# Supplementary Material for DoveNet: Deep Image Harmonization via Domain Verification

Wenyan Cong[1], Jianfu Zhang[1], Li Niu[1]*, Liu Liu[1], Zhixin Ling[1], Weiyuan Li[2], Liqing Zhang[1]
[1] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University [2] East China Normal University
[1]{plcwyam17320,c.sis,ustcnewly,Shirlley,1069066484}@sjtu.edu.cn
[2]10162100162@stu.ecnu.edu.cn [1]zhang-lq@cs.sjtu.edu.cn

In this Supplementary file, we will introduce the attention block used in our proposed DoveNet in Section 1, analyze our constructed iHarmony4 dataset *w.r.t.* foreground ratio, color transfer method, and semantic category in Section 2, 3, 4. Besides, We will show samples of manually filtered images and final images in our dataset in Section 5, 6. Finally, we will exhibit the results of different methods on all 99 real composite images in Section 7.

## 1. Details of Attention Block

U-Net utilizes skip-connections to leverage information from encoder for decoder. Inspired by [11], we leverage attention blocks to enhance U-Net. The detailed structure is depicted in Fig. 1 (b). We concatenate encoder and decoder features, based on which full attention maps [13] (integration of spatial attention and channel attention) are learnt for encoder feature and decoder feature separately. Specifically, to obtain encoder attention map and decoder attention map, we apply $1 \times 1$ convolution layer on the concatenation of encoder and decoder features, followed by Sigmoid activation. After that, we multiply encoder (*resp.*, decoder) attention map to the encoder feature (*resp.*, decoder) feature element-wisely. We expect the encoder attention map to pay more attention to the background of encoder feature, because the foreground of encoder feature may not be fully harmonized yet. Finally, we concatenate the attended encoder feature and decoder feature as the output of attention block.

Note that in [11], attention maps are learnt for foreground an background separately with explicit mask control. We argue that since the mask is included in the input to the generator, the attention block can utilize the mask information automatically. Compared with [11], our attention block is simple yet effective.

*Corresponding author.

## 2. Analyses of Foreground Ratio

Our iHarmony4 dataset has a wide range of foreground ratios, *i.e.*, the area of foreground to be harmonized over the area of the whole image. We report the distribution of foreground ratios on our four sub-datasets (*i.e.*, HCOCO, HAdobe5k, HFlickr, Hday2night) and the whole dataset in Figure 2a, from which we can observe that the foreground ratios are mainly distributed in the range $[0\%, 70\%]$ and have a long-tail distribution. We also observe that the foreground ratios on four sub-datasets are quite different, which is caused by different acquisition process of four sub-datasets. Next, we will analyze each sub-dataset separately.

For COCO dataset [6] with provided segmentation masks, we naturally leverage its segmentation annotation and ensure that each foreground occupies larger than 1% and smaller than 80% of the whole image. Since we apply color transfer methods to generate synthesized composites and filter out unqualified ones, synthesized composites with larger foreground regions are more prone to be removed due to more likely low quality. Thus, our HCOCO sub-dataset has relatively small foreground regions.

For the other three datasets (*i.e.*, Adobe5k[1], Flickr, and day2night[4]), there are no segmentation masks, so we have to manually select and segment one or more foreground regions in an image. The images in Adobe5k are mostly taken by professional photographers with a prominent subject in the image, so it is more likely to select a relatively large foreground. For the crawled Flickr images, we remove those images without obvious foreground and those with blurred background, and thus the remaining images are similar to those in Adobe5k. Therefore, our HAdobe5k and HFlickr sub-datasets have relatively large foreground regions.

For day2night dataset, as discussed in Section 3.1 in the main paper, we need to satisfy more constraints to select a reasonable foreground. Specifically, moving or deformable objects (*e.g.*, person, animal, car) or objects with essential changes are not suitable to be chosen as foreground regions.

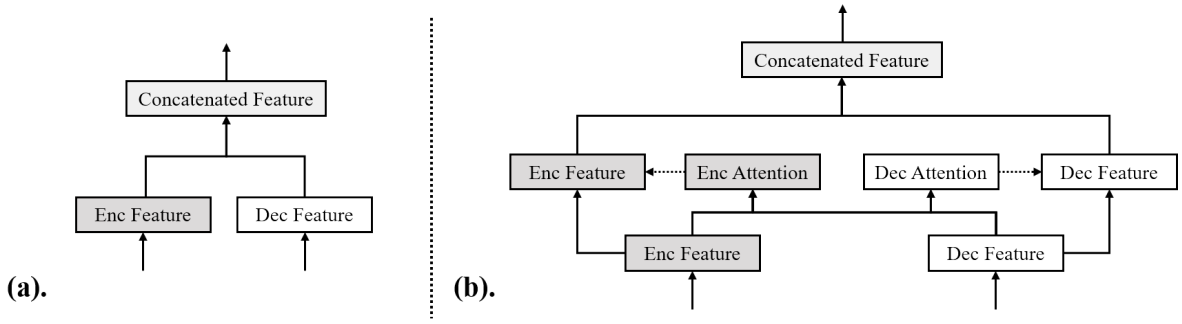
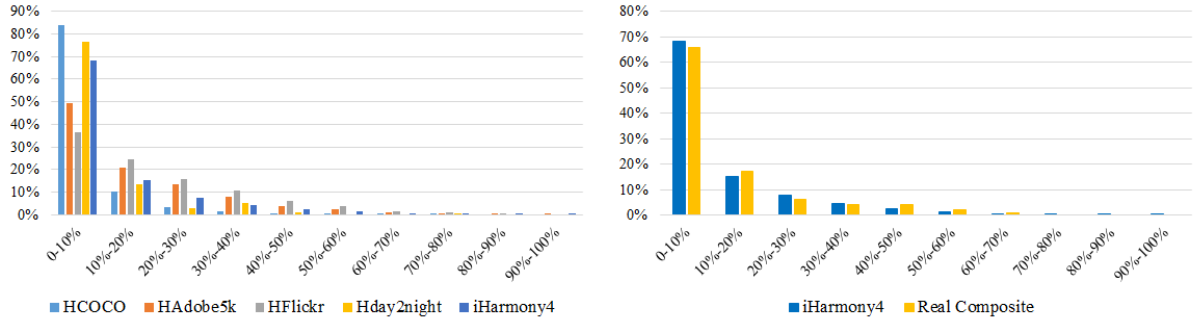



Figure 1: Illustration of our proposed attention module. (a). original U-Net structure without attention module; (b). U-Net with our attention module.




(a) Comparison between four sub-datasets

(b) Comparison between iHarmony4 and real composites

Figure 2: The distributions of foreground ratios. (a) Comparison between four sub-datasets shows that HCOCO and Hday2night have more images with small foreground regions while HAdobe5k and HFlickr have more images with large foreground regions. (b) Comparison between iHarmony and real composites in [12, 9] shows that the distribution of foreground ratios of iHarmony4 dataset is close to that of real composite images.

Hence, we prefer to select static objects that remain consistent across multiple capture conditions, resulting in relatively small foreground regions.

Actually, the composite images in real-world applications also have a wide range of foreground ratios. We show the distribution of 99 real composite images (48 images from Xue *et al.* [12] and 51 images from Tsai *et al.* [9]) and compare with our whole dataset in Figure 2b. From Figure 2b, it can be seen that the distribution of foreground ratios in our whole dataset is close to that of real composite images, which means that the foreground ratios of our constructed dataset are reasonable for real-world image harmonization tasks.

## 3. Analyses of Color Transfer Methods

When constructing HCOCO and HFlickr sub-datasets, we apply color transfer methods to adjust the foreground to make it incompatible with background.

Existing color transfer methods can be categorized into parametric and non-parametric methods. Parametric methods assume the parametric format of the color mapping function. Assumed parametric format is compact but may be unreliable. Instead, non-parametric methods have no parametric format of the color transfer function, and most of them directly record the mapping of the full range of color/intensity levels using a look-up table, which is usually computed from the 2D joint histogram of image feature correspondences.

From the perspective of color space, existing color transfer methods can be applied to either correlated color space or decorrelated color space. Typically, images are encoded using RGB color space, in which three channels (R, G, B) are highly correlated. This implies that if we want to change the appearance of a pixel in a coherent way, we must modify all three color channels. That complicates any color modification process and may have unpredictable results [8]. However, by shifting, scaling, and rotating the axes in RGB color space, we can construct a new color space (*e.g.*, CIELAB, Yuv, HSV). If different channels in

this new color space are near-independent or independent, image processing can be done in each channel independently. Nonetheless, decorrelated color space may fail to capture some subtleties including local color information and interrelation [10].

Based on the abovementioned parametric or non-parametric methods as well as correlated or decorrelated color space, existing color transfer methods can be grouped into four quadrants, and each quadrant has its own advantage and drawback. To enrich the diversity of synthesized composite images, we choose one representative method in each quadrant.

**Parametric method in decorrelated color space:** Based on global color distribution of two images, Reinhard *et al.* [8] proposed a linear transformation in decorrelated color space $L\alpha\beta$, by transferring mean and standard deviation between each channel of two images:

$$I_o = \frac{\sigma_r}{\sigma_t}(I_t - \mu_t) + \mu_r,$$

where ($\mu_t$, $\sigma_t$) and ($\mu_r$, $\sigma_r$) are the mean and standard deviation of the target and reference images in $L\alpha\beta$ space, $I_t$ and $I_o$ are the color distribution of the target and output images.

**Parametric method in correlated color space:** Xiao *et al.* [10] extended [8] by transferring mean and covariance between images in correlated RGB space. It replaces the rotation to $L\alpha\beta$ with the axes defined by the principal components of each image, leading to a series of matrix transformation:

$$I_o = T_r R_r S_r S_t R_t T_t I_t,$$

in which $T$, $R$ and $S$ denote the matrices of translation, rotation, and scaling derived from the target and reference images accordingly.

**Non-parametric method in decorrelated color space:** Fecker *et al.* [2] proposed to use cumulative histogram mapping in decorrelated color space $YC_bC_r$. They used nearest neighbor mapping scheme to set the corresponding color level of the source image to each level of the target. In this way, the shape of the target histogram can be matched to the reference histogram, and thus the transferred image has the same color as the reference.

**Non-parametric method in correlated color space:** By treating 3D color distribution as a whole, Pitié *et al.* [3] proposed iterative color distribution transfer by matching 3D distribution through an iterative match of 1D projections. Iterative color distribution transfer can increase the graininess of the original image, especially if the color dynamics of two images are very different. So Pitié *et al.* [7] proposed a second stage to reduce the grain artifact through an efficient post-processing algorithm.

In summary, we adopt four color transfer methods: global color transfer in $L\alpha\beta$ space [8], global color transfer in RGB space [10], cumulative histogram matching [2], and iterative color distribution transfer [7]. When generating synthetic composite images for HCOCO and HFlickr, we randomly choose one color transfer method from the above. By taking HCOCO dataset as an example, after automatic and manual filtering, the number of remaining composite images obtained using method [8] [10] [7] [2] are 9581, 8119, 17009, and 8119 respectively, which indicates that iterative color distribution transfer [7] is better at producing realistic and reasonable composites.

We split the test set of HCOCO into four subsets according to four color transfer methods, and report the results on four subsets in Table 1. We can see that the statistics (MSE, PSNR) of input composites obtained by different color transfer methods are considerably different, which shows the necessity of applying multiple color transfer methods to enrich the diversity of synthesized composite images. Moreover, our proposed method achieves the best results on four subsets, which demonstrates the robustness of our proposed method.

## 4. Analyses of Semantic Category

COCO dataset is associated with semantic segmentation masks, so we can easily obtain the category labels of foreground regions in our HCOCO sub-dataset. To explore the difference between different categories, we report fMSE (foreground MSE) of input composites and our harmonized results on different categories in Table 2, in which Table 2a (*resp.*, 2b) shows the hard (*resp.*, easy) categories. We define easy or hard categories based on the relative improvement of our method compared with input composite.

From Table 2, we find that for categories with small intra-category variance (*e.g.*, mouse, keyboard), fMSE could be improved significantly, while for categories with large intra-category variance (*e.g.*, person), the improvement is relatively small.

## 5. Examples of Manual Filtering

After generating composite images, two steps of automatic filtering and an additional manually filtering are applied to HCOCO, HFlickr, and Hday2night sub-datasets to eliminate low-quality synthesized composites. In the step of manual filtering, we pay close attention to the cases that are harmful to image harmonization task.

For HCOCO sub-dataset, foreground regions are obtained based on the semantic segmentation annotation provided in COCO dataset. However, some foreground regions are highly occluded and not very meaningful for image harmonization task. For example, in Figure 3a, an occluded person with only a hand or a shoulder is annotated as "person" in COCO dataset, but it is not very meaningful to harmonize a highly occluded person.

| Color Transfer Methods | [8] | | [10] | | [7] | | [2] | |
|---|---|---|---|---|---|---|---|---|
| Evaluation metric | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ |
| Input composite | 75.81 | 33.39 | 66.84 | 33.90 | 73.83 | 33.73 | 57.55 | 34.92 |
| Lalonde and Efros [5] | 127.82 | 30.83 | 107.44 | 31.26 | 110.88 | 31.04 | 92.02 | 31.51 |
| Xue *et al.* [12] | 90.31 | 32.74 | 69.97 | 33.78 | 74.74 | 33.26 | 72.78 | 33.62 |
| Zhu *et al.* [14] | 84.31 | 32.62 | 77.93 | 33.06 | 85.36 | 32.78 | 67.81 | 33.89 |
| DIH [9] | 57.77 | 34.32 | 51.02 | 34.70 | 53.90 | 34.63 | 42.91 | 35.21 |
| $S^2$AM [11] | 41.89 | 35.29 | 37.33 | 35.81 | 47.00 | 34.99 | 33.93 | 36.13 |
| Ours | **38.21** | **35.62** | **34.42** | **36.17** | **40.92** | **35.38** | **30.51** | **36.47** |

Table 1: MSE and PSNR on four sub test sets of HCOCO corresponding to different color transfer methods. The best results are denoted in boldface.


(a) Composite Images with highly occluded foregrounds.

(b) Composite Images with dramatically changed hues.

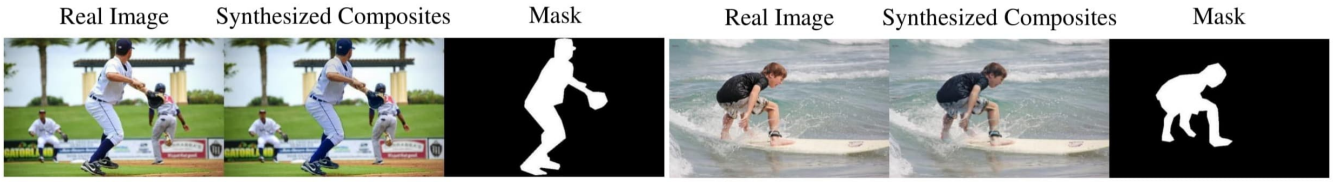
(c) Composite Images with essential object changes.

Figure 3: Sample composite images that are discarded during manual filtering. From top to bottom, we show undesirable examples with highly occluded foregrounds, dramatically changed hues, and essential changes of the objects themselves that are not caused by capture condition.

Besides, for HCOCO and HFlickr sub-datasets, color transfer is applied to different foreground objects of the same category between reference image and target image, so the hue of foreground object may be dramatically changed, especially for the categories with large intra-category variance like "person". For example, in Figure 3b, the shirt color is changed from red in the real image to yellow in the composite image. It does not make sense to harmonize a yellow shirt into a red shirt, so we remove such images from our dataset.

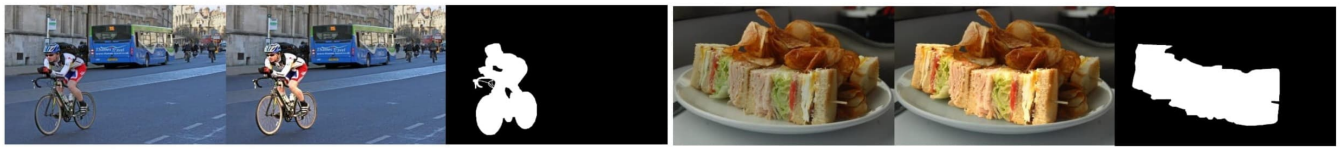
Furthermore, for Hday2night sub-dataset, when overlaying the foreground from reference image on the target image, some essential changes which are not caused by capture condition may happen to the foreground object. For example, in Figure 3c, the lights of Eiffel Tower are switched on (see left subfigure) and the snow appears on the roof (see right subfigure) in the composite image. We argue that these changes do not belong to the scope of image harmonization task, and thus filter out these images from our dataset.

By filtering out the unqualified images in the above cases, we ensure the high quality of our iHarmony4 dataset to the utmost.




(a) Example images of HCOCO sub-dataset

(b) Example images of HAdobe5k sub-dataset



(c) Example images of HFlickr sub-dataset

(d) Example images of Hday2night sub-dataset

Figure 4: Example images of our contributed dataset iHarmony4. From top to bottom, we show examples from our HCOCO, HAdobe5k, HFlickr, and Hday2night sub-datasets. From left to right, we show the real image, the synthesized composite image, and the foreground mask for each example.

| | baseball glove | snow-board | kite | person | surf-board |
|---|---|---|---|---|---|
| Input | 1402.62 | 570.07 | 2391.73 | 428.59 | 1409.41 |
| DoveNet | 1301.21 | 523.18 | 1857.24 | 321.48 | 972.09 |

(a) Categories with slightest fMSE improvement.

| | mouse | keyboard | oven | pizza | zebra |
|---|---|---|---|---|---|
| Input | 1530.63 | 1624.38 | 1257.21 | 1316.40 | 959.08 |
| DoveNet | 423.84 | 521.24 | 481.37 | 532.38 | 388.66 |

(b) Categories with largest fMSE improvement.

Table 2: fMSE improvement of different categories of HCOCO sub-dataset.

## 6. Examples of Our iHarmony4 Dataset

In Figure 4, we show some examples of our four sub-datasets with each row corresponding to one sub-dataset. For each example, we show the original real image, synthesized composite image, and foreground mask.

## 7. Results on Real Composite Images

In Figure 5 to 15, we present all results of 99 real composite images used in our user study (see Section 5.6 in the main paper), including 48 images from Xue *et al.* [12] and 51 images from Tsai *et al.* [9]. We compare the real composite images with harmonization results generated by our proposed method and other existing methods, including Lalonde and Efros [5], Xue *et al.* [12], Zhu *et al.* [14], DIH [9], and $S^2$AM [11]. Based on Figure 5 to 15, we can see that our proposed method could generally produce satisfactory harmonized images across various scenes and objects.

## References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 1

[2] U. Fecker, M. Barkowsky, and A. Kaup. Histogram-based prefiltering for luminance and chrominance compensation of multiview video. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1258–1267, 2008. 3, 4

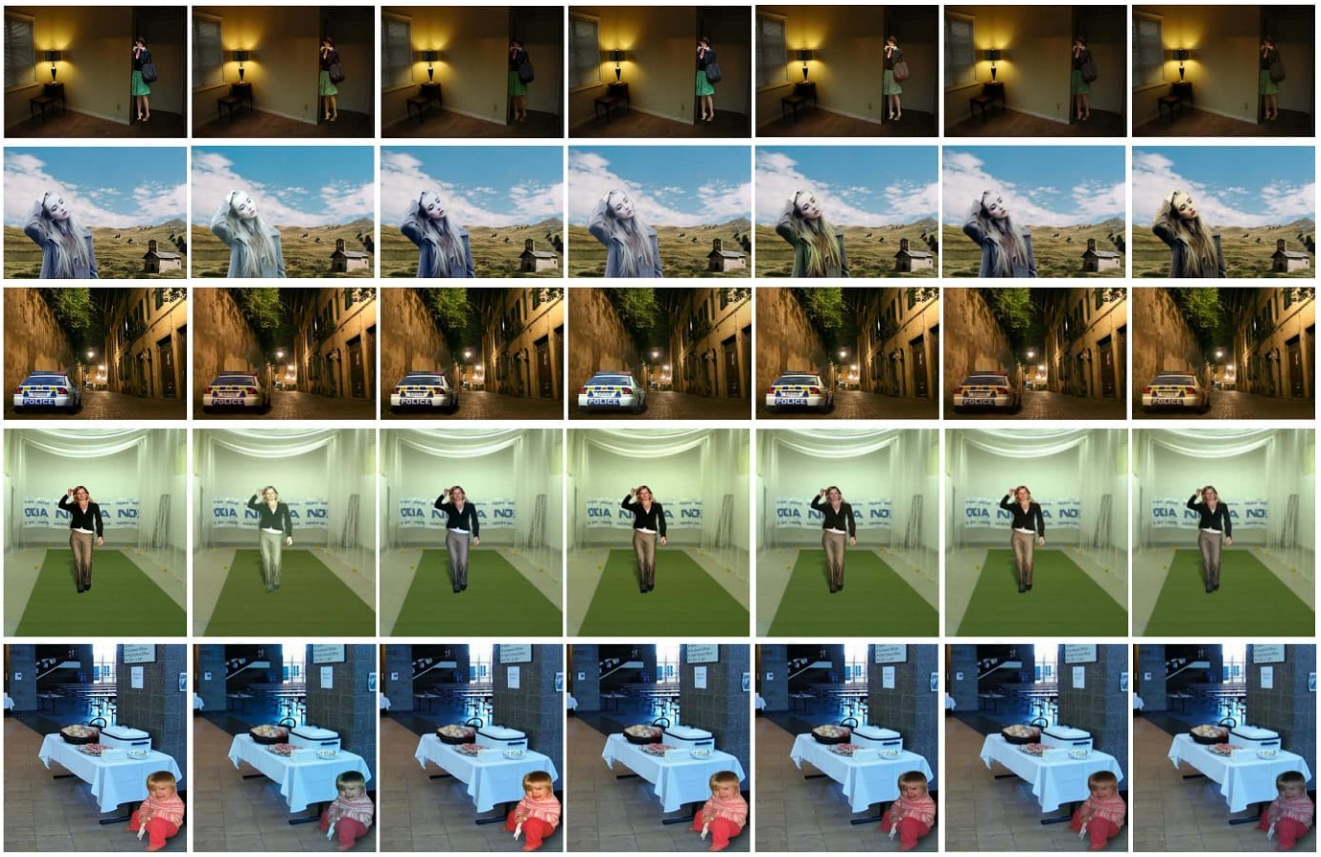



Figure 5: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.


[3] Anil C. Kokaram and Rozenn Dahyot. N-dimensional probability density function transfer and its application to colour transfer. In *ICCV*, 2005. 3

[4] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 33(4), 2014. 1

[5] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 4, 5

[6] 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 *ECCV*, 2014. 1

[7] François Pitié, Anil C. Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123 – 137, 2007. 3, 4

[8] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 2, 3, 4

[9] Yi-Hsuan Tsai, Xiaohui Shen, Zhe L. Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. *CVPR*, 2017. 2, 4, 5

[10] Xuezhong Xiao and Lizhuang Ma. Color transfer in correlated color space. In *VRCIA*, 2006. 3, 4

[11] Cun Xiaodong and Pun Chi-Man. Improving the harmony of the composite image by spatial-separated attention module. *arXiv preprint arXiv:1907.06406*, 2019. 1, 4, 5

[12] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics*, 31(4):84, 2012. 2, 4, 5

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1

[14] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 4, 5

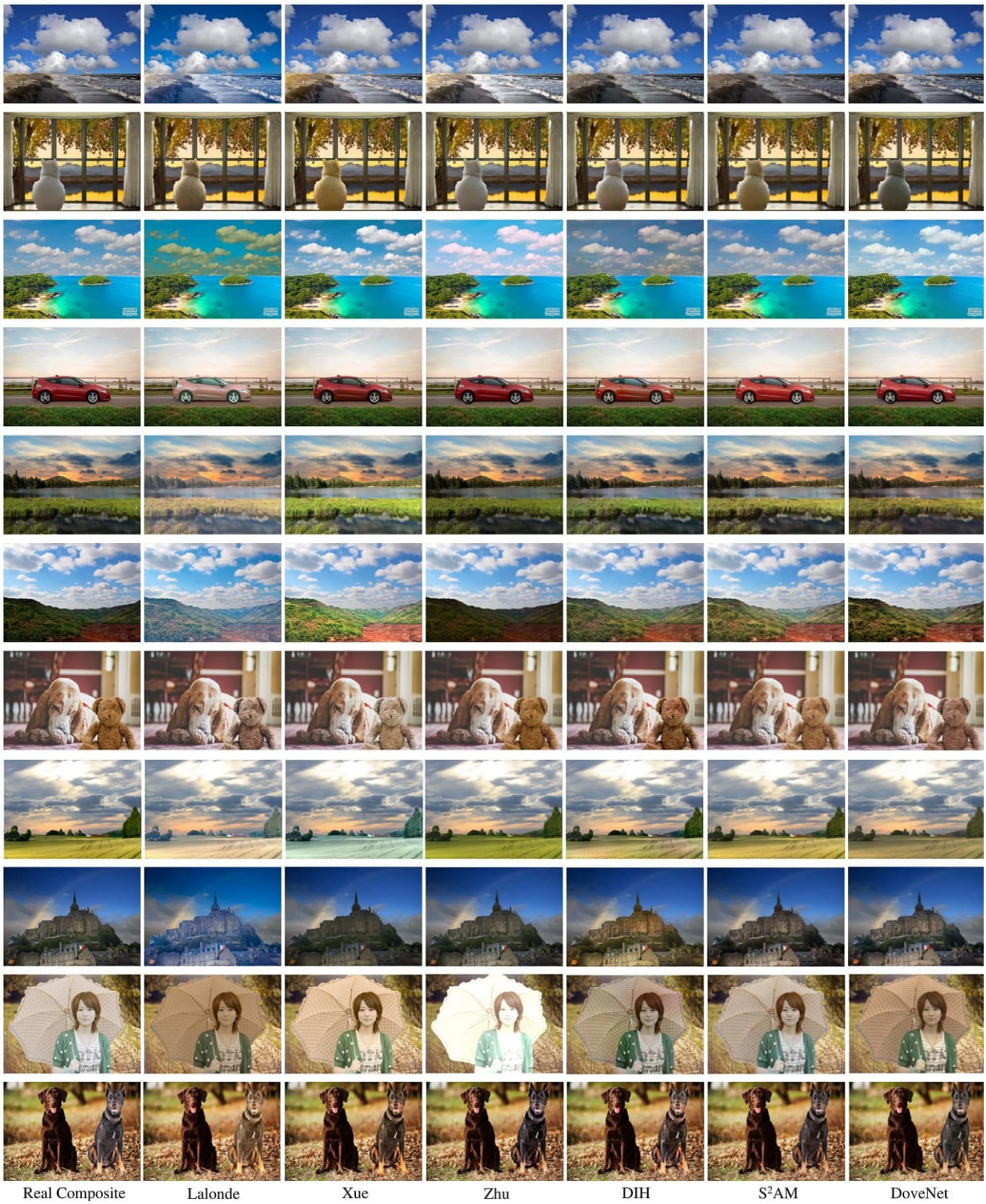



Figure 6: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

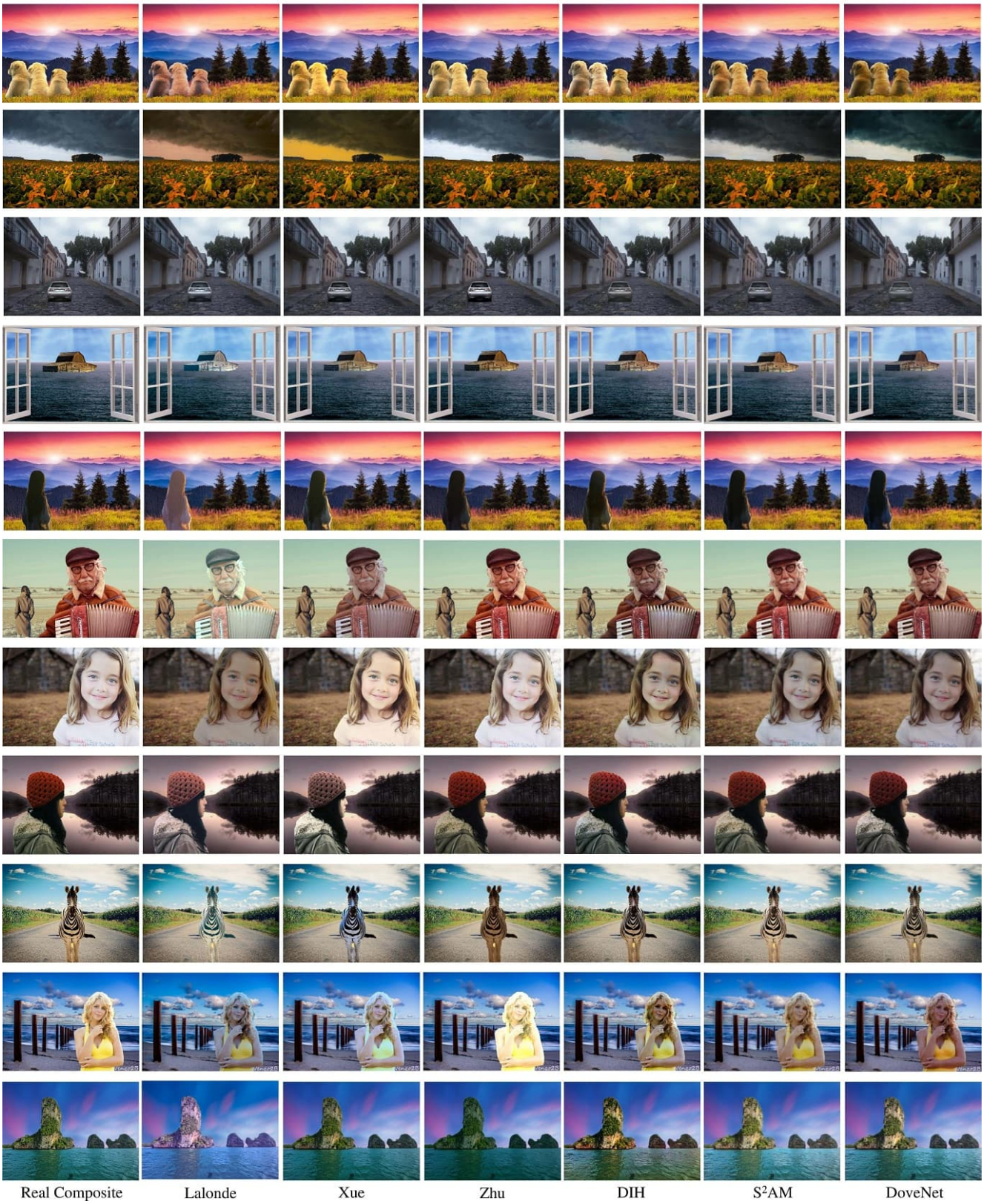



Figure 7: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

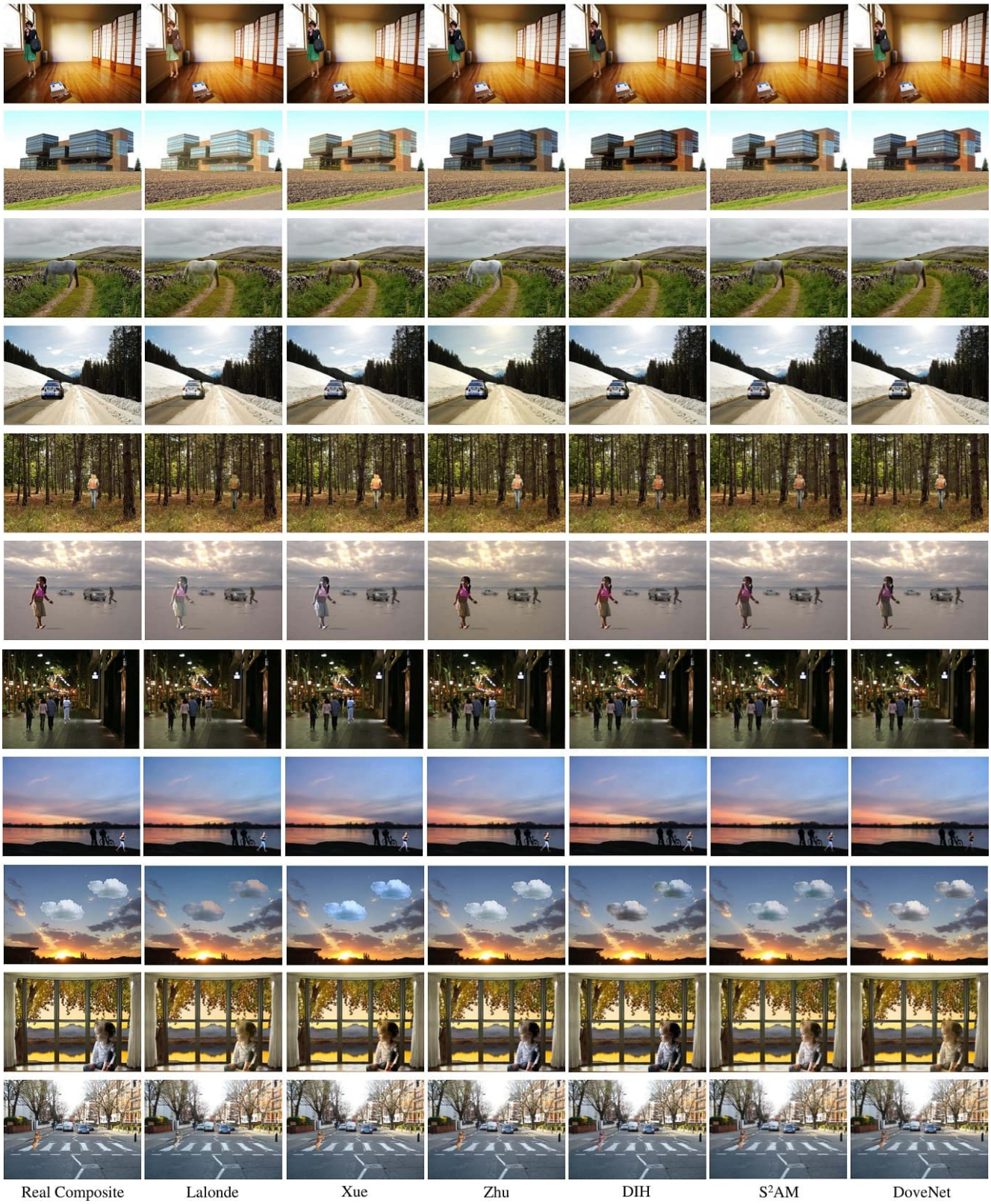



Figure 8: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

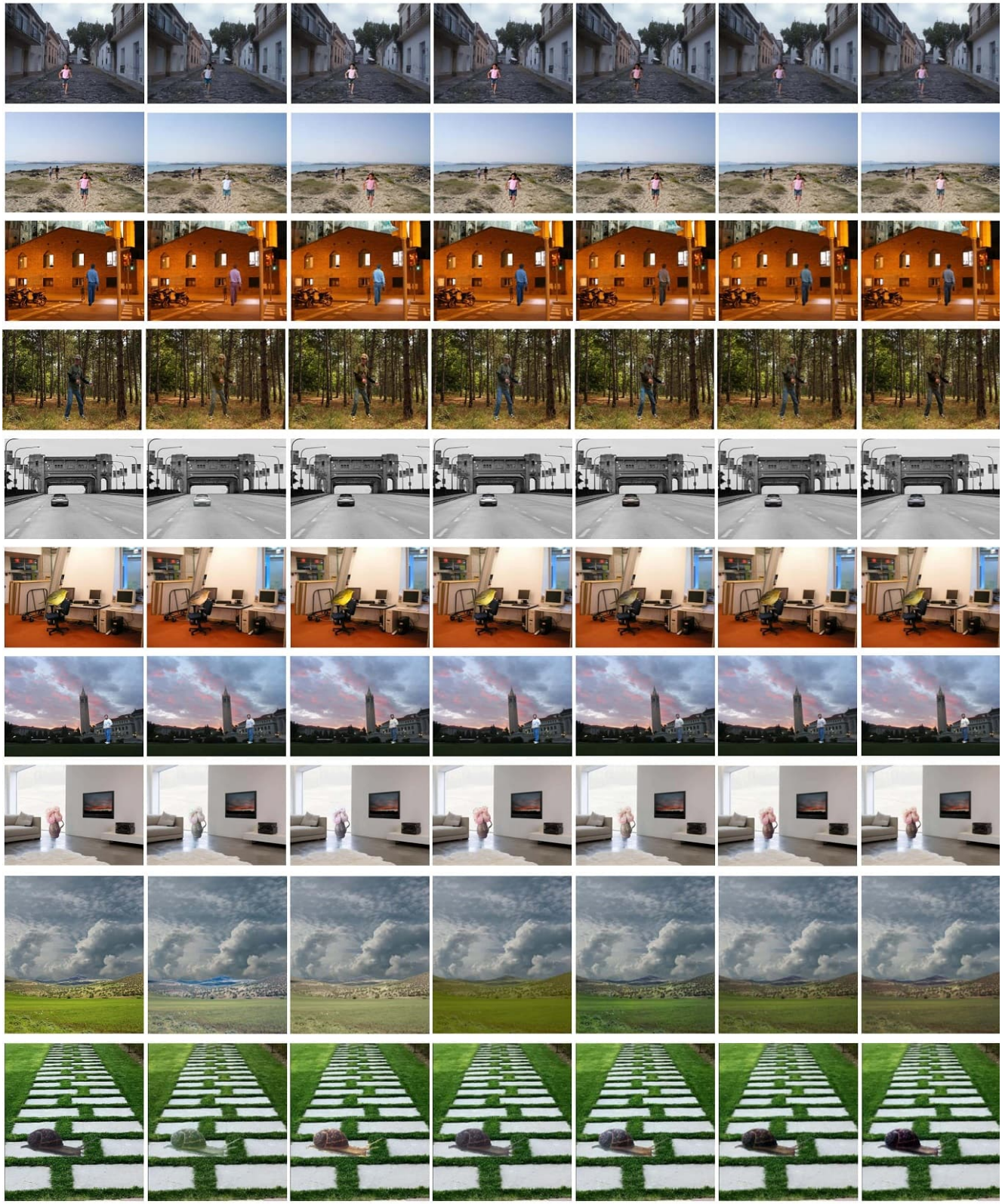



Figure 9: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

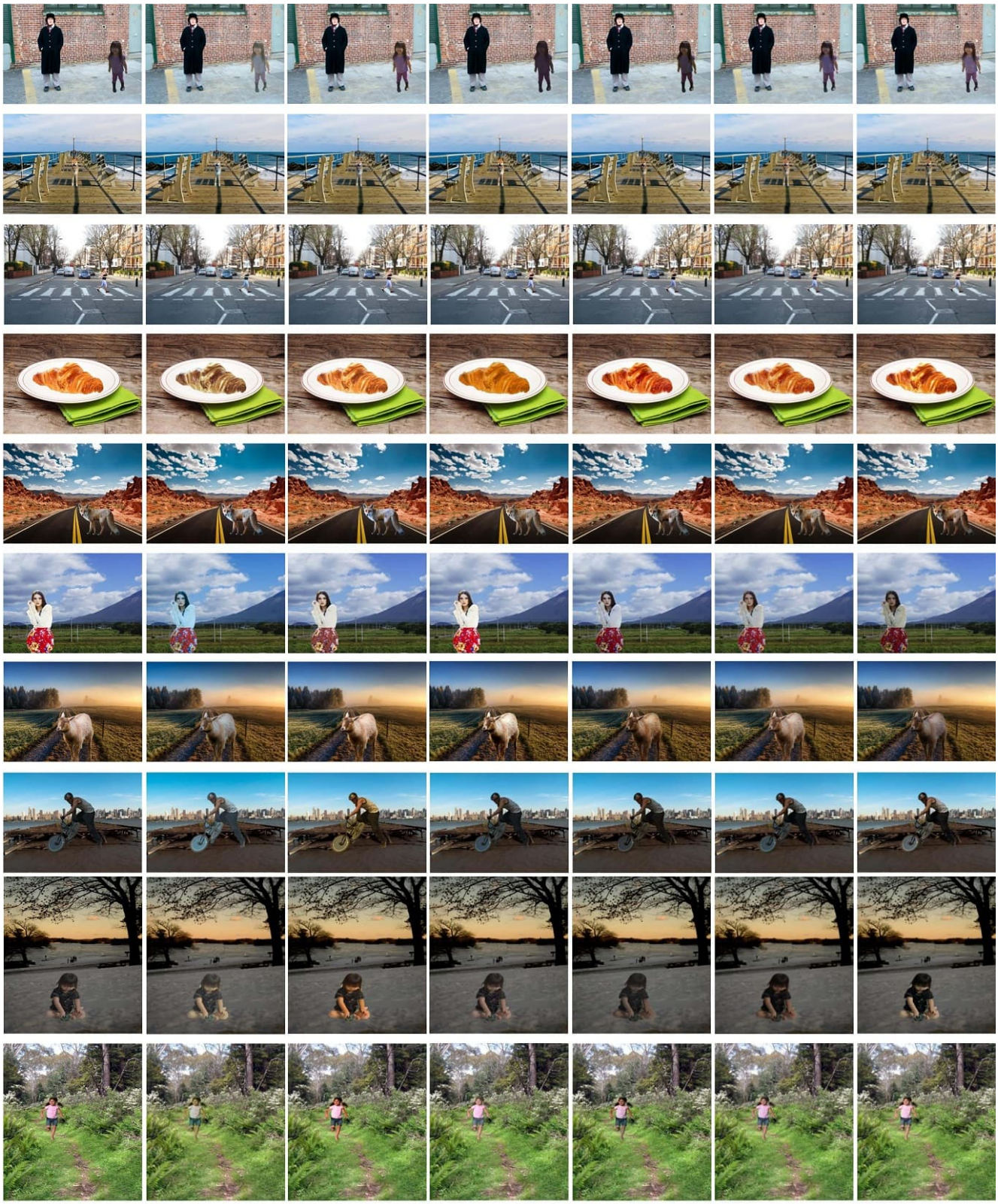



Figure 10: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

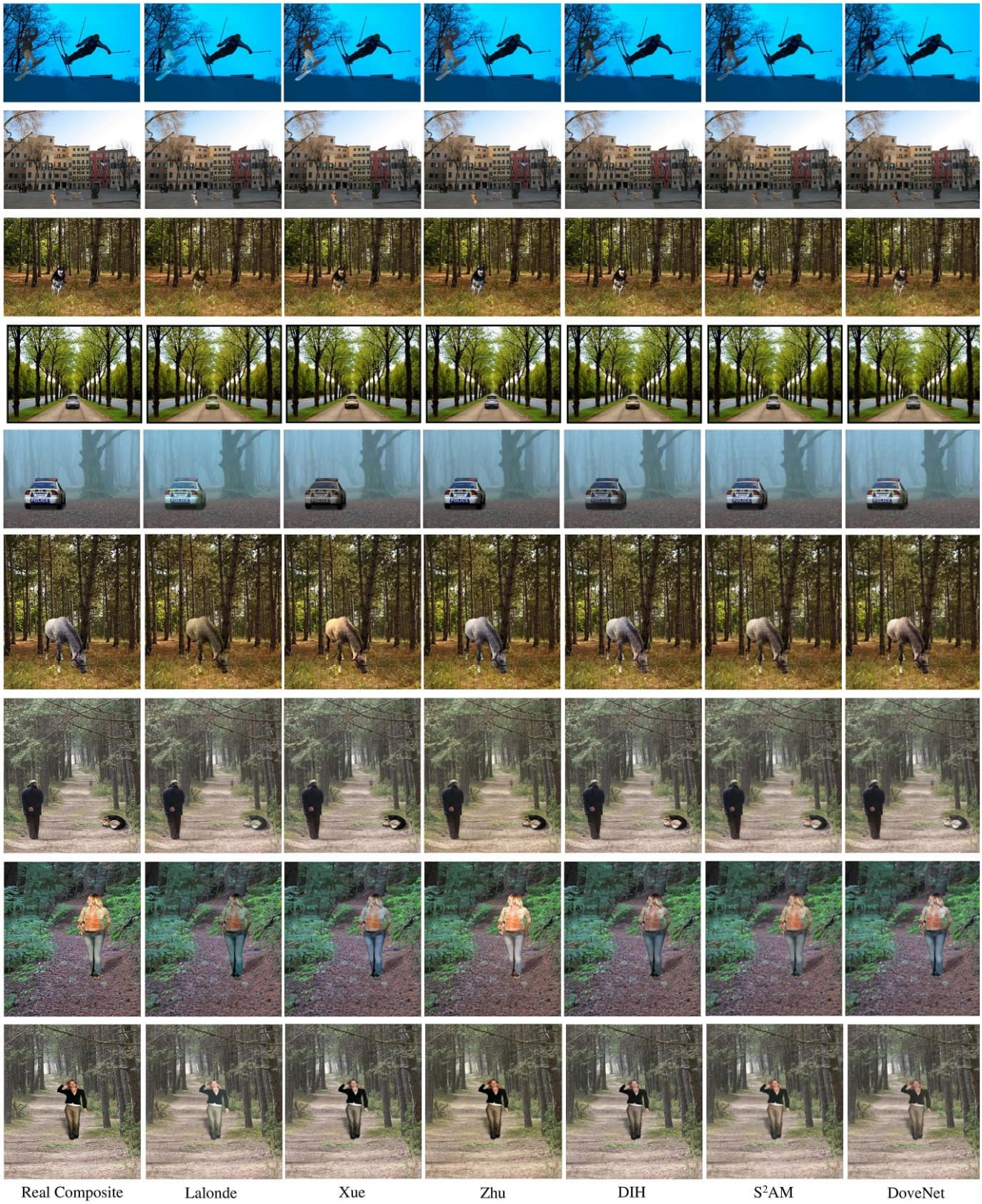



Figure 11: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

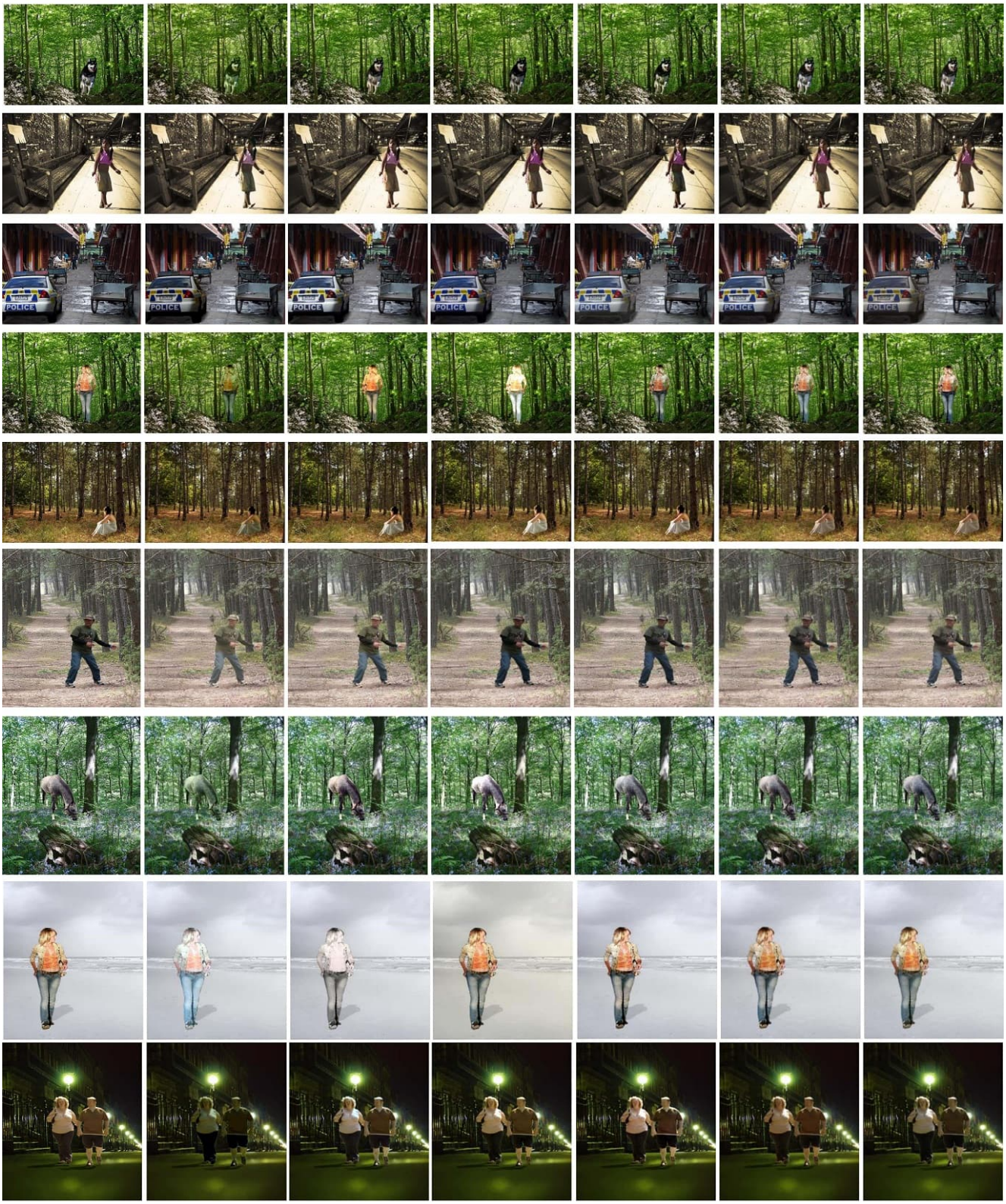



Figure 12: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

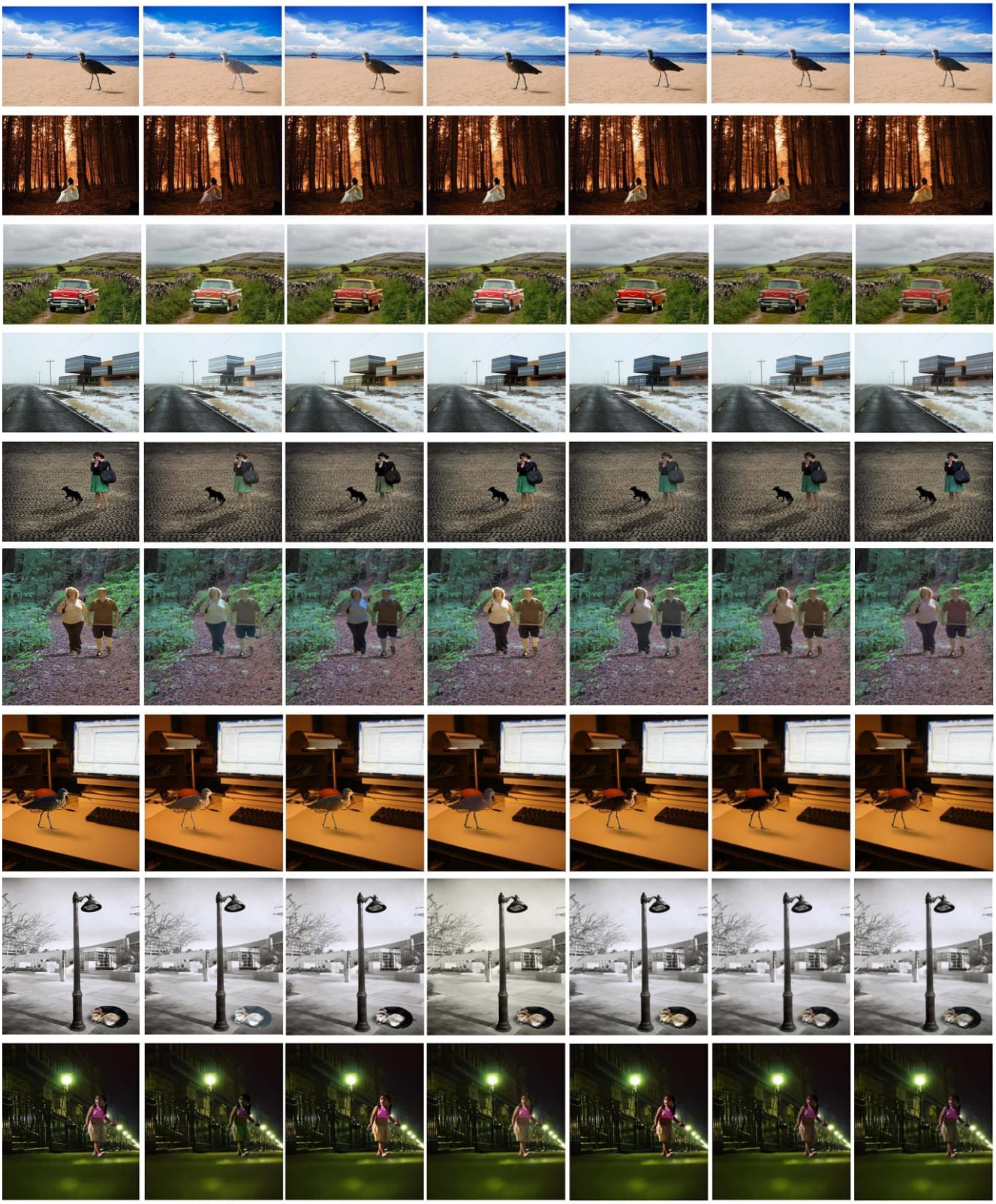



Figure 13: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.

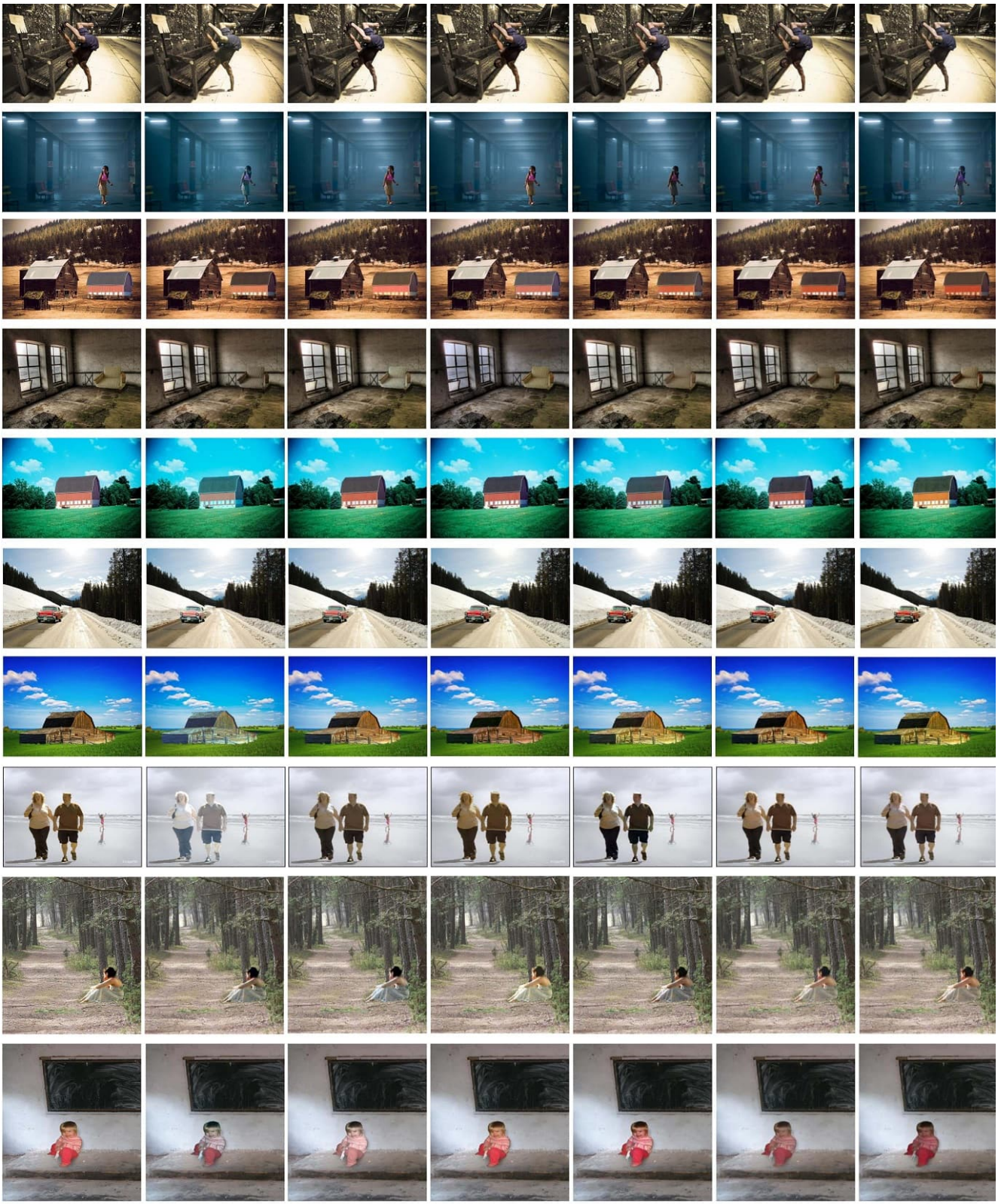



Figure 14: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.


Figure 15: Results on real composite images, including the input composite, five state-of-the-art methods, and our proposed DoveNet.