Supplementary Material for DeFLOCNet

In this supplementary material, we first introduce the training details of the proposed model. Then, more visual comparisons are also shown on the Places2 natural image dataset and the CelebA-HQ face dataset.

1. Details of network training and inference

Training. The pseudo code of the training process is shown in Algorithm 1. We denote generator as G, discriminator as D. Our inputs are from two aspects. First, we set $I_{in} = I_{gt} \odot (1 - M)$ as the input image, where I_{gt} is an original image, M is the binary mask indicating hole regions and these binary masks are from PConv [17]. We concatenate $[I_{in}, N, M]$ as the inputs to the encoder-decoder, where N is the random noise image, and N is sampled from the Gaussian distribution. Second, we concatenate $[S, I_n]$ as the inputs to the sketch line generation branch where S is the sketch line image. In the training process, we extract the contour lines of I_{gt} to generate S following [11] to use HED[22]. The $[C, I_n]$ are sent to the color propagation branch at the same time where C is the input colors. In the training process, we generate the colors C of Places2 and CelebeA-HQ datatsets by using RTV [27] and GFC [24], respectively.

Inference. We select the image I_{gt}, and draw the mask M to generate the $I_{in} = I_{gt} \odot (1 - M)$. Meanwhile, we draw the sketch S and colors C in the mask regions, and the random noise N is automatically generated. Finally, we concatenate $[I_{in}, N, M]$ as the inputs to the encoder-decoder, and $[S, N]$ and $[C, N]$ are as the inputs to the structure generation blocks. The image manipulated results are generated via the network output I_{out}.

2. Visual Comparisons on Places2 and CelebA-HQ

We show more comparisons with PConv, DF2, P2P and SGAN on Places2 and CelebA-HQ datasets in Fig. 1 and Fig. 2. Overall, our method is able to effectively create semantic contents conveying user intentions and produce visually pleasant results.

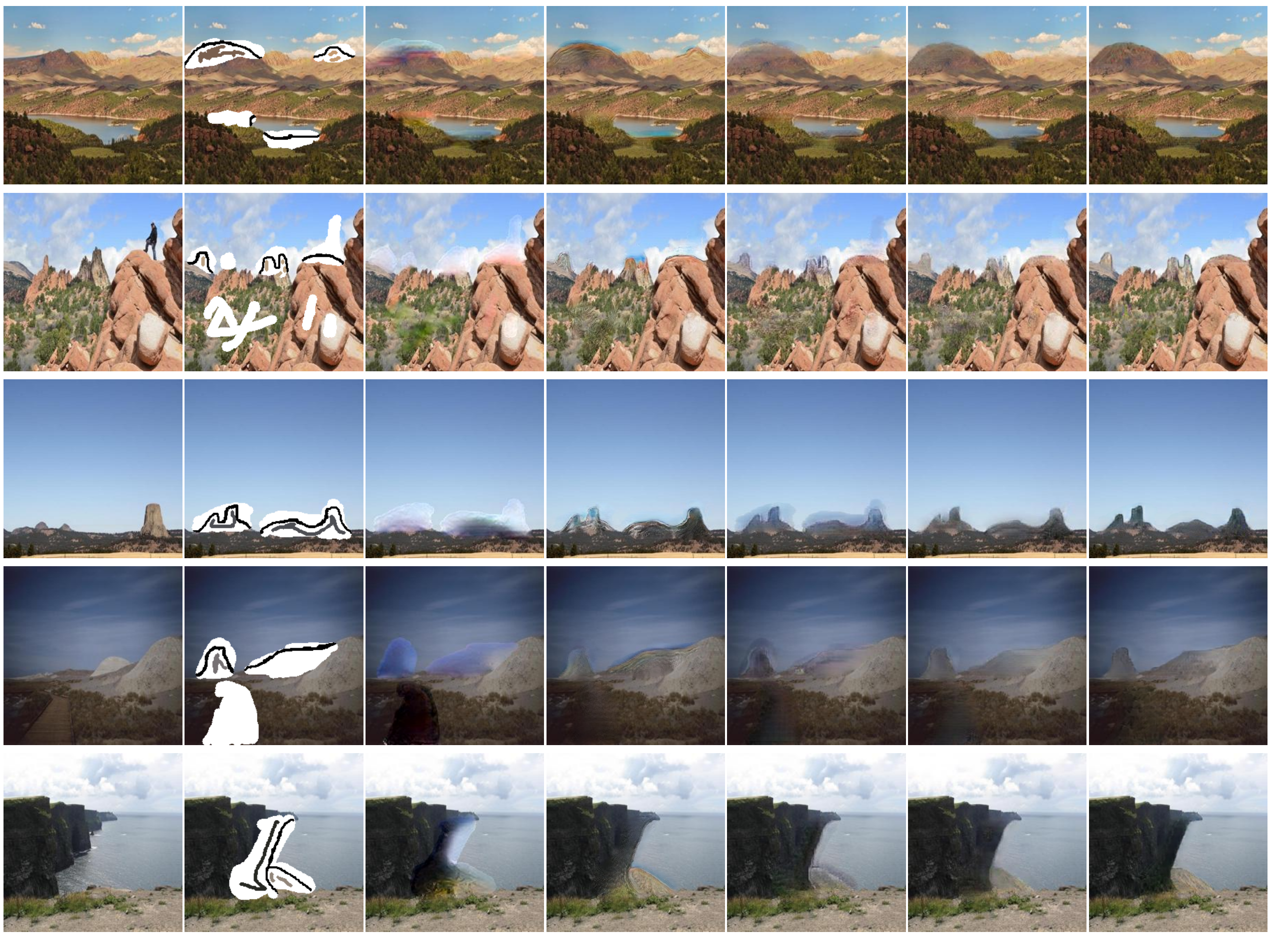
Algorithm 1 Network Training

Input: A set of original images, mask images with arbitrary hole regions, sketch images, color images.

while G has not converged **do**
  Sample batch original images I_{gt}, mask M, sketch images S and color images C;
  Generate batch noisy images N;
  Generate input images $I_{in} = I_{gt} \odot (1 - M)$;
  Get predictions $I_{out} = G([I_{in}, M, I_n]$;
  Calculate the adversarial loss in Equation (9);
  Update D;
  Calculate the loss in Equation (11);
  Update G;
end while

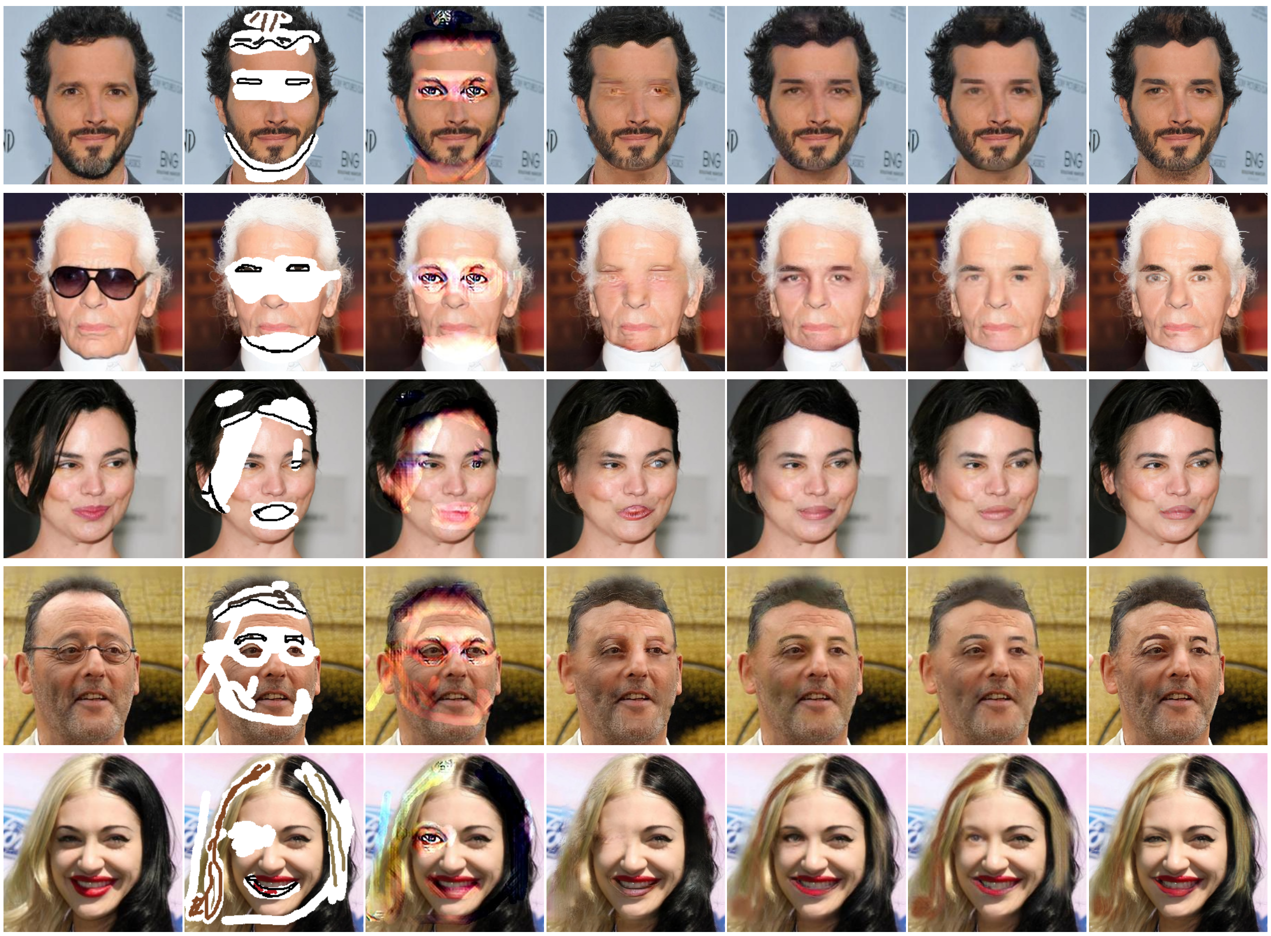
3. Visual results of flexible low-level controls

We show more results generated by our method with different low-level controls in both natural image and face image datasets (i.e., Places2 and CelebA-HQ) in Fig. 3 and Fig. 4, respectively. The results indicate that our method is able to produce semantic contents conveying different user intentions via low-level controls.



(a) Original	(b) Input	(c) PConv [13]	(d) DF2 [25]	(e) P2P [9]	(f) SGAN [8]	(g) Ours

Figure 1: Visual comparisons on the Places2 dataset (i.e., natural images). Our results contain more semantic contents and reduce blurry and artifacts. The readers are suggested to view this figure on a high resolution digital display.



(a) Original (b) Input (c) PConv [13] (d) DF2 [25] (e) P2P [9] (f) SGAN [8] (g) Ours

Figure 2: Visual comparisons on the CelebA-HQ dataset (i.e., face images). Our results produce more faithful human facial components. The readers are suggested to view this figure on a high resolution digital display.

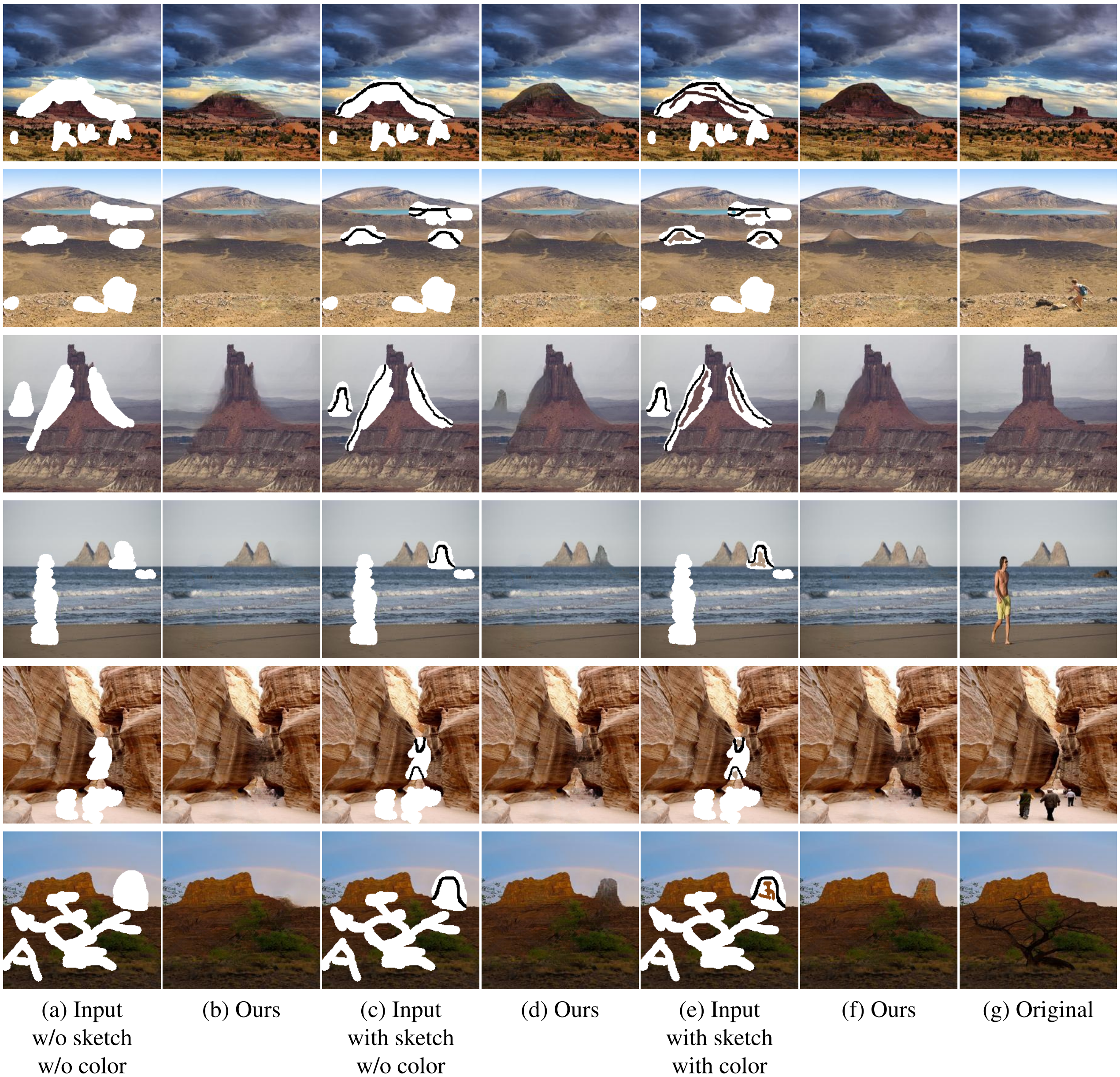


Figure 3: Natural image manipulation with flexible low-level controls. The inputs with no controls are shown in (a). Our method produces faithful results in (b) compared to the original images in (g). Given only partial sketch lines in (c), and the combination of partial sketch lines and colors in (e), our methods produces semantic contents conveying user intentions in (d) and (f). The readers are suggested to view this figure in a high resolution digital display.

(a) Input w/o sketch w/o color | (b) Ours | (c) Input with sketch w/o color | (d) Ours | (e) Input with sketch with color | (f) Ours | (g) Original

Figure 4: Face image manipulation with flexible low-level controls. The inputs with no controls are shown in (a). Our method produces faithful results in (b) compared to the original images in (g). Given only partial sketch lines in (c), and the combination of partial sketch lines and colors (e), our method produces semantic contents conveying user intentions in (d) and (f). The readers are suggested to view this figure in a high resolution digital display.