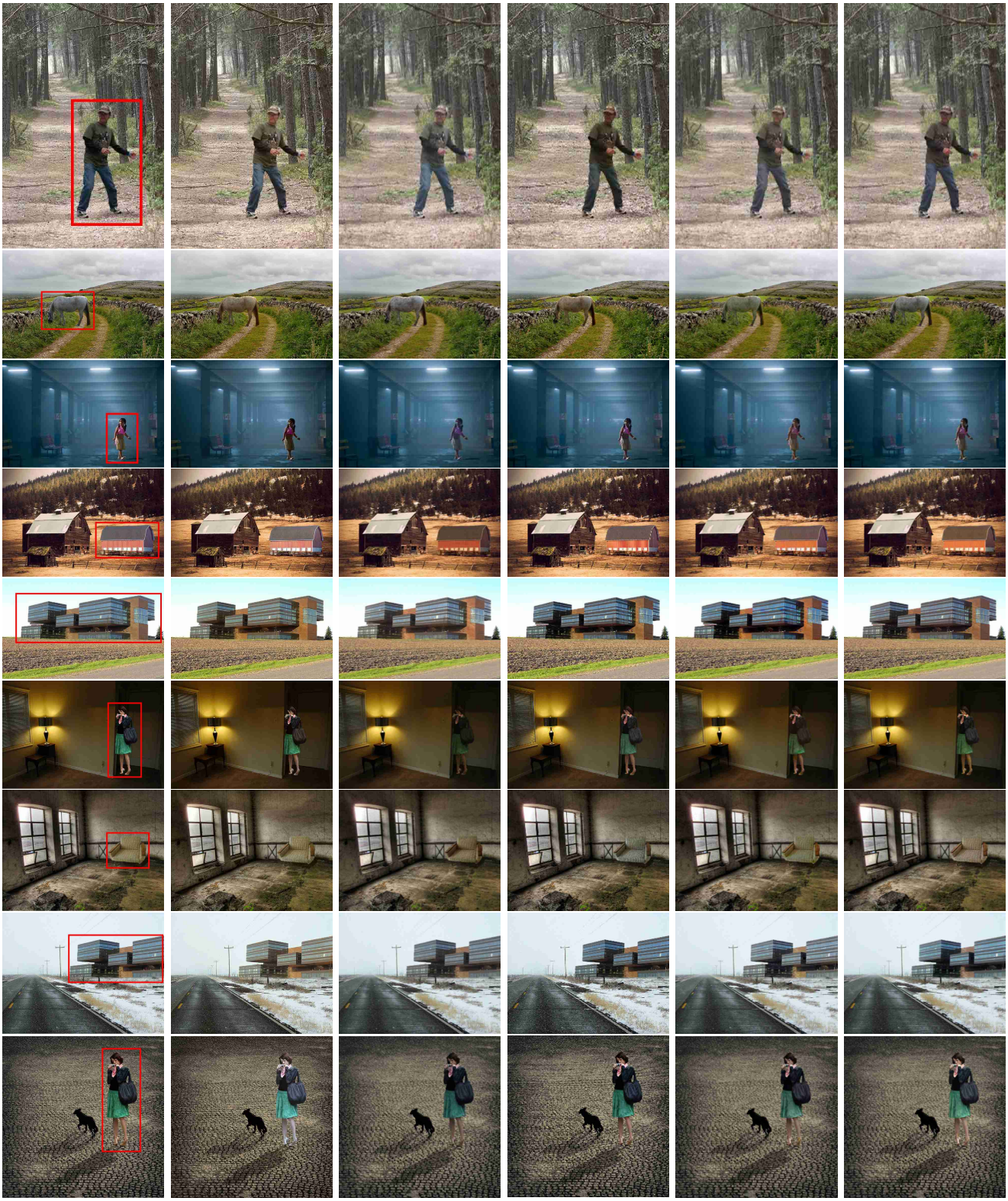

Figure 3. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

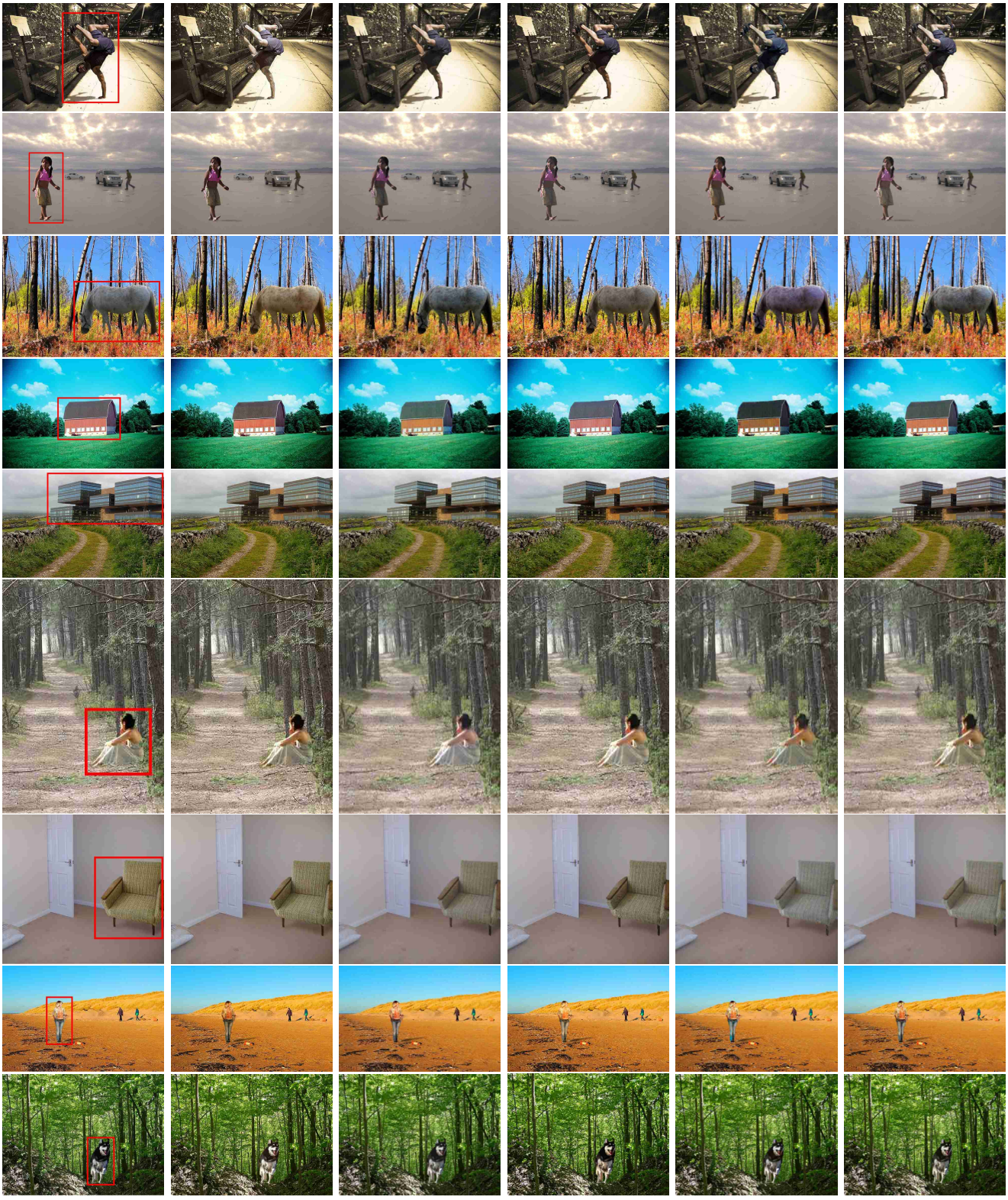

Figure 4. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

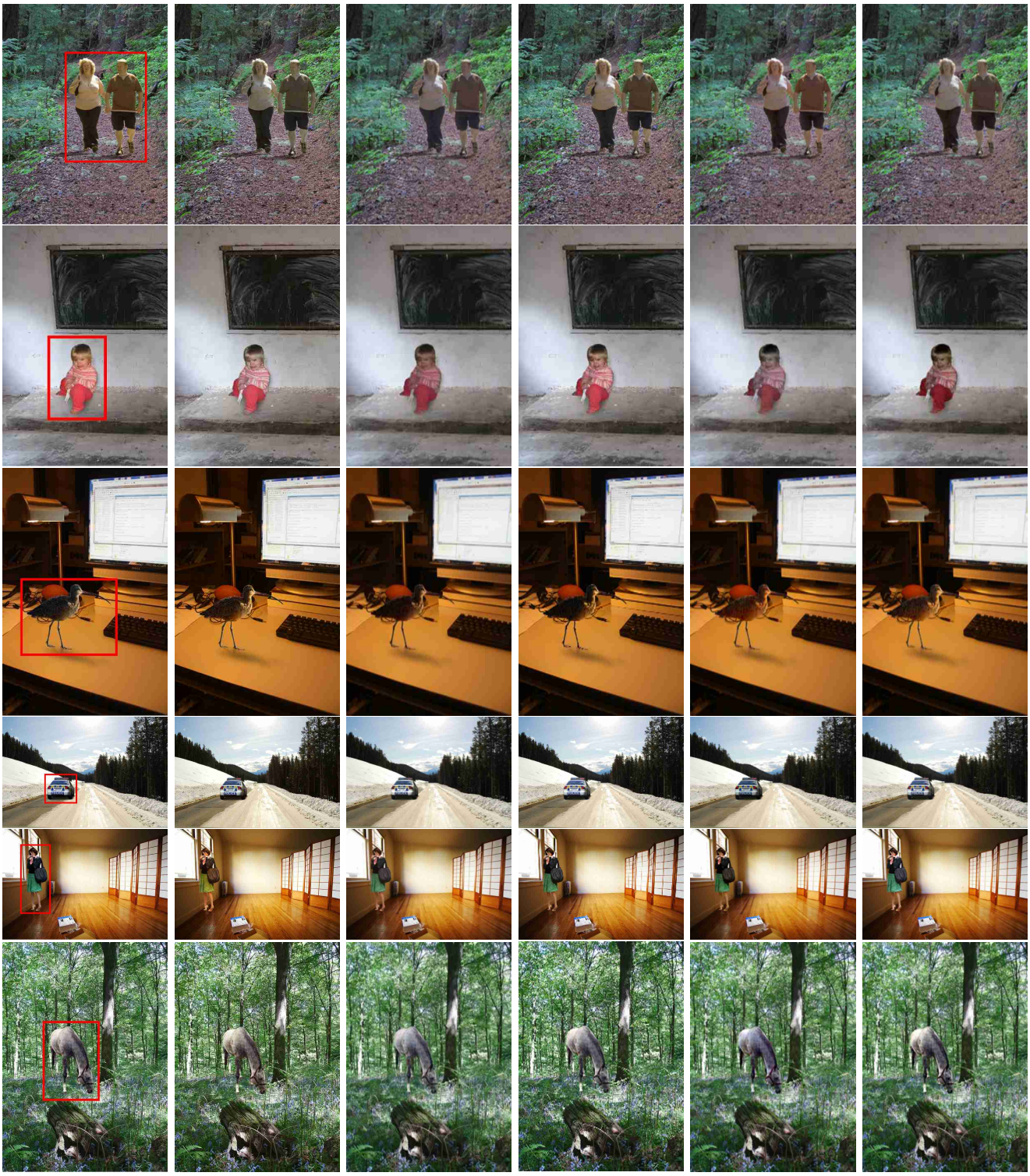

Figure 5. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

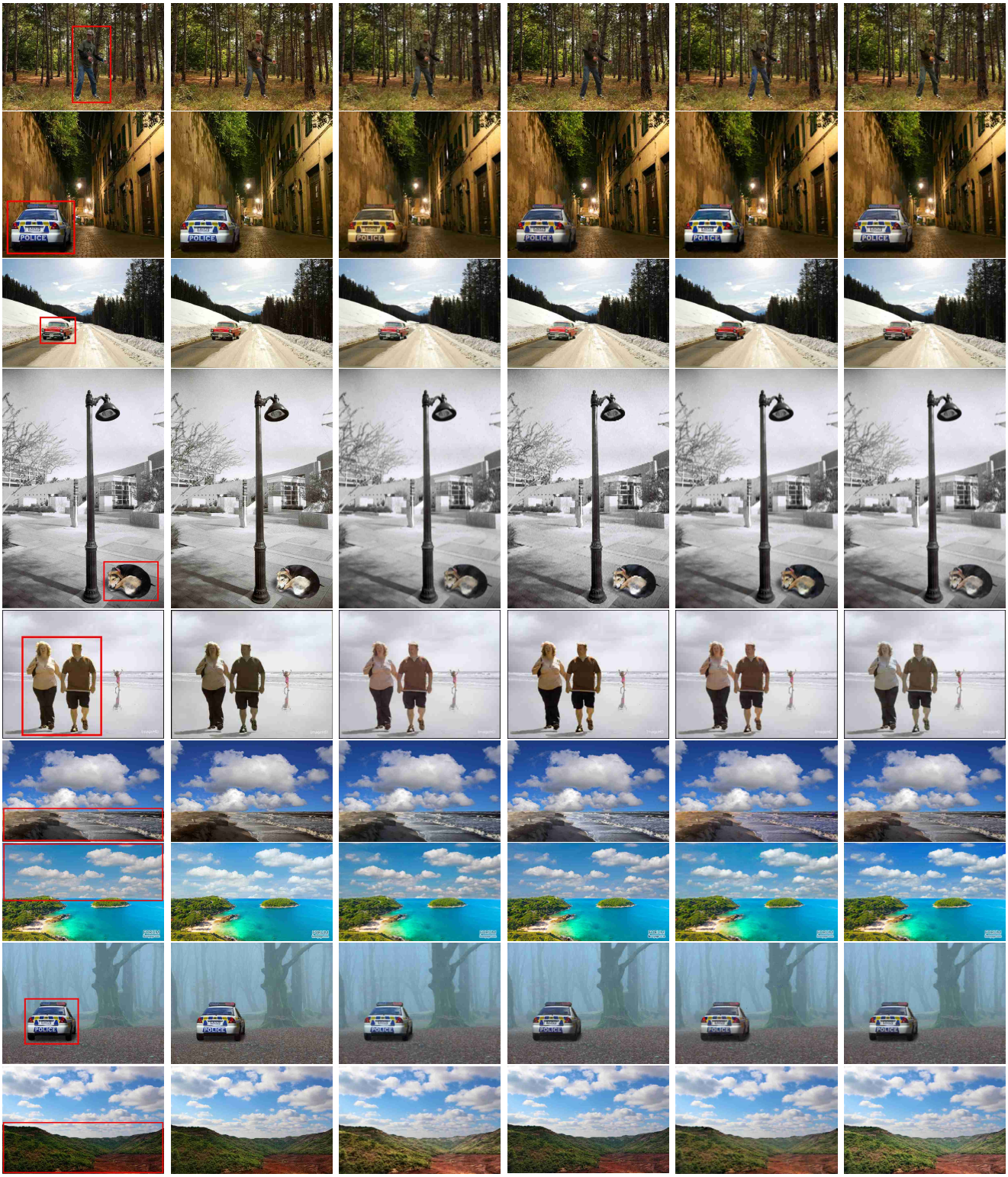

Figure 6. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

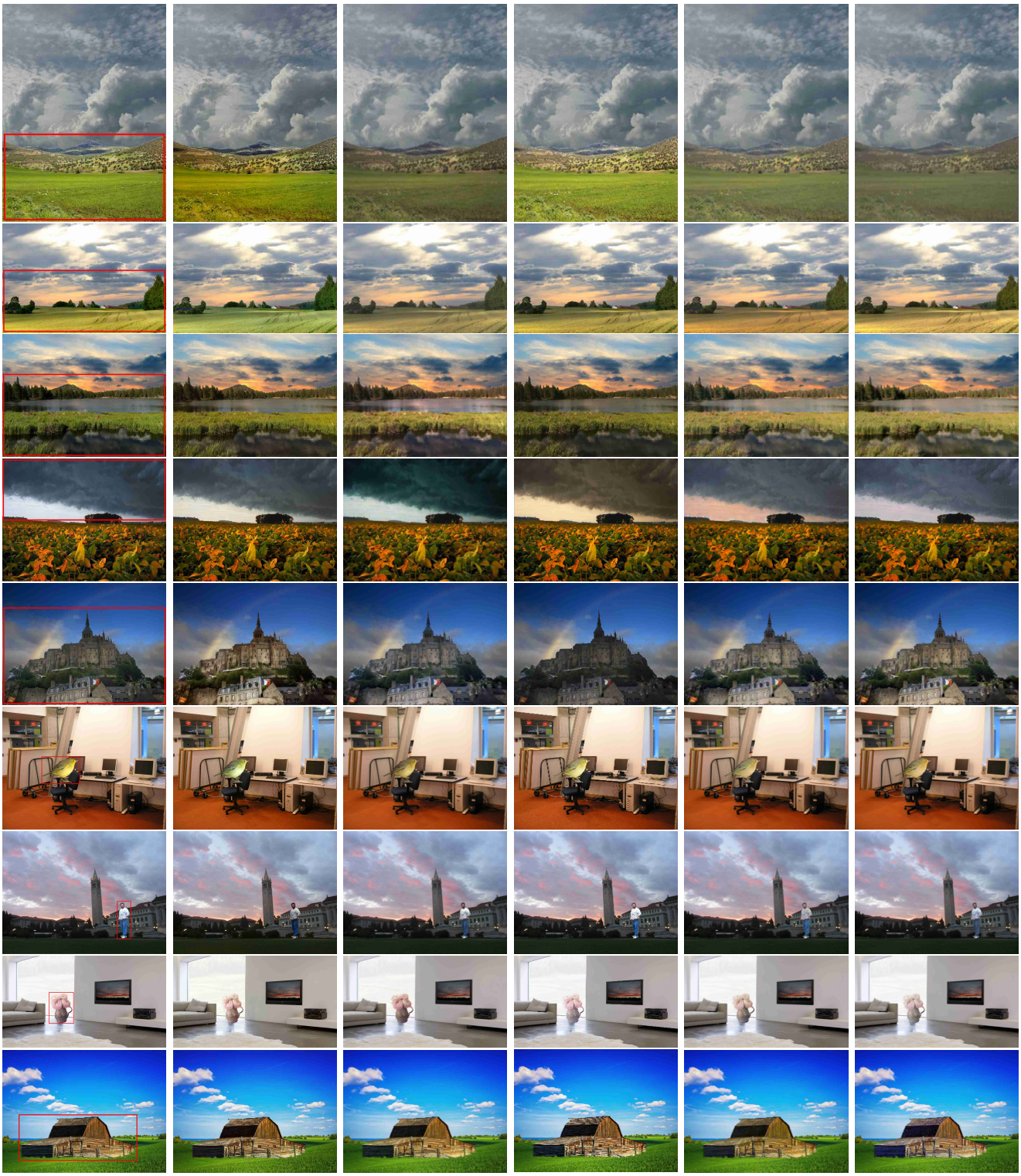

Figure 7. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

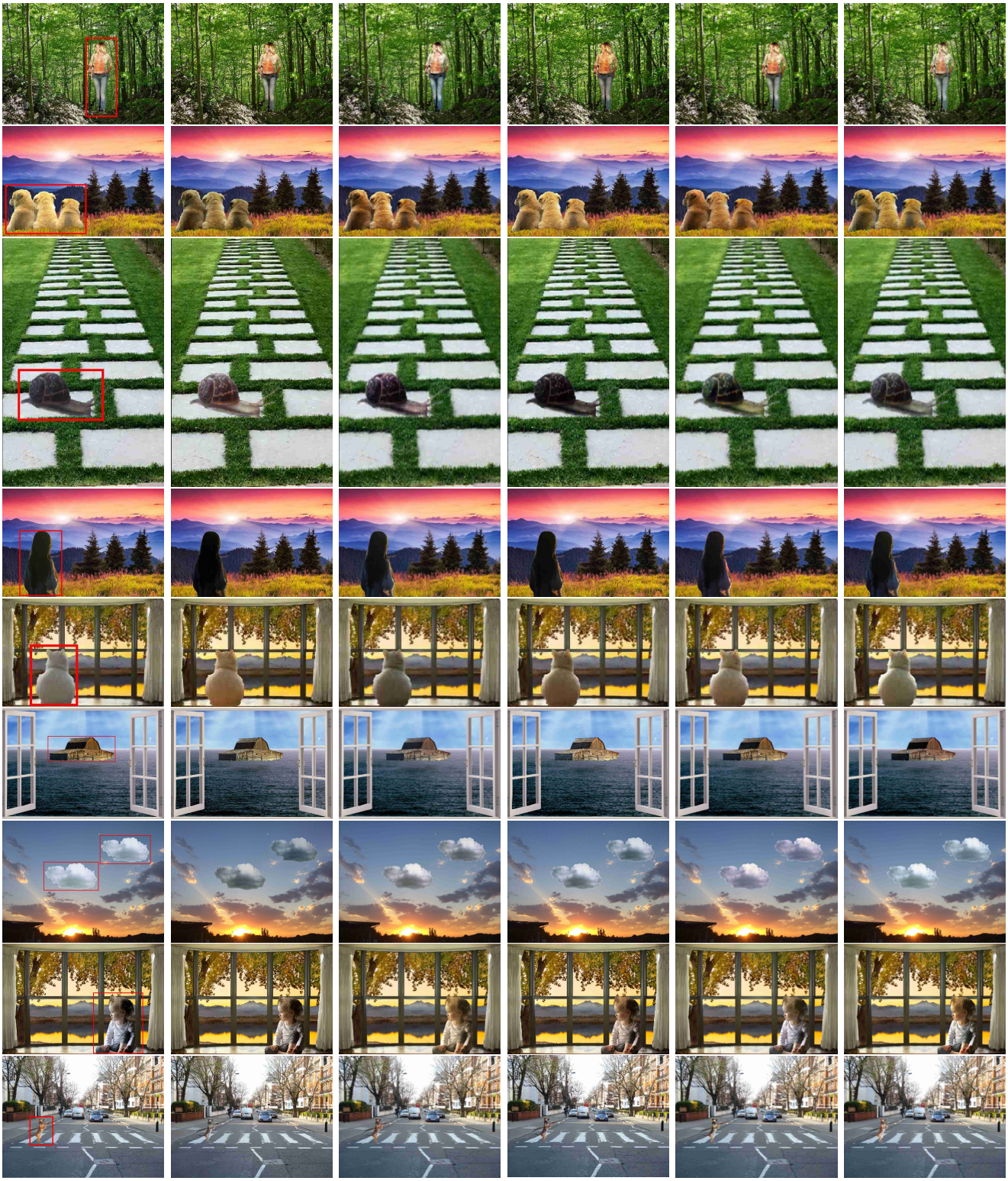

Figure 8. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

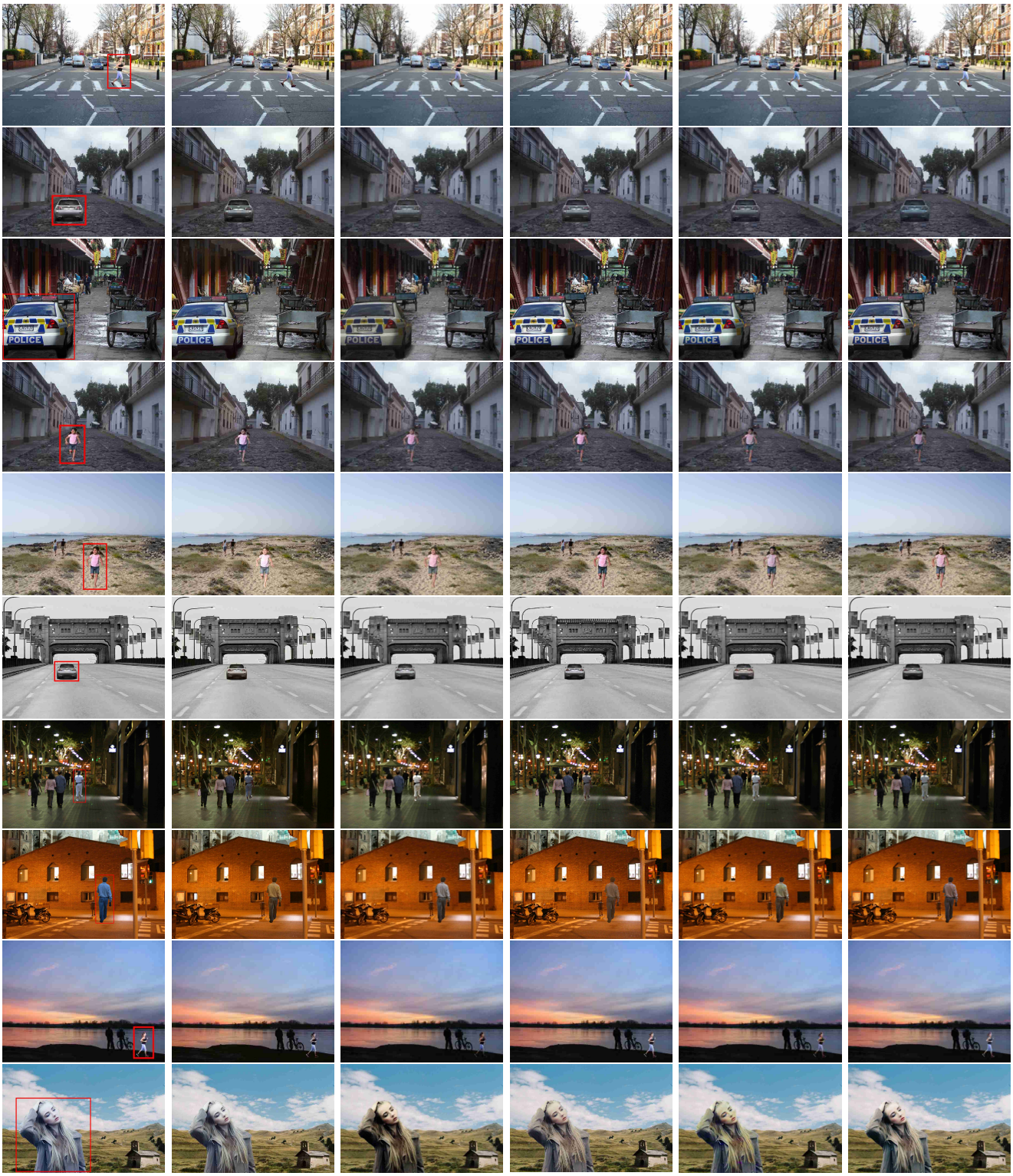

Figure 9. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

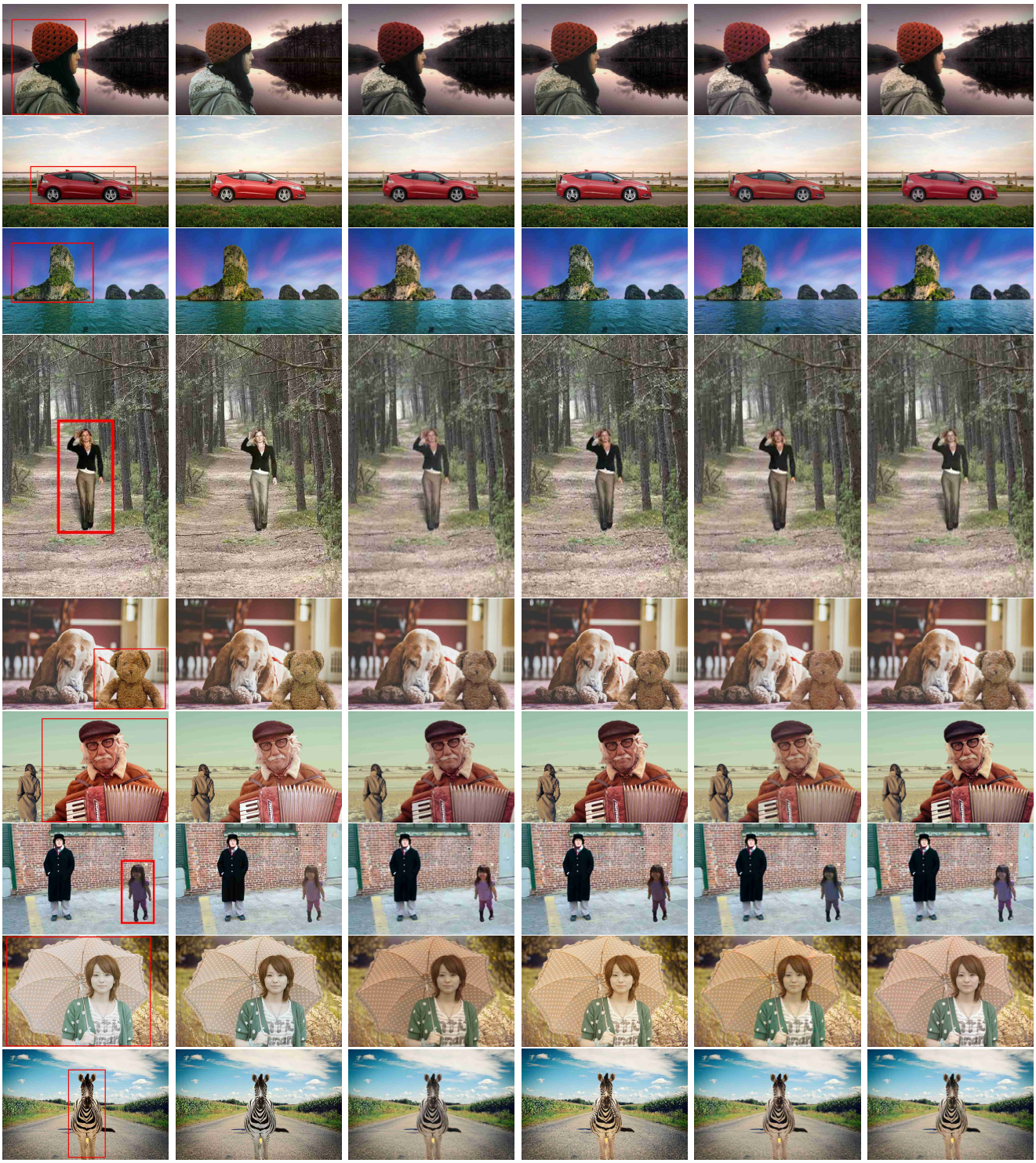

Figure 10. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

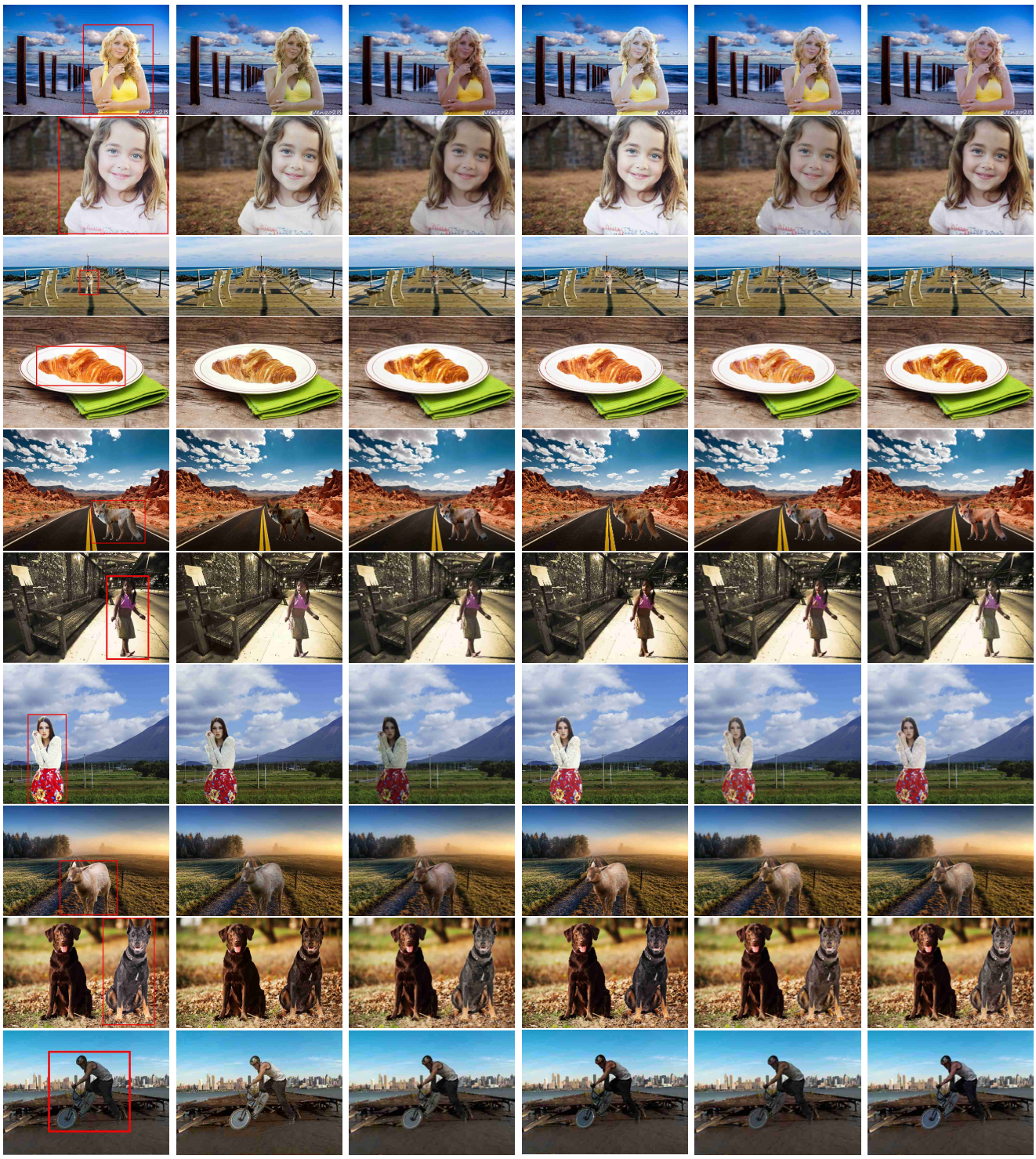

Figure 11. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.

Figure 12. Visual comparison results on real composite images released by [5]. Red boxes in composite images mark foreground.