# Supplementary Material for PD-GAN

In this supplementary material, we first introduce the training details. Then, we present more analysis and ablation study to analyze the proposed SPDNorm and perceptual diversity loss. Finally, we show more quantitative and qualitative comparisons on the benchmark datasets: CelebA-HQ, Place2 and Paris Street View.

## 1. Training Details

### 1.1. Objective Function

Except for the proposed perceptual diversity loss, we follow the SPADE [26] and utilize several objective loss functions to train the PD-GAN end-to-end. These functions include reconstruction loss (perceptual loss [15]), feature matching loss [37] and hinge adversarial loss [20]. We denote the generator of PD-GAN as $G$, $I_{out1} = G(z_1, P, M)$ and $I_{out2} = G(z_2, P, M)$ are two generated images conditioned on same coarse prediction $P$ and image mask $M$ but different latent vector $z_1$ and $z_2$. The loss terms can be written as follows.

**Reconstruction loss.** We consider high-level feature representation and human perception to utilize the perceptual loss as the reconstruction loss, which is based on an ImageNet-pretrained VGG-19 network. The reconstruction loss can be written as:

$$L_{\text{re}} = \sum_{k=1,2} \sum_{i} \|F_i(I_{outk}) - F_i(I_{gt})\|_1, \tag{1}$$

where $F_i$ is the feature map of the $i$-th layer of the VGG-19 network. In our work, $F_i$ corresponds to the activation maps from layers ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1.

**Feature matching loss.** The feature matching loss $L_{fm}$ compares the activation maps in the intermediate layers of the discriminator to stabilize training:

$$L_{\text{fm}} = \sum_{k=1,2} \sum_{i=1}^{L} \|D_{(i)}(I_{outk}) - D_{(i)}(I_{\text{gt}})\|_1, \tag{2}$$

where $L$ is the index of final convolution layer of the discriminator. $D^{(i)}$ is the activation in the $i$'-th layer of the discriminator.

**Hinge adversarial loss.** The hinge adversarial loss is adopted for our PD-GAN. The adversarial objectives of discriminator $D$ and generator $G$ are respectively defined as:

$$\begin{aligned} L_{\text{adv}}^{\text{D}} &= \sum_{k=1,2} -\mathbb{E}\left[h(D(I_{gt}))\right] - \mathbb{E}\left[h(-D(I_{outk}))\right] \\ L_{\text{adv}}^{\text{G}} &= \sum_{k=1,2} -\mathbb{E}\left[D(I_{outk})\right] \end{aligned} \tag{3}$$

where $h(t) = min(0, -1 + t)$ is a hinge function used to regularize the discriminator.

**Total losses.** The whole objective function of generator G can be written as:

$$\begin{aligned} L_{\text{Total}} = &\lambda_{re} \cdot L_{\text{re}} + \lambda_{fm} \cdot L_{\text{fm}} + \lambda_{adv} \cdot L_{\text{adv}}^{\text{G}} + \\ &\lambda_{pdiv} \cdot L_{\text{pdiv}} \end{aligned} \tag{4}$$

where $\lambda_{re}$, $\lambda_{fm}$, $\lambda_{adv}$ and $\lambda_{pdiv}$ are the scalars controlling the influence of each loss term. The $L_{\text{pdiv}}$ is defined in Equation 6 in the main paper. We empirically set $\lambda_{re} = 10$, $\lambda_{fm} = 10$, $\lambda_{adv} = 1$ and $\lambda_{pdiv} = 1$.

### 1.2. Pseudo code of network training

**Training.** The pseudo code of the training process is shown in Algorithm 1. We denote the input image as $I_{in}$, the pretrained partial convolution model [21] as PC, our generator as G, and discriminator as D. The input image $I_{in} = I_{gt} \odot M$, prior information $P = PC(I_{in})$. To calculate the perceptual diversity loss, for each training pair $I_{in}$ and $P$, we random sample $z_1$ and $z_2$ from a standard Gaussian distribution result in two final outputs $I_{out1} = G(z_1, P, M)$, $I_{out2} = G(z_2, P, M)$ .

## 2. More Analysis and Ablation Study

### 2.1. Analysis of Hard and Soft SPDNorm

The hard SPDNorm increases the probability of getting diverse results but reduces the quality of the results. In contrast, the soft SPDNorm can stabilize the training and dynamically learn the condition of the prior information but reduces the diversity. We combine the hard and soft SPDNorm by the proposed SPDNorm ResBlock to balance

**Algorithm 1** Network Training

**while** G has not converged **do**
  Sample batch $I_{gt}$ and $M$ from training data, sample batch $z_1$ and $z_2$ from a standard Gaussian distribution.
  Generate $I_{in} = I_{gt} \odot M$
  Send input image $I_{in}$ to $PC$;
  Get prior information $P = PC(I_{in})$.
  Get predictions $I_{out1} = G(z_1, P, M)$ and $I_{out2} = G(z_2, P, M)$;
  Calculate the $L_{adv}^{D}$ in Equation (3);
  Update D;
  Calculate the loss in Equation (4);
  Update G;
**end while**

the diversity and quality of generated content. In order to show the characteristics and performance of hard and soft SPDNorm, after we get the trainied PD-GAN, we set the outputs of hard SPDNorm and soft SPDNorm to 0 respectively and show the corresponding results in Fig. 4. When setting the soft SPDNorm branch to 0, The generated content of hard SPDNorm branch is diverse, but the quality is poor (see Hard$_{1\sim5}$). In contrast, when we disable the hard SPDNorm, the output of soft SPDNorm branch has plausible structure but lacks of diversity (see Soft$_{1\sim5}$). To sum up , PD-GAN combines the hard and soft SPDNorm by SPDNorm ResBlock to get better results (see Out$_{1\sim5}$).

### 2.2. Analysis of random latent code

We randomly sample two vectors $z_1$ and $z_2$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then we generate the inpainting results via linear interpolation $z = z_1 + \lambda \frac{z_2 - z_1}{||z_2 - z_1||}$, where $0 \leq \lambda \leq 1$. As shown in Fig. 1, we find a direction which can change the openness of the mouth. In the final version, we will investigate deeper into the directions of interpolation so that we can better control both the textural level and semantic level diversity of results. Moreover, we also find the diversity of inpainted results could be controlled via adjusting the $\sigma$ of sampled normal distribution as shown in Fig. 2.

### 2.3. Effects of the coarse prior and failure cases

We empirically find that given more accurate prior (GT image), the diversity is weaken but the quality improves (Ours$_{2\sim3}$ in Fig. 3). While the outputs of the PC network is bad, the diversity increases while the quality decrease (Ours$_{1\sim2}$ in Fig. 3). Since the PD-GAN should generate diverse and realistic content, so lack of accuracy or diversity are the main failure cases (Ours$_{1\sim4}$ in Fig. 3).

### 2.4. Ablation study

Table 1 shows the quantitative comparison of our ablations on the CelebA-HQ dataset. We choose the center mask for testing. In contrast to our method without soft SPDNorm (w/o s), the SSIM, PSNR and FID of our method without hard SPDNorm (w/o h) are better but the LIPIS is poor. These comparison results suggest that hard SPDNorm can ensure the diversity but reduce the quality of final results. While soft SPDNorm has the opposite effect. When we abandon the diversity loss (w/o diver), our method can gets high-quality predictions, but the diversity of results decreases. The conventional diversity loss (w/ CDL) can increase the diversity of generated content, but the recovered content tend to be all black or all white. SPADE [26] is not suitable for diverse image inpainting task, and the performance is unsatisfactory. In sum, our method with soft SPDNorm, hard SPDNorm and perceptual diversity loss can guarante the diversity and authenticity of final results.


Figure 1: Interpolation along the direction ($\frac{z_2 - z_1}{||z_2 - z_1||}$) reliably opens mouth.

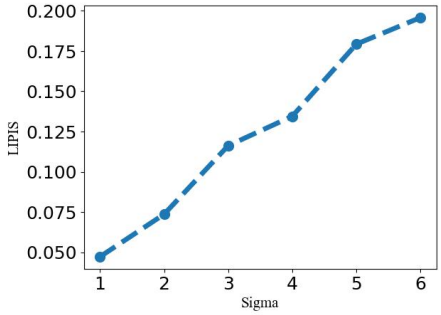



Figure 2: The diversity (measured by LIPIS) increases when we increase $\sigma$ of the sampled normal distribution.

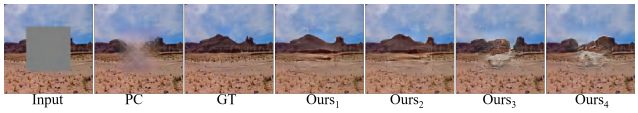



Figure 3: PC denotes one bad result of PConv. GT is the ground truth. Our$_{1\sim2}$ use GT as the prior. Our$_{3\sim4}$ use PC as the prior.

## 3. More Qualitative and Quantitative comparisons

**Quantitative and quantitative comparisons between PIC [50] and our method on center mask.** As shown in Fig 5 and Fig 6, we make quantitative comparisons for center mask on Places2 dataset. In contrast to the free form mask,

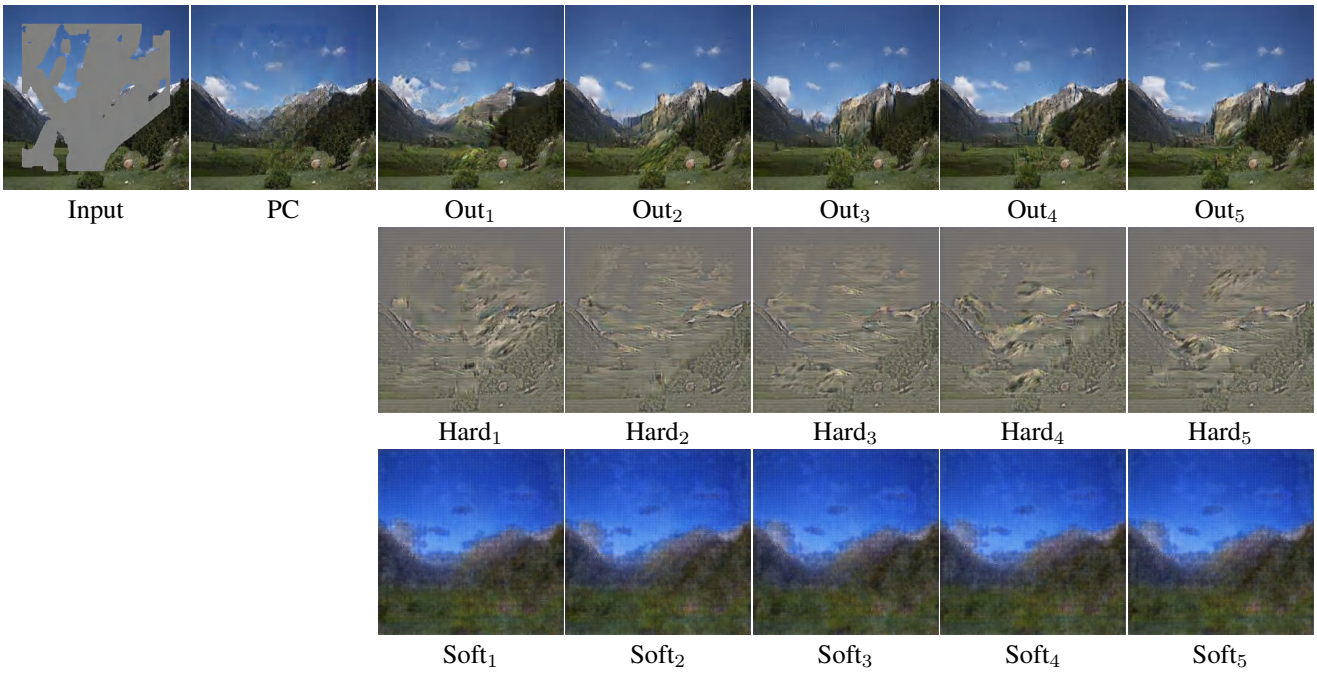



Figure 4: Analysis of hard and soft SPDNorm. The diverse outputs of hard SPDNorm branch are $\text{Hard}_{1\sim5}$. The diverse outputs of soft SPDNorm branch are $\text{Soft}_{1\sim5}$. The diverse outputs under complete SPDNorm residual block are $\text{Our}_{1\sim5}$

Table 1: Quantitative comparison of ablation study on the CelebA-HQ dataset.

| | PSNR ↑ | SSIM ↑ | FID ↓ | LIPIS ↑ |
|---|---|---|---|---|
| w/ CDL | 26.12 | 0.915 | 18.55 | 0.0490 |
| w/o diver | **26.37** | **0.918** | 17.90 | 0.0172 |
| w/o h | 26.35 | 0.913 | 18.24 | 0.0516 |
| w/o s | 26.25 | 0.910 | 20.51 | **0.0623** |
| SPADE | 25.79 | 0.903 | 24.71 | 0.0479 |
| Ours | 26.32 | 0.915 | **16.83** | 0.0590 |

the effect of PIC to restore the center hole is relatively better. This is because the proportion of background regions in each masked image is fixed, thus PIC can found a balance between reconstruction (image quality) and diversity. However, for the free form mask, the location and size of the background regions are random. As the hole area decreases, the influence of context increases and the diversity of generated content decreases. PIC cannot find the balance between the reconstruction and diversity in this complex scene. Unlike PIC, our method can handle free from mask and center mask at the same time.

**Qualitative comparisons of BicycleGAN [52], PIC [50], CVAE [35] and ours on CelebA-HQ [17].** We show more comparison results in Fig 7 and Fig 8. The results indicate that our method is able to produce diverse and realistic results.

**Qualitative comparisons of GC[47], PC [21], EC [25], SF [28], PIC [50] and ours.** As shown in Fig 9, Fig 10 and Fig 11, we show more comparison results on Paris Street View [17], CelebA-HQ [7] and Places2 [51] respectively. Compared with other methods, our method can generate reasonable content and various details.

## 4. More visual results

We show more visual results on Paris Street View, CelebA-HQ and Places2 in Fig 12, Fig 13 and Fig 14 respectively. The results are obtained by our full model with irregular masks. As shown in these visual results, our method can not only produce reconstruction images with high quality but also generate diverse content.

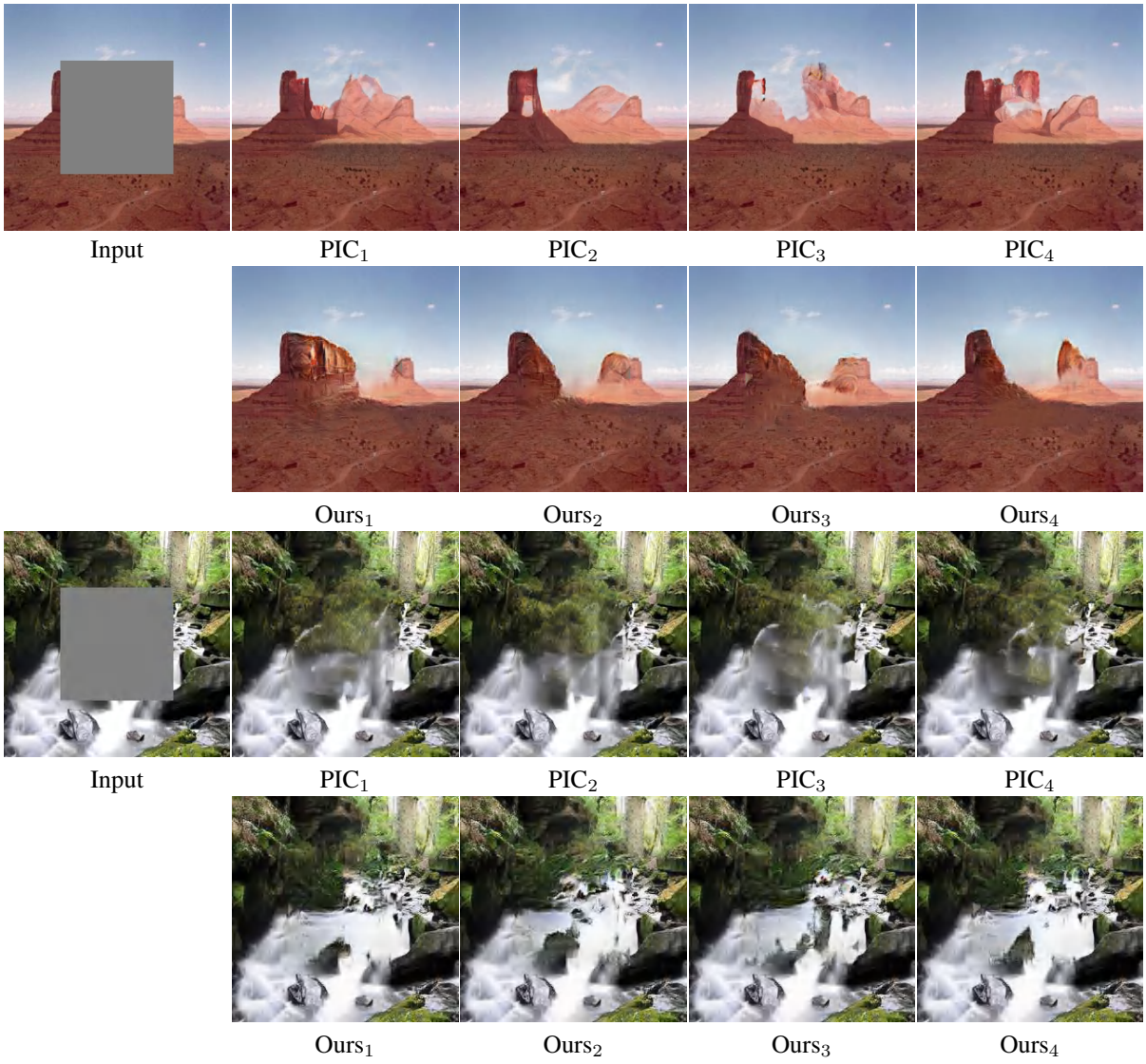



Figure 5: Qualitative comparisons on Places2 with center mask. The predictions of PIC are in PIC$_{1\sim4}$. The results of our method are in Ours$_{1\sim4}$

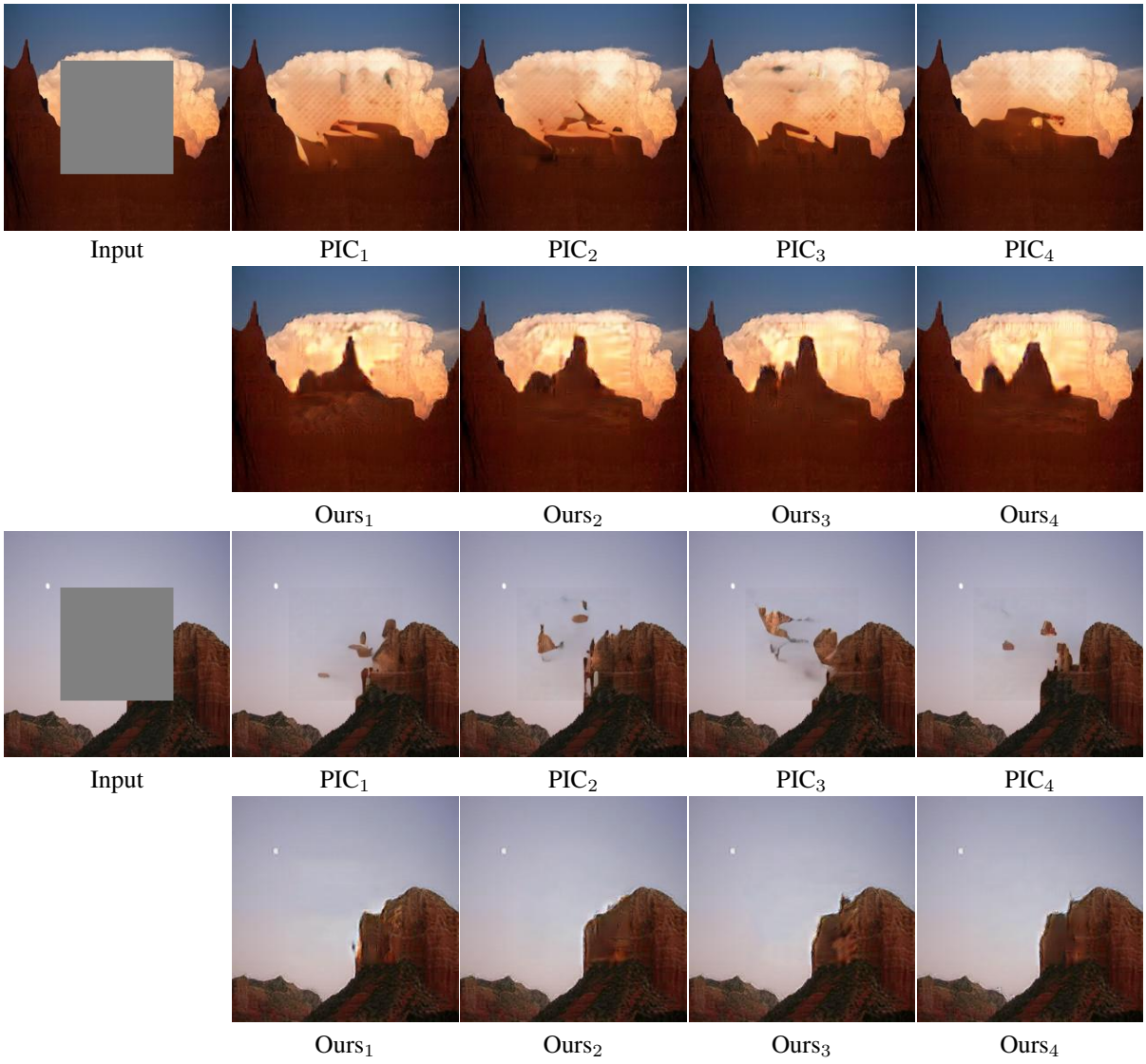



Figure 6: Qualitative comparisons on Places2 with center mask. The predictions of PIC are in PIC$_{1\sim4}$. The results of our method are in Ours$_{1\sim4}$

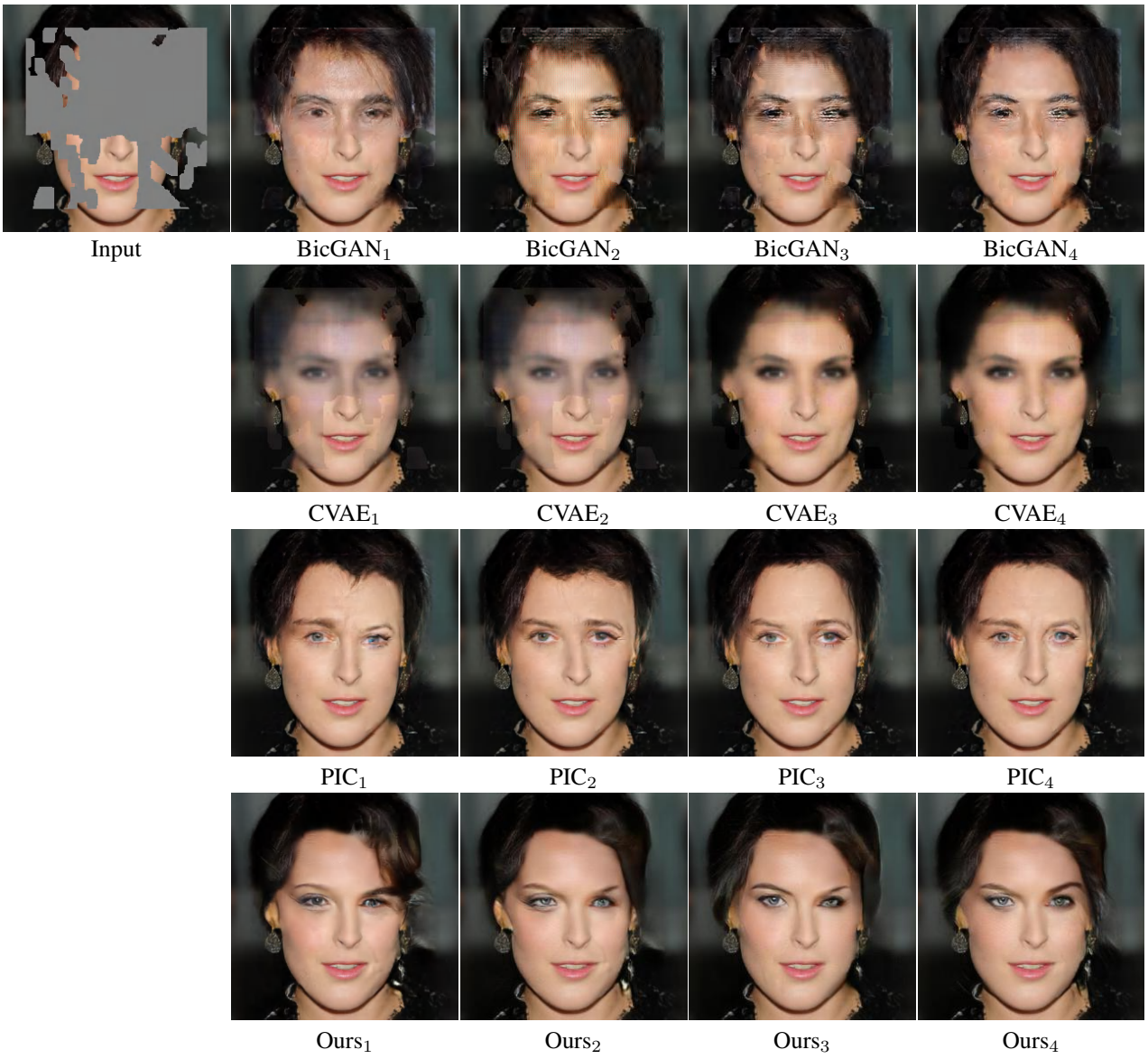



Figure 7: Qualitative comparisons on CelebA-HQ. The results of BicycleGAN are in BicGAN$_{1\sim4}$. The generated content of CVAE are in CVAE$_{1\sim4}$. The predictions of PIC are in PIC$_{1\sim4}$. The results of our method are in Ours$_{1\sim4}$

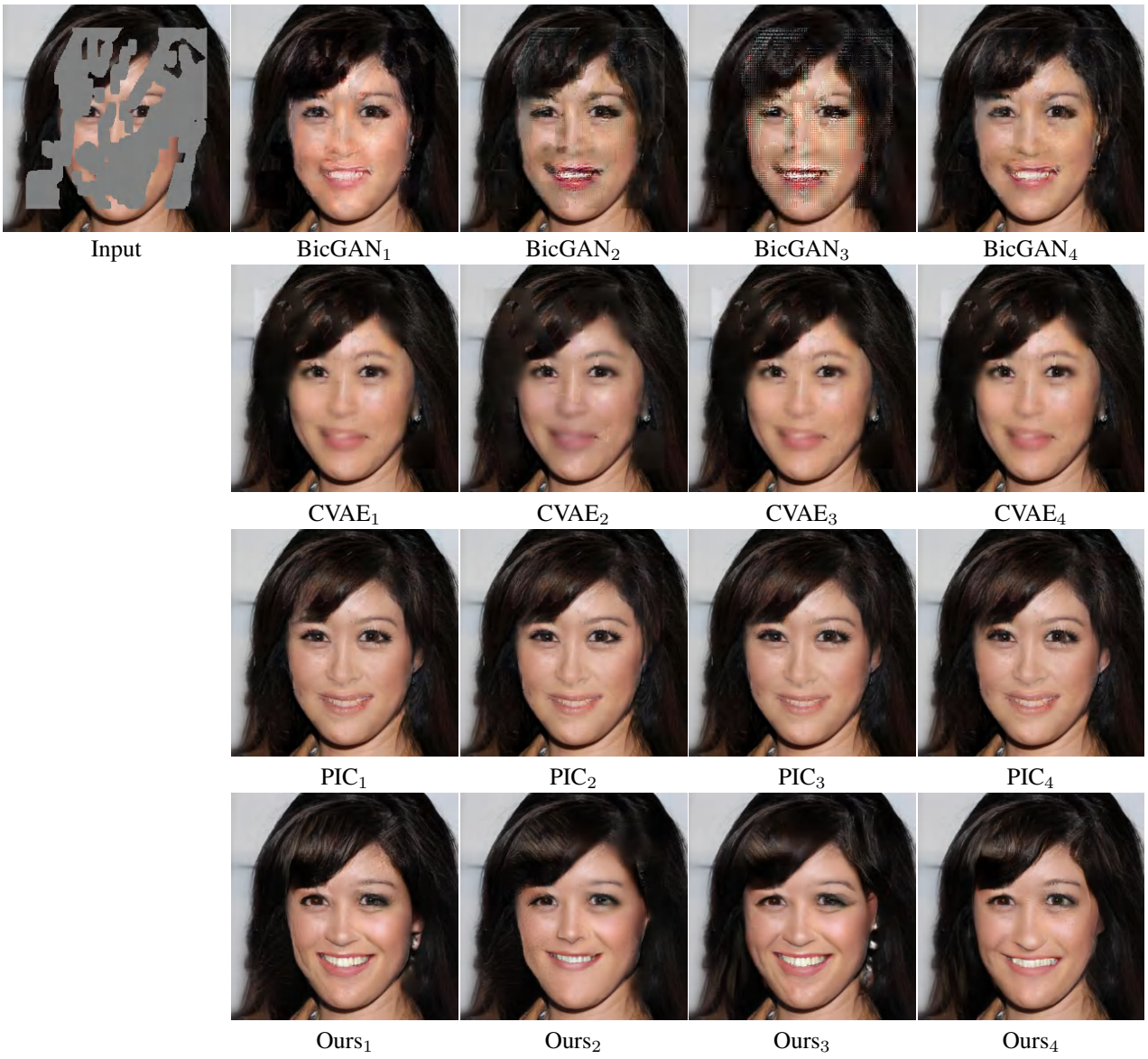



Figure 8: Qualitative comparisons on CelebA-HQ. The results of BicycleGAN are in BicGAN$_{1\sim4}$. The generated content of CVAE are in CVAE$_{1\sim4}$. The predictions of PIC are in PIC$_{1\sim4}$. The results of our method are in Ours$_{1\sim4}$

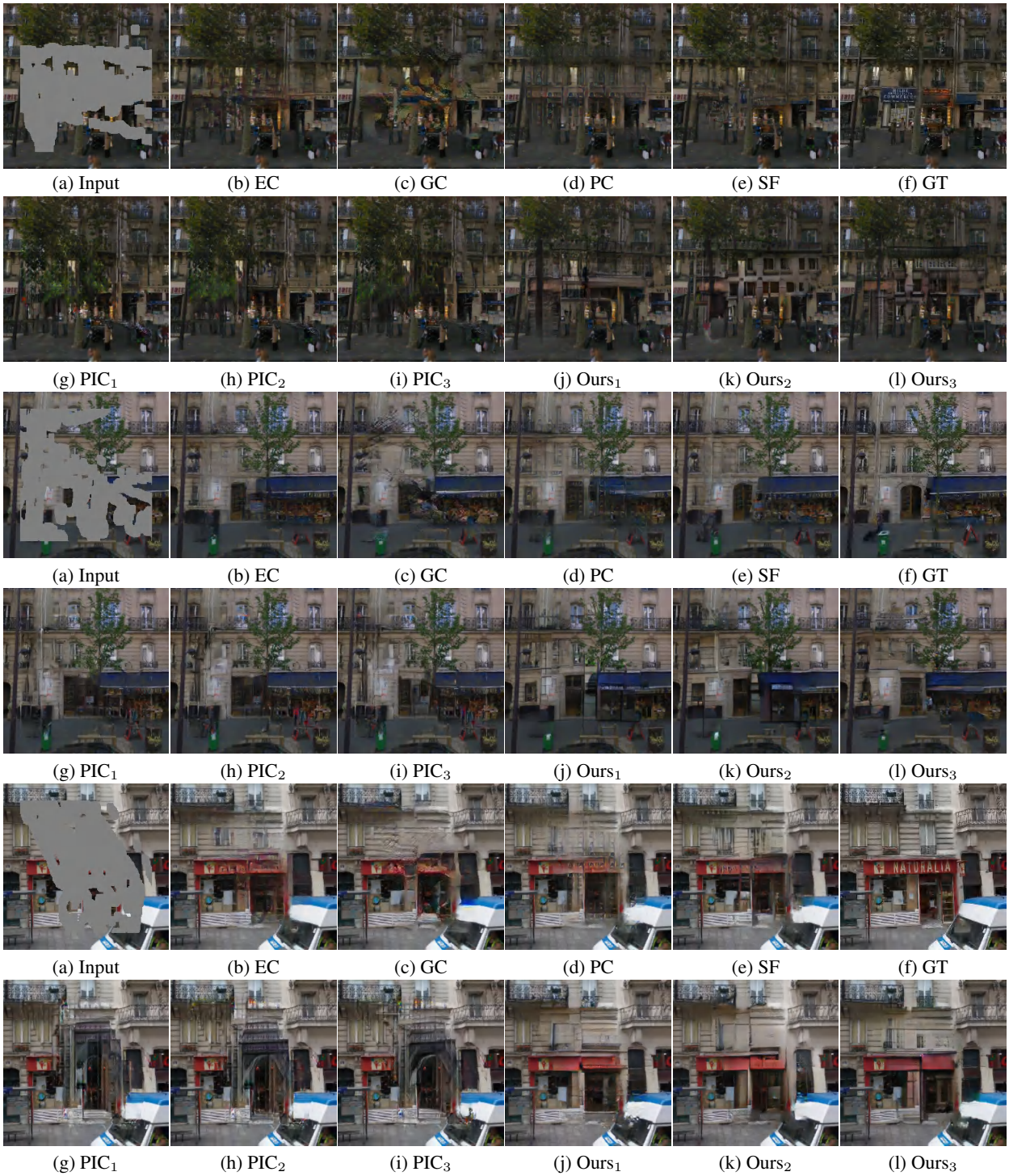



Figure 9: Qualitative comparisons with state-of-the-art methods on Paris Street View. Original images are in (f). Input images are in (a). The prior information is the output of PC in (d). The diverse outputs of PIC are in (g)-(i). The diverse outputs of our method are in (j)-(l).

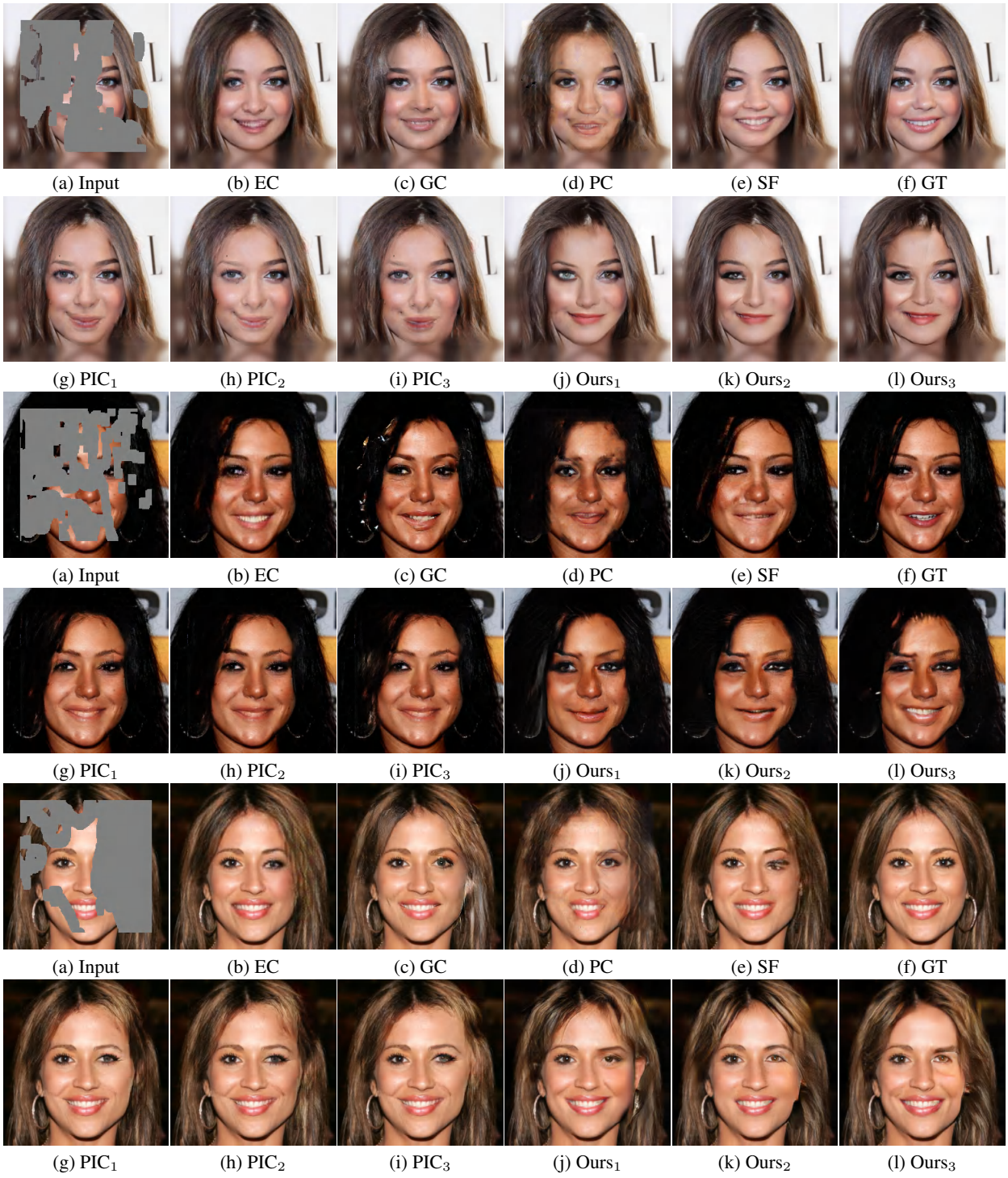



Figure 10: Qualitative comparisons with state-of-the-art methods on CelebA-HQ. Original images are in (f). Input images are in (a). The prior information is the output of PC in (d). The diverse outputs of PIC are in (g)-(i). The diverse outputs of our method are in (j)-(l).

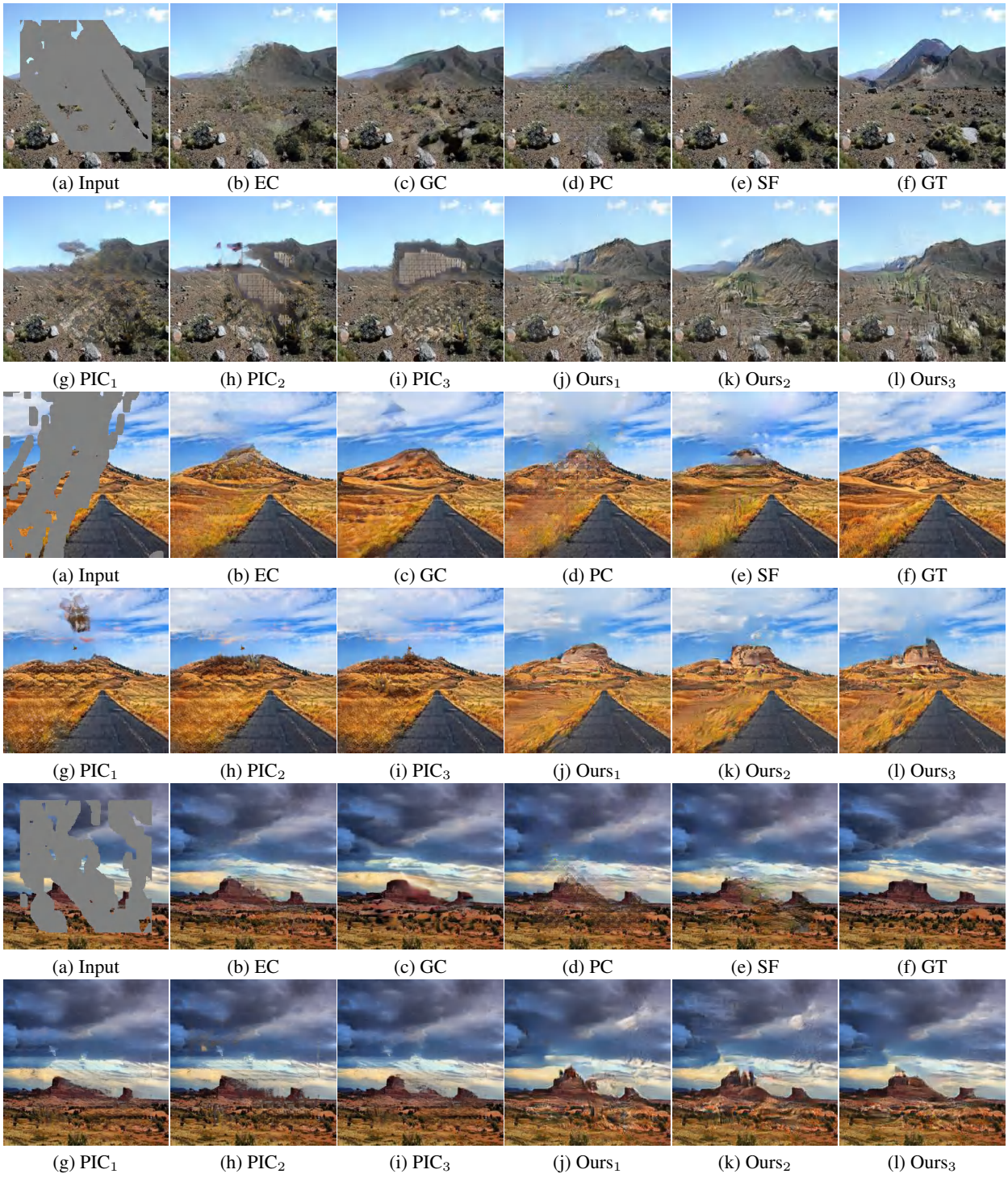
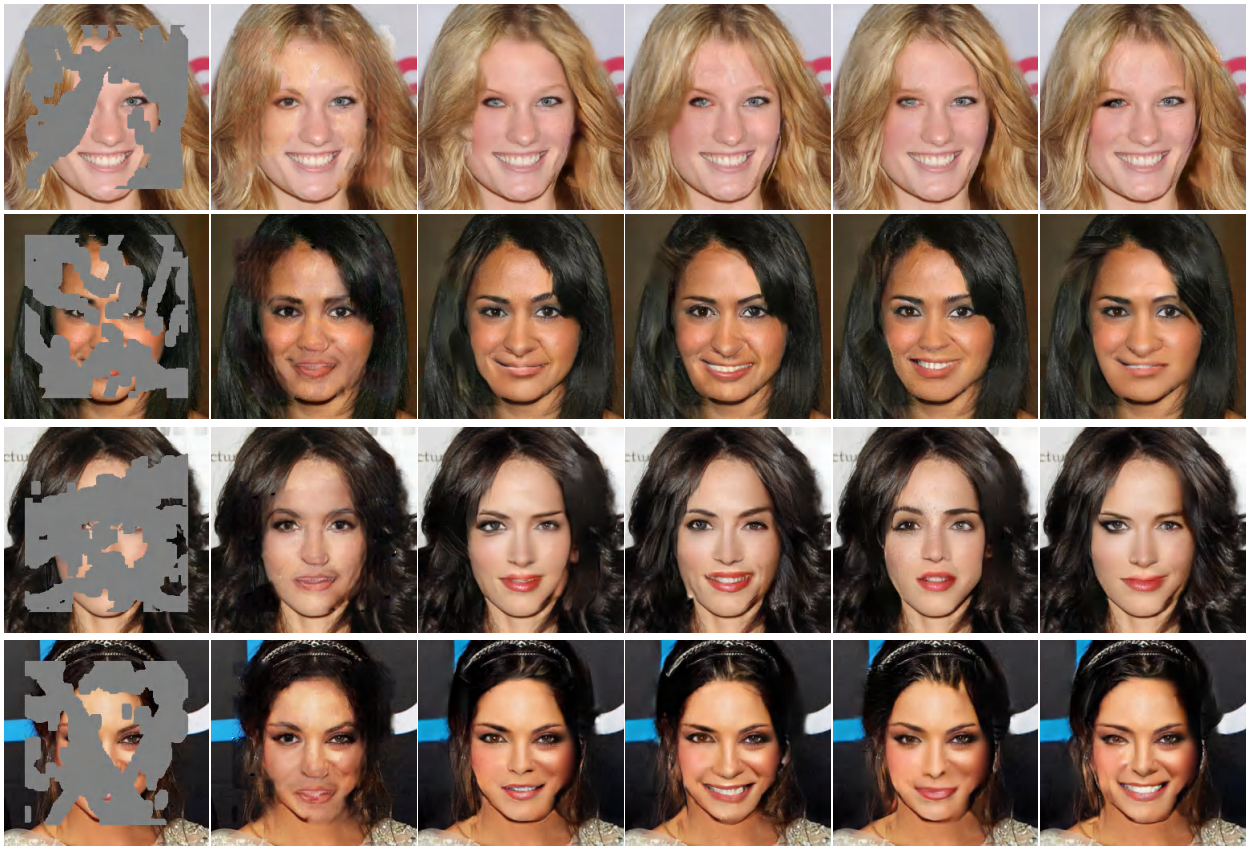



Figure 11: Qualitative comparisons with state-of-the-art methods on Places2. Original images are in (f). Input images are in (a). The prior information is the output of PC in (d). The diverse outputs of PIC are in (g)-(i). The diverse outputs of our method are in (j)-(l).

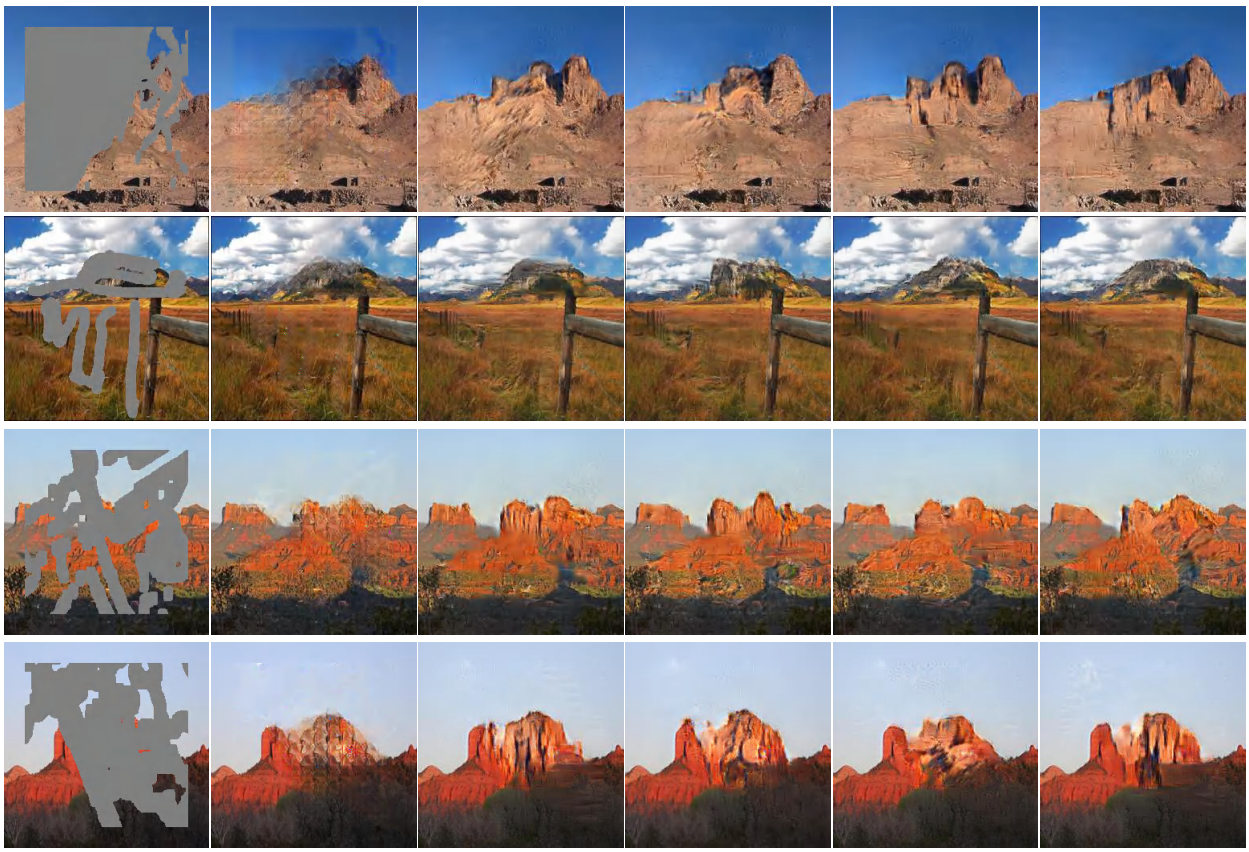
(a) Input (b) PC (c) Ours$_1$ (d) Ours$_2$ (e) Ours$_3$ (f) Ours$_4$

Figure 12: Visual results on Parry Street View.



(a) Input (b) PC (c) Ours$_1$ (d) Ours$_2$ (e) Ours$_3$ (f) Ours$_4$

Figure 13: Visual results on CelebA-HQ.



(a) Input (b) PC (c) Ours$_1$ (d) Ours$_2$ (e) Ours$_3$ (f) Ours$_4$

Figure 14: Visual results on Places2.