Geometry-Aware Face Completion and Editing

Linsen Song, Jie Cao, Lingxiao Song, Yibo Hu and Ran He
Institute of Automation, Chinese Academy of Sciences

Abstract
Face completion is a challenging generation task because it requires generating visually pleasing new pixels that are semantically consistent with the unmasked face region. This paper proposes a geometry-aware Face Completion and Editing NETwork (FCENet) by systematically studying facial geometry from the unmasked region. Firstly, a facial geometry estimator is learned to estimate facial landmark heatmaps and parsing maps from the unmasked face image. Then, an encoder-decoder structure generator serves to complete a face image and disentangle its mask areas conditioned on both the masked face image and the estimated facial geometry images. Besides, since low-rank property exists in manually labeled masks, a low-rank regularization term is imposed on the disentangled masks, enforcing our completion network to manage occlusion area with various shape and size. Furthermore, our network can generate diverse results from the same masked input by modifying estimated facial geometry, which provides a flexible mean to edit the completed face appearance. Extensive experimental results qualitatively and quantitatively demonstrate that our network is able to generate visually pleasing face completion results and edit face attributes as well.

1 Introduction
Face completion, also known as face inpainting, aims to complete a face image with a masked region or missing content. As a common face image editing technique, it can also be used to edit face attributes. The generated face image can either be as accurate as the original face image, or content coherent to the context so that the completed image looks visually realistic. Most traditional methods (Barnes et al. 2009; Huang et al. 2014; Darabi et al. 2012) rely on low-level cues to search for patches to synthesize missing content. These image completion algorithms are skilled in filling backgrounds which contain similar textures but not excel at specific image object like face since prior domain knowledge isn’t well incorporated.

Recently, CNN based image completion methods with adversarial training strategy have already significantly boosted the image completion performance (Pathak et al. 2016; Yeh et al. 2017; Li et al. 2017). These methods set a masked image as the network input to learn context features, then restore the missing content from the learned features. Some challenges still exist in these methods when they are applied to real-world face completion problems. First of all, different from general objects, human faces have distinct geometry distribution, and hence face geometry prior is more likely to facilitate face completion (Yeh et al. 2017). Most of the existing methods don’t well utilize the prior knowledge to boost face completion. Second, these algorithms are incapable of modifying the face attributes of the filled region (Li et al. 2017), e.g., editing the eye shape or mouth size in figure 1.

To address these two problems, this paper studies a geometry-aware face completion and editing network (FCENet) by exploring facial geometry priors including...
3. Our face completion generator simultaneously accom-
2. The FCENet allows interactive facial attributes editing
1. We design a novel network called facial geometry estima-
erate realistic outputs to fool the discriminator. At the same
versarial training method encourages the generator to gen-
train a generator and a discriminator alternatively. The ad-
most significant image generation techniques, GAN (Good-
follows:
FCENet are shown in figure 1.
face image. A few face completion and editing examples of
on the facial geometry images concatenated with masked
generate face images with a variety of attributes depends
from the corrupted face image. Our FCENet is efficient to
generate face images with a variety of attributes depends
in the third stage, two discriminators distinguishe
real ones globally and locally to force the generated face images as realistic as possible. Furthermore, the low-rank regularization boosts
from the corrupted face image. Our FCENet is efficient to
generate face images with a variety of attributes depends
on the facial geometry images concatenated with masked
a few face completion and editing examples of
FCENet are shown in figure 1.

The main contributions of this paper are summarized as follows:

1. We design a novel network called facial geometry estima-
tor to estimate reasonable facial geometry from masked
face images. Such facial geometry is leveraged to guide
the face completion task. Several experiments systematically
demonstrate the performance improvements from
different facial geometry, including facial landmarks, fa-
cial parsing maps and both together.

2. The FCENet allows interactive facial attributes editing
of the generated face image by simply modifying its in-
ferred facial geometry images. Therefore, face comple-
tion and face attributes editing are integrated into one uni-
ified framework.

3. Our face completion generator simultaneously accomplis-
hes face completion and mask disentanglement from
the masked face image. A fancy low-rank loss regularizes
the disentangled mask to further enhance similarity be-
tween the disentangled mask and the original mask, which
enables our method handle various masks with different
shapes and sizes.

4. Experiments on the CelebA (Liu et al. 2015b) and Multi-
PIE (Gross et al. 2010) dataset demonstrate the superiority
of our approach over the existing approaches.

2 Related Work

Generative Adversarial Network (GAN). As one of the
most significant image generation techniques, GAN (Good-
fellow et al. 2014) utilizes the mini-max adversarial game to
train a generator and a discriminator alternatively. The ad-
versarial training method encourages the generator to gen-
erate realistic outputs to fool the discriminator. At the same
time, the discriminator tries to distinguish real and generated
images. In this way, the generator can generate samples
that obey the target distribution and look plausibly realistic.
Radford et al. propose a deep convolutional GAN (DCGAN)
(Radford, Metz, and Chintala 2016) to generate high-quality
images on multiple datasets. It is the first time to integrate
the deep models into GAN. To address training instability
of GAN, Arjovsky et al. (Arjovsky, Chintala, and Bottou
2017) analyze the causes of training instability theoretically
and propose Wasserstein GAN. Mao et al. (Mao et al. 2017)
put forward LSGAN to avoid vanishing gradients problem
during the learning process. Recently, improved GAN archi-
tectures have achieved a great success on super-resolution
(Ledig et al. 2017), style transfer (Li and Wand 2016), image
inpainting (Pathak et al. 2016), face attribute manipulation
(Shen and Liu 2017), and face rotation (Huang et al. 2017;
Hu et al. 2018). Motived by these successful solutions, we
develop the FCENet based on GAN.

Image Completion. The image completion techniques can
be broadly divided into three categories, The first cate-
gory exploits the diffusion equation to propagate the low-
level feature from the context region to the missing area
along the boundaries iteratively (Bertalmio et al. 2000;
Elad et al. 2005). These methods excel at inpainting small
holes and superimposed text or lines but having limitations
to the reproduction of large textured regions. The second
category is patch-based methods, which observe the con-
text of the missing content and search similar patch from the
same image or external image databases (Darabi et al. 2012;
Bertalmio et al. 2003; Criminisi, Pérez, and Toyama 2004;
Hays and Efros 2007). These methods can achieve ideal
completion results on backgrounds like sky and grass, but
they fail to generate semantic new pixels if these patches
don’t exist in the databases. The third category is CNN-
based, which train an encoder-decoder architecture network
to extract image features from the context and generate miss-
ing content according to the image features (Pathak et al.
2016; Iizuka, Simo-Serra, and Ishikawa 2017). Deepak et
al. (Pathak et al. 2016) present an unsupervised visual fea-
ture learning algorithm called context encoder to produce a
plausible hypothesis for the missing part conditioned on its
surroundings. A pixel-wise reconstruction loss and an adver-
sarial loss regularize the filling content to bear some resem-
blance to its surrounding area both on appearance and se-
semantic segmentation. Satoshi et al. (Iizuka, Simo-Serra,
and Ishikawa 2017) put forward an approach which can gen-
erate inpainting images that are both locally and globally con-
sistent by training a global and local context discriminator.
Their method is able to complete images of arbitrary res-
olutions and various obstructed shapes by training a fully-
convolutional neural network.

Face Completion. Human face completion is much more
challenging than general image completion tasks because fa-
cial components (e.g., eyes, nose, and mouth) are of highly
structurization and contain large appearance variations. Be-
sides, the symmetrical structure is reflected on human faces
like many natural objects. Second, compared to general ob-
ject image completion, face completion need to pay more
attention to preserving face identity. S. Zhang et al. (Zhang
et al. 2018; 2016) develop models to complete face images with structural obstructions like wavy lines. Y. Li et al. (Li et al. 2017) propose a generative face completion model assisted by a semantic regularization term. Their algorithm fails to fill missing region for unaligned faces and cannot manipulate face attributes of the missing areas. P. Liu et al. (Liu et al. 2017) integrate perceptual loss into their network and replace the unmasked region of the generated face image with that from the original face image. Their approach produces high-quality face images with fine-grained details. Nevertheless, how to take advantage of facial geometry and control the face attributes of the filled region is still an open challenge, and that is the motivation of our FCENet.

3 Proposed Method

In this section, we introduce the FCENet for face completion and attributes editing. The FCENet consists of three parts: first, a facial geometry estimator learns to infer reasonable and natural facial parsing maps and landmark heatmaps. Second, an encoder-decoder structural generator restores the completion face image and disentangles the mask from the concatenation of the mask face image and facial geometry images. Third, global and local discriminators are introduced to determine whether the generated face images are sampled from ground truth distribution. The overall framework of our algorithm is shown in figure 2.

3.1 Network Architecture

**Facial Geometry Estimator** Geometry is the most distinct feature of the most real-world object, including human faces. Human faces contain various visual appearances due to factors like gender, illumination conditions, makeup, etc. However, similar facial geometry information still exists in these faces. As prior domain knowledge, facial geometry is rarely exposed to these influences. Most existing face completion algorithms have not yet explored the benefit of facial geometry prior. Thus we propose facial geometry estimator \( P \) to exploit facial geometry in face completion. The facial geometry serves as the guidance information for completion task, which is different from (Li et al. 2017) whose method treats the parsing map as a semantic regularization by closing the distance between the parsing maps of the original image and the generated image.

Details of the facial geometry estimator can be seen in figure 3. We apply the hourglass(HG) structure to estimate facial geometry inspired by its successful application on human pose estimation (Chen et al. 2017; Newell, Yang, and Deng 2016). The HG block adopts a skip connection mechanism between symmetrical layers, which consolidates features across scales and preserves spatial information in different scales. Stacked hourglass blocks are adopted in our approach, more stacked hourglass blocks lead to better learning ability of estimating facial geometry. However, considering more stacked HG blocks consume more inference time and more computational power, two stacked hourglass(HG) blocks are adopted as a trade-off between inference time and facial geometry estimation performance. The input masked image of \( W \times H \) is pre-processed by several residual blocks. Width and height of feature maps extracted by each HG block is half of its input. Thus the size of fa-
Facial geometry feature maps output by the last HG block is \( W/4 \times H/4 \). Then, a 1 \( \times \) 1 convolutional layer is attached to post-process the obtained features. Finally, two branches with same structure are stacked on top of the post-process layer to generate landmark heatmaps \( I_m^p \) and facial parsing maps \( I_m^g \). Both the branches contain two \( 2 \times \) upsample layers so that their output facial geometry images holding the same size with the input masked image. The landmark heatmaps \( I_m^p \) encode more accurate facial geometry information while the facial parsing maps \( I_m^g \) carry finer granularity.

**Face Completion Generator and Discriminators** Our generator \( G \) has an encoder-decoder architecture that contains four parts: one encoder, two decoders, and one non-parametric fusion operation. Given the masked face images concatenated with inferred facial geometry images, the encoder extracts a feature vector that can be split into context feature vector and mask feature vector. These two feature vectors are fed into the face image decoder and mask decoder respectively. Our encoder and decoders have the same architecture, composing of 9 residual blocks evolved from ResNet (He et al. 2016). The structure of encoder is symmetrical to that of the two decoders, and they bear almost the same network architecture except for the input layer. By splitting the latent code inferred by the encoder into two feature vectors, face context feature and mask feature can be well disentangled. The generated face completion result is denoted as \( G^f(I_m) \) and the disentangled mask is denoted as \( G^m(I_m) \) where we omit \( I_m^p \) and \( I_m^g \) for simplicity. At last, the input masked face image is recovered as \( I_r \) by a simple arithmetic operation conducted on \( G^f(I_m) \) and \( G^m(I_m) \) to facilitate the disentangling task further. In our framework, each pixel value is normalized between -1 and 1, and the disentangled mask is expected to be a matrix filled with -1 and 1, where 1 represents the masked area and -1 represents the unmasked area. Thus, the recovered masked face image is formulated as follows,

\[
I_r = \max(G^f(I_m), G^m(I_m))
\]  

where the \( \max(\cdot, \cdot) \) is an element-wise maximum function.

The generator completes face images under the guidance of inferred facial geometry from the masked input image. The confidence values they assign to the entire image and masked region patch are denoted as \( D^f(\cdot) \) and \( D^m(\cdot) \) respectively. Then under the guidance of modified facial geometry, our FCENet generates face images with desired facial attributes. Thus, by editing the inferred facial geometry of filled regions, as highlighted by the red box in figure 2. The generator is able to produce a diversity of face images with different attributes. Facial attributes editing cases are presented in section 4.4.

Our facial geometry-aware FCENet has already produced visually realistic face images in this way. However, some attributes of generated face images might not be satisfactory. Editing facial attributes by changing its guidance information is a natural idea, e.g., editing facial attributes by modifying inferred facial landmark heatmaps and parsing maps. Multiply ways can be explored in modifying facial geometry images. One alternative way is the copy-and-paste strategy since abundant facial attributes exist in facial geometry images from the training set, thus image patches that possess desired facial attributes can be utilized. Another alternative way is modifying facial geometry images directly by moving landmark points or changing edges of parsing maps.

The generator can only capture visually realistic face images in this way. However, some attributes of generated face images might not be satisfactory. Editing facial attributes by changing its guidance information is a natural idea, e.g., editing facial attributes by modifying inferred facial landmark heatmaps and parsing maps. Multiply ways can be explored in modifying facial geometry images. One alternative way is the copy-and-paste strategy since abundant facial attributes exist in facial geometry images from the training set, thus image patches that possess desired facial attributes can be utilized. Another alternative way is modifying facial geometry images directly by moving landmark points or changing edges of parsing maps.

The global and local discriminators are trained to compete with the generator. By optimizing the generator and discriminators alternatively, the generator \( G \) produces complete face images which are photo-realistic and of high-quality (Li et al. 2017). The global and local discriminators are composed of same network structure apart from the input layer. The confidence values they assign to the entire image and masked region patch are denoted as \( D^g(\cdot) \) and \( D^l(\cdot) \) respectively. The generator \( G \) is trained to generate visually realistic face images through adversarial training with the two discriminators, whose object is to solve the following min-max

![Figure 3: Detailed architecture of proposed facial prior estimator. The stacked hourglass blocks extract facial geometry features, the two branches infer natural landmark heatmaps and parsing maps based on shared facial geometry features.](image-url)
problem,
\[
\min_{\theta_G} \max_{\theta_D, \phi_P} \mathbb{E}_{I \sim P(I)} [\log D^g(I) + \log D^l(\text{crop}(I))] + \\
\mathbb{E}_{I \sim P(I^m)} [\log(1 - D^g(G^l(I^m))) + \\
\log(1 - D^l(\text{crop}(G^l(I^m))))]
\]
where we denote \(\text{crop}(\cdot)\) as the function to crop the patch of the original missing region from the generated complete face image.

### 3.2 Loss Functions

#### Low-rank Regularization

The mask image we adopt in our approach is a black-background gray image with white squares representing masked regions. Thus, the mask matrix contains only 1 and -1, where 1 for the masked region and -1 for the unmasked region. The rank values of several typical mask matrices are presented in figure 4. Therefore, low-rank regularization is beneficial to the denoising of the disentangled mask matrix. When the low-rank regularization is incorporated into the proposed FCENet, it is required to be back-propagated, whereas simple \(\text{rank}(\cdot)\) function doesn't satisfy such a condition. Thus, an approximate variant is applied in our algorithm in practice. We notice that if the mask matrix \(M\) is a low-rank matrix, its elements are linearly correlated. Then \(M^TM\) tends to be a block-diagonal matrix. Thus, its nuclear norm can be used to measure its rank, i.e.
\[
||M||_* = \text{tr}(\sqrt{M^TM})
\]
Given the SVD decomposition of \(M = U\Sigma V^T\), we can obtain:
\[
L_{\text{rank}} = ||M||_* = \text{tr}(\sqrt{\Sigma^TV^TU\Sigma V^T}) = \text{tr}(\sqrt{\Sigma^2})
\]
Since all the elements of singular value matrix are non-negative, the gradient of the nuclear norm can be written as:
\[
\frac{\partial L_{\text{rank}}}{\partial M} = \frac{\partial \text{tr}(\Sigma)}{\partial M} = UV^T
\]
Gradient of the low rank loss can be back-propagated into the generator according to equation 5

![Figure 4: Ranks of some typical mask matrixes. The size of the above matrixes is 128 × 128. In these matrixes, black represents -1 and white represents 1. From (a) to (e) we notice that the more complex the mask is, the larger its rank value will be. Thus low-rank loss regularizes disentangled masks since the human labeled mask won’t be much too complicated.](image)

#### Pixel-wise Reconstruction Loss

In our approach, the pixel-wise loss is adopted to accelerate optimization and boost the superior performance. For the generated image \(I_{\text{gen}}\) and the target image \(I_{\text{target}}\), the pixel-wise loss is written as:
\[
L_{\text{rec}} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |I_{\text{gen},x,y} - I_{\text{target},x,y}|^p
\]
where \(p\) is set as 1 for \(L1\) loss or 2 for \(L2\) loss. The pixel-wise reconstruction loss is measured at five generated image: the inferred facial geometry images \(I^g_{f}\) and \(I^g_{p}\), the generated complete face \(G^l(I_m)\), the disentangled mask \(G^m(I_m)\), and the reconstructed masked face image \(I_r\).

#### Adversarial Loss

The adversarial losses for both the complete face image and its masked region are calculated. The adversarial loss pushes the synthesized images to reside in the manifold of target image distribution. We denote \(N\) as batch size, the adversarial loss for the complete face image is calculated as:
\[
L_{\text{adv}}^g = \frac{1}{N} \sum_{n=1}^{N} - \log D^g(G^l(I_m))
\]
Similarly, the adversarial loss for missing region is formulated as:
\[
L_{\text{adv}}^l = \frac{1}{N} \sum_{n=1}^{N} - \log D^l(\text{crop}(G^l(I_m)))
\]
The overall adversarial loss is:
\[
L_{\text{adv}} = \alpha L_{\text{adv}}^g + \beta L_{\text{adv}}^l
\]
The parameters \(\alpha\) and \(\beta\) are set as 1.

#### Symmetric Loss

Symmetry is the prior geometry knowledge widely existing in human faces, especially reflected on frontal face images. To preserve such a symmetrical structure, a symmetric loss is imposed to constraint the synthesized face images and accelerate the convergence of our algorithm. For a face completion result \(I^g\), its symmetric loss takes the form:
\[
L_{\text{sym}} = \frac{1}{W/2 \times H} \sum_{x=1}^{W/2} \sum_{y=1}^{H} |I^g_{x,y} - I^g_{W-(x-1),y}|
\]
However, the limitation of the above pixel-level symmetric loss is obvious and can be divided into three folds: the illumination changes, the intrinsic texture difference, and human face poses. Hence its weight in the overall loss is not heavy.

#### Overall Loss

The final object loss to be optimized is the weighted sum of all the loss mentioned above:
\[
L = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{adv}} + \lambda_3 L_{\text{rank}} + \lambda_4 L_{\text{sym}} + \lambda_5 L_{\text{tv}}
\]
where \(L_{\text{tv}}\) is a regularization term on generated face images to reduce spiky artifacts (Johnson, Alahi, and Fei-Fei 2016).

### 4 Experiments and Analysis

#### 4.1 Experimental Settings

**Datasets.** We evaluate our model under both controlled and in-the-wild settings. To this end, two publicly available
datasets are employed in our experiments: Multi-PIE (Gross et al. 2010) and CelebA (Liu et al. 2015b). The Multi-PIE is established for studying on the PIE (pose, illumination, and expression) invariant face recognition. It consists 345 subjects captured in controlled environments. The Multi-PIE dataset contains face images with variational face poses, illumination conditions, and expressions. We choose face images with the frontal view and balanced illumination, resulting in 4539 images of 345 subjects. We use images from the first 250 subjects for training and the rest for testing, and there is no overlap between the training and testing sets. Thus, the training set contains 3627 images belonging to 250 individuals, and the testing set contains 912 images belonging to 95 individuals. The face regions are cropped by their eyes locations and resized to $128 \times 128$.

CelebA consists of 202,599 celebrity images with large variations in facial attributes. These images are obtained from unconstrained environments. The standard split for CelebA is employed in our experiments, where 162,770 images for training, 19,867 for validation and 19,962 for testing. Following (Radford, Metz, and Chintala 2016), we crop and roughly align all the face images by the locations of the centers of eyes.

**Implementation Details.** We use colorful images of size $128 \times 128 \times 3$ in all the experiments. The rectangle masks are randomly selected with mean shape $64 \times 64$ for training so that at least one sense organ is under the obstruction. Whereas, the position is randomly selected to prevent the model from putting too much attention on completing a certain facial part. The images are random flipped horizontally by probability 0.5. We set the learning rate as 0.0002 and deploy Adam optimizer (Kingma and Ba 2015) for the facial geometry estimator, the generator and the two discriminators. The FCENet is trained for 200 epochs on the Multi-PIE and 20 epochs on the CelebA. The weights $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, $\lambda_5$ in overall loss are set as 10, 1, 0.001, 0.01, 0.0001 in practice, respectively. We apply end-to-end training in FCENet and facial attributes manipulation is only conducted on testing phase.

**Ground Truth Facial Geometry Extractors.** Inferring facial landmark heatmaps and parsing maps is a significant step in the FCENet. In our algorithm, Facial heatmaps of 68 landmarks and parsing maps of 11 components are used to supervise the facial geometry estimator. But these two facial geometry information is not provided in the Multi-PIE and CelebA datasets. Therefore, we deploy open source state-of-the-art face alignment (Bulat and Tzimiropoulos 2017) and face parsing (Liu et al. 2015a) tools to extract facial landmarks and parsing maps from original complete face images as the ground truth geometry, then the facial geometry estimator is trained to recover these ground truth facial geometry from masked input face images.

### 4.2 Comparison Results

To demonstrate the effectiveness of our algorithm, we make comparisons with several state-of-the-art methods: Patch-Match (PM) (Barnes et al. 2009), Context Encoder (CE) (Pathak et al. 2016), Generative Face Completion (GFC) (Li et al. 2017) and Semantic Image Inpainting (SII) (Yeh et al. 2017). The performances are evaluated both qualitatively and quantitatively. Specifically, three evaluation metrics are considered: visual quality, Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM). PSNR directly measures the difference in pixel values and SSIM estimates the holistic similarity between two images.

| Method | PSNR | SSIM | PSNR | SSIM |
|--------|------|------|------|------|
| PM     | 29.960 | 0.906 | 29.591 | 0.906 |
| CE     | 23.052 | 0.678 | 24.499 | 0.732 |
| SII    | 19.366 | 0.682 | 18.963 | 0.685 |
| GFC    | 27.208 | 0.889 | 24.281 | 0.836 |
| Ours   | 27.714 | 0.904 | 24.944 | 0.871 |

Table 1: Quantitative results on the Multi-PIE and CelebA testing sets. Higher values are better.

| Dataset | Ours>PM | Ours>SII | Ours>CE | Ours>GFC |
|---------|---------|----------|---------|----------|
| MultiPIE | 100% | 100% | 88.6% | 98.1% |
| CelebA | 100% | 100% | 85.7% | 87.4% |

Table 2: User study results. The rate where our result is preferred is listed. 50% means two methods achieve same performance.

**Quantitative Results.** Comparisons on PSNR and SSIM are shown in table 1. For fair comparisons, we retrain the CE model on the Multi-PIE and CelebA since it is not trained for face completion, and the GFC model is retrained on the Multi-PIE. The publicly available SII implementation is trained only on CelebA and doesn’t support training on other datasets. We find these indexes are basically consistent with the visual quality except PM. However, images generated by PM don’t have realistic visual quality. We conduct user study experiment as table 2 shown to better compare these methods. On quantitative results, our FCENet outperforms SII, CE, and GFC. Compared to general image completion models(CE, SII), performances of models considering facial domain knowledge(GFC, ours) are substantially better. We also investigate performance improvement from GFC to FCENet by comparing their networks. Our FCENet exploits facial geometry as guidance information for face completion while GFC treats them as semantic regularization. More valid facial geometry information is utilized by feeding them into the face completion generator than by regularizing face semantic loss. Our FCENet makes use of facial landmark heatmaps and parsing maps that carry richer facial geometry information than GFC which only deploys facial parsing maps.

**Qualitative Results.** The visual comparisons are summarized in figure 5 and 6. We mainly show typical examples with masks covering key face components (e.g., eyes, mouths, and noses). The PM method searches the most similar patches from the masked input and fails to complete appealing visual results. It is not surprising because the miss-
Figure 5: Visual quality of completion results on the MultiPIE. Our completion results vastly outperform other methods on visual quality.

Figure 6: Visual quality of completion results on the CelebA. For SII, we use its public implementation on CelebA to complete face images. Note that SII completes face images cropped at the center to $64 \times 64$.

Figure 7: Hard cases selected from the CelebA. The face images in the three rows pose difficulty of exceptional context (red background), inpainting face with occlusion (hat) and posed face. Our FCENet generates visually realistic face images while reconstruction artifacts can be easily observed in PM, SII, CE and GFC.

Figure 8: Some failure cases. Image completed by our method still achieve the better visual quality.

Figure 9: Visual quality of generated images. The visual quality of testing results under different experiment settings is demonstrated as figure 10.

From table 3 and figure 10, our observations are fivefold. (1) The low-rank regularization term improves the face completion performance slightly. In our testing protocol, complete face images overlaid by square masks are set as input. This result is reasonable since only pixel-wise reconstruction loss may have already been sufficient for disentanglement of rectangle masks. The low-rank loss will play a much more important role in filling regions with irregular shape and size. (2) The performance of experiment setting 3, 4, 5 are substantially higher than the performance of experiment setting 2, which validates the effects of facial geometry guidance on face completion task. (3) When facial landmark heatmaps and parsing maps are simultaneously exploited, the best testing results are acquired. The performance under setting 5 is better than setting 3 and 4 can be attributed to two aspects: Multi-task training (facial landmark heatmaps and parsing maps) forces the stacked HG blocks

4.3 Ablation Study

Effects of Low-rank Regularization and Different Facial Geometry Images. To validate the effects of the low-rank loss, estimated facial landmark heatmaps and facial parsing maps. Different experiment settings are applied as table 3 presents.

The experiment setting 1 and 2 are used to verify the effect of the low-rank regularization. The experiment setting 3, 4 and 5 respectively reflect the performance boosting of facial landmark heatmaps, facial parsing maps as well as both together in comparison to experiment setting 2 that doesn’t contain any facial geometry. Two metrics, PSNR and SSIM are calculated at the testing set as shown in table 3. Figure 9 demonstrates that the low-rank regularization slightly improves visual quality of generated images. The visual quality of testing results under different experiment settings is demonstrated as figure 10.
Table 3: Different experiment settings and their testing results (PSNR, SSIM) on the Multi-PIE for ablation study, higher values are better.

| Exp. Setting | 1       | 2       | 3       | 4       | 5       |
|--------------|---------|---------|---------|---------|---------|
| Low-rank Loss| ×       | ✓       | ✓       | ✓       | ✓       |
| Landmarks    | ×       | ×       | ✓       | ×       | ✓       |
| Parsing Maps | ×       | ×       | ×       | ✓       | ✓       |
| PSNR         | 24.926  | 25.470  | 27.256  | 27.482  | 27.714  |
| SSIM         | 0.841   | 0.845   | 0.898   | 0.900   | 0.904   |

Figure 9: Visual quality of completion results with and without low-rank loss.

to produce face feature maps that are more representative. In addition, more facial geometry information is exploited for completing masked face image under setting 5. (4) The testing performance under setting 4 outperforms that under setting 3. In comparison with the facial landmark heatmaps, the facial parsing maps contain finer granularity information for face completion thus bring much more improvement. (5) Comparing performance under setting 5 to that under setting 4, the performance improvement is marginal. Such a result is caused by information redundancy between facial landmarks and parsing maps.

Upper Boundary of Performance Improvement In this section, we study the upper boundary of performance improvements that different facial geometry images bring. In these experiments, the facial geometry estimator is removed since the landmark heatmaps or parsing maps are directly obtained from the original complete face images. The experiments about upper boundary of performance improvement are conducted on setting 3, 4 and 5 to evaluate upper bound of performance improvements brought by landmarks, parsing maps and both together respectively. The testing results on the Multi-PIE dataset are presented as table 4:

The above results demonstrate the upper bound performance improvement that FCENet achieve if the facial geometry is estimated perfectly. Under the same experiment setting, our algorithm that adopts stacked HG blocks is able to achieve performance comparable with corresponding upper boundary. Thus, the proposed facial geometry estimator excels at recovering complete facial geometry information such as facial landmark heatmaps and parsing maps from the masked input. Table 4 shows that our FCENet even outperforms its upper bound performance on both PSNR and SSIM under setting 3 and 4, which proves that FCENet is robust to inaccurate guidance information inferred by facial geometry estimator when FCENet strives to exploit them effectively.

4.4 Face Attributes Manipulation

After filling the masked regions, users may not be satisfied with the face attributes of these parts and tend to modify them. Our estimated facial geometry images allow the users to perform interactive face attributes editing on the resulting complete face. Examples are presented in figure 1. Our algorithm also supports face attributes modifications for complete face images by manually pasting a white mask on the unsatisfied region of a complete face image. Then operations like copy-and-paste are used to modify facial geometry images with desired attributes. Hence we can simply change the facial geometry images to create novel human portraits. For example, changing the shape of generated faces by simply editing their geometry images (row 1 and 2 in figure 11), modifying one attribute while other attributes are kept similar to the target image (row 2, 3 and 4 in figure 11).

Figure 10: Visual quality of completion results under different experiment settings. (a) and (g) are masked input and ground truth complete face images respectively, (b)-(f) are completion results on the Multi-PIE testing set under experiment setting 1-5 respectively. Effects of low-rank regularization term and facial geometry are presented intuitively.

Table 4: Performance comparison when the guidance is inferred facial geometry or ground truth facial geometry, upper bound means ground truth facial geometry is applied to the face completion generator.

| Exp. Settings | PSNR    | SSIM    |
|---------------|---------|---------|
| Setting 3/upper bound | 27.256 / 27.221 | 0.898 / 0.896 |
| Setting 4/upper bound | 27.482 / 26.494 | 0.900 / 0.872 |
| Setting 5/upper bound | 27.714 / 28.776 | 0.904 / 0.916 |
Figure 11: Attributes editing results. (b) and (c) are modified facial geometry. The face shapes in the 1st and 2nd rows are changed. Row 3, 4 and 5 present that nose, mouth and eyes are edited respectively while other attributes are nearly unchanged.

5 Conclusion

This work has systematically studied facial geometry for face completion and editing, resulting in a new geometry-aware network, named FCENet. Our network is composed of a face geometry estimator for inferring facial geometry from a masked face, a generator for inpainting face images and disentangling their masks. A low-rank loss is imposed to regularize the mask matrix to denoise. Different from most existing face completion methods, our network naturally supports facial attributes editing by interactively modifying facial geometry images. Extensive experimental results on the widely used Multi-PIE and CelebA datasets demonstrate FCENet's superior performance over state-of-the-art face completion methods.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. In ICML.

Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. TOG 28(3):24.

Bertalmio, M.; Sapiro, G.; Caselles, V.; and Ballester, C. 2000. Image inpainting. In SIGGRAPH.

Bertalmio, M.; Vese, L.; Sapiro, G.; and Osher, S. 2003. Simultaneous structure and texture image inpainting. TIP 19(3):340–358.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In ICCV.

Chen, Y.; Shen, C.; Wei, X.-S.; Liu, L.; and Yang, J. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In ICCV.

Criminisi, A.; Pérez, P.; and Toyama, K. 2004. Region filling and object removal by exemplar-based image inpainting. TIP 13(9):1200–1212.

Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D. B.; and Sen, P. 2012. Image melding: Combining inconsistent images using patch-based synthesis. TOG 31(4):82–1.

Dolhansky, B., and Ferrer, C. C. 2018. Eye in-painting with exemplar generative adversarial networks. In CVPR.

Elad, M.; Starck, J.-L.; Querre, P.; and Donoho, D. L. 2005. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). ACHA 19(3):340–358.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In NIPS.

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; and Baker, S. 2010. Multi-pie. IVC 28(5):807–813.

Hays, J., and Efros, A. A. 2007. Scene completion using millions of photographs. TOG 26(3):4.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR.

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018. Pose-guided photorealistic face rotation. In CVPR.

Huang, J.-B.; Kang, S. B.; Ahuja, N.; and Kopf, J. 2014. Image completion using planar structure guidance. TOG 36(4):107.

Huang, R.; Zhang, S.; Li, T.; He, R.; et al. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In ICCV.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2017. Globally and locally consistent image completion. TOG 36(4):107.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In ECCV.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In ICLR.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In CVPR.

Li, C., and Wand, M. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In CVPR.

Li, Y.; Liu, S.; Yang, J.; and Yang, M.-H. 2017. Generative face completion. In CVPR.

Liu, S.; Yang, J.; Huang, C.; and Yang, M.-H. 2015a. Multi-objective convolutional learning for face labeling. In CVPR.
Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015b. Deep learning face attributes in the wild. In ICCV.
Liu, P.; Qi, X.; He, P.; Li, Y.; Lyu, M. R.; and King, I. 2017. Semantically consistent image completion with fine-grained details. In arXiv preprint arXiv:1711.09345.
Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In ICCV.
Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In ECCV.
Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In CVPR.
Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR.
Shen, W., and Liu, R. 2017. Learning residual images for face attribute manipulation. In CVPR.
Yeh, R. A.; Chen, C.; Lim, T. Y.; Schwing, A. G.; Hasegawa-Johnson, M.; and Do, M. N. 2017. Semantic image inpainting with deep generative models. In CVPR.
Zhang, S.; He, R.; Sun, Z.; and Tan, T. 2016. Multi-task convnet for blind face inpainting with application to face verification. In ICB.
Zhang, S.; He, R.; Sun, Z.; and Tan, T. 2018. Demeshnet: Blind face inpainting for deep meshface verification. TIFS 13(3):637–647.