arXiv:1812.03252v1 [cs.CV] 8 Dec 2018

Face Completion with Semantic Knowledge and Collaborative Adversarial Learning

Haofu Liao[1*], Gareth Funka-Lea[2], Yefeng Zheng[3*], Jiebo Luo[1], and S. Kevin Zhou[4*]

[1] Department of Computer Science, University of Rochester, Rochester, USA
hliao6@cs.rochester.edu
[2] Digital Technology and Innovation, Siemens Healthineers, Princeton, USA
[3] Tencent X-Lab, Shenzhen, China
[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

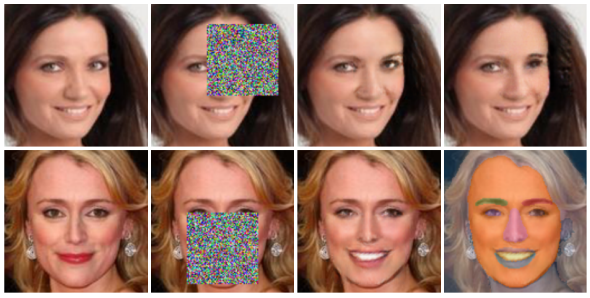
Abstract. Unlike a conventional background inpainting approach that infers a missing area from image patches similar to the background, face completion requires *semantic knowledge* about the target object for realistic outputs. Current image inpainting approaches utilize generative adversarial networks (GANs) to achieve such semantic understanding. However, in adversarial learning, the semantic knowledge is learned implicitly and hence good semantic understanding is not always guaranteed. In this work, we propose a *collaborative adversarial learning* approach to face completion to explicitly induce the training process. Our method is formulated under a novel generative framework called collaborative GAN (collaGAN), which allows better semantic understanding of a target object through collaborative learning of multiple tasks including face completion, landmark detection and semantic segmentation. Together with the collaGAN, we also introduce an inpainting concentrated scheme such that the model emphasizes more on inpainting instead of autoencoding. Extensive experiments show that the proposed designs are indeed effective and collaborative adversarial learning provides better feature representations of the faces. In comparison with other generative image inpainting models and single task learning methods, our solution produces superior performances on all tasks.

Keywords: Face completion · Image inpainting · Generative Adversarial Networks · Multitask Learning.

1 Introduction

Image inpainting is the process of reconstructing a missing region in an image such that the inpainted area is visually consistent with its neighboring pixels and the inpainted image overall looks realistic. Traditional approaches to this problem either require that the filling information is available in the image [6,9,1] or rely on the availability of a large photo database to retrieve the missing

[*] The work was done when the authors were with Siemens Healthineers.



(a) (b) (c) (d)

Fig. 1: Inpainting results using our method. (a) original image. (b) masked image. (c) inpainted image with collaborative learning. (d) inpainted image without collaborative learning (top) and segmentation mask from our method superimposed on the inpainted image (bottom).

region [32,5]. These approaches work well for the *background inpainting* problem, i.e., filling the missing part with image patches similar to its background, as inpainting can be performed through pattern matching. However, when it comes to complete a missing part of an object where no existing patches can be matched or retrieved, the traditional approaches may fail. For example, if a mouth is missing, it is not possible to synthesize the mouth using image patches from other face parts. Instead, image inpainting in this case requires *semantic knowledge* about faces, e.g., location, shape, color and texture of face parts.

To address this *object completion* problem in image inpainting, recent models [25,34,20] propose to use generative adversarial networks (GANs) for more semantically consistent results. However, for generative models, the semantic understanding is implicitly learned through adversarial training. There are no direct constraints on the structure of the target object and hence the inherent semantic understanding is not always guaranteed. Fortunately, in recent years, the success of deep learning has made the semantic labels of objects accessible. In this work, we investigate the possibility of introducing the semantic knowledge of face labels to the adversarial training of face completion for better induction of semantic understanding.

We focus on helping the inpainting model better understand the underlying structure of faces through the collaborative learning of other face related tasks. We argue that current approaches using generative inpainting models alone may not be able to produce structurally realistic results in some cases. For example, when an eye is missing from the image, the inpainting model should be able to predict the missing eye's location and shape based on the facial symmetry. However, as shown in the first row of Figure 1, the generative image inpainting model trained without using our proposed collaborative method produces a structurally unrealistic face (Figure 1(d) top) with the inpainted eye smaller and darker than the eye outside the corrupted region. In contrast, a collabora-

tively trained model can keep the structural consistency between the inpainted region and the nearby context (Figure 1(c) top). In addition, we also find that models trained in this manner tend to produce visually consistent results among tasks. As demonstrated in Figure 1(d) bottom, the segmentation result is closely aligned with the inpainting result other than the ground truth. This provides a clear evidence that they are inherently helping each other during training and the knowledge is shared instead of individually learned.

To this end, we propose an innovative image-to-image generative network for face completion. The proposed method formulates a collaborative GAN to facilitate the direct learning of multiple tasks. For the generator, the network outputs multiple channels for each task and has them share most of the network parameters for better collaborative learning. We also stand apart from the existing inpainting models by introducing skip connections between the encoder and decoder [29,12]. For the discriminators, we apply conditional GAN (cGAN) [24] for better transformation quality and have dedicated discriminators for each task. For the loss function, we introduce an inpainting concentrated scheme to allow the model focusing on the inpainting itself instead of autoencoding the context. Our experimental results demonstrate the effectiveness of the proposed design and better feature representations can be obtained with the proposed collaborative GAN. Comparing with other generative models without using collaborative adversarial learning, our approach consistently produces remarkably more realistic inpainting results. Comparing with single task adversarial learning, our joint approach produces better performances on all tasks.

2 Related Work

Generative Image to Image Transformation Image inpainting can be seen as a special case of the image-to-image transformation problem in that image inpainting tries to transform a cropped image to a reconstructed one. One of the typical image-to-image transformation problems is autoencoding [8]. In relation to inpainting, a seminal work in this area is the denoising autoencoder [31]. It introduces an inpainting-like scheme to autoencoding, hoping the autoencoder can learn a better feature representation by recovering the damage to the input image. Another related work is [16]. It incorporates GANs into a variational autoencoder (VAE) [14] and argues that the network trained using the adversarial loss and VAE loss can give a more representative feature vector of the input image.

For other image-to-image transformation problems, [18] proposes to use GANs for image super-resolution. It has the standard generative image-to-image model setting: a generator that maps the low resolution input to high resolution and a discriminator that tells if the input image is a high resolution one or not. It also includes perceptual loss [18] to further regularize the realism. [12] proposes a general framework for image-to-image translation. It improves the generative image-to-image networks by introducing image-conditional GANs and PatchGAN.

Semantic Inpainting The term *semantic inpainting* is first introduced by context encoder (CE) [25] to address the challenging inpainting case where a large region of the image is missing and the inpainting generally requires semantic understanding of the context. The paper proposes to learn a better autoencoder by having it recover the missing part of an input image. To our best knowledge, it is the first work that introduces adversarial loss into the inpainting problem. Another approach to semantic inpainting is [34]. The method is based on a pretrained GAN model that maps a noise vector to the generator manifold. The algorithm finds the noise vector such that the generated image through the pretrained GAN model minimizes both a contextual loss and a perceptual loss. This work is an indirect approach to image inpainting. Due to the limitation of the pretrained GAN model, it may fail to produce good results when the image resolution is high or the scene is complex. [33] proposes a multi-scale neural patch synthesis network for high-resolution image inpainting. Since our work focus on the benefit of collaborative adversarial training, high-resolution image inpainting is not within the cope of this study. In fact, our work can be easily extended to high-resolution settings for better performance.

The closest work to ours is [20]. It advances the state of the art [25] by introducing a parsing network to further regularize the inpainting through semantic parsing/segmentation and a local discriminator to ensure the generated contents are semantically coherent. However, the parsing network is pretrained and independent of the generator. Thus, the semantic parsing information is not shared with the generator during training. The semantic parsing loss is also not applied directly to the generator. As a result, the accuracy of computing such a loss is limited by the parsing network which further limits the final performance of the generator. Meanwhile, for the local discriminator, it requires the mask to be rectangular and hence cannot be generalized to other inpainting cases such as noise inpainting. [10] also proposes a similar global and local discriminator design and it suffers from the same disadvantage as [20].

Multi-Task Learning Multi-task learning (MTL) [2] refers to the process of learning multiple tasks jointly in order to improve the generalization performance of the model. It has been widely used in deep neural networks (DNNs) for various tasks, such as object detection [4], action recognition [30], landmark detection [35], etc. In terms of generative models, many works have introduced adversarial learning in a multi-task fashion to improve the learning of the main tasks. [3] uses adversarial learning for better domain adaption of the main task. [19] proposes a perceptual GAN that learns super-resolution together with object detection for better performance of small object detection. [21] leverages GANs to generate shared features that are independent of different text classification tasks. In this work, we contribute to the literature with a novel image-to-image generative framework that collaboratively learns multiple tasks for better semantic understanding and, ultimately, yields better image inpainting.

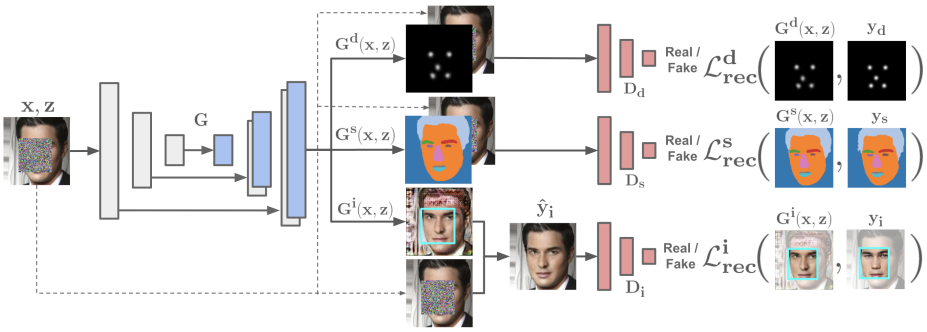


Fig. 2: The architecture of the proposed method. The network is trained collaboratively with three tasks: inpainting i, segmentation s and (landmark) detection d. The generator G takes a masked image x as input and outputs the inpainted image $G^i(x, z)$, segmentation mask $G^s(x, z)$ and detection heatmap $G^d(x, z)$ simultaneously. The discriminators D_i, D_s and D_d are used for adversarial learning. In addition, reconstruction losses $\mathcal{L}^i_{rec}$, $\mathcal{L}^s_{rec}$ and $\mathcal{L}^d_{rec}$ are also applied to the three tasks, respectively.

3 Collaborative Face Completion

The proposed collaborative face completion method is formulated under a novel GAN framework which we call *collaborative GAN* (collaGAN). The proposed framework aims to inductively improve the main generation task (face completion in our case) by incorporating the additional knowledge embedded in other tasks. In this section, we give the formal definition of collaGAN and its connection to face completion.

3.1 Collaborative GAN

Let Ω be a finite set of tasks and y_t be a data sample of task $t \in \Omega$. A collaGAN learns a mapping from an input image x and a random noise z to a set of outputs $\{y_t|t \in \Omega\}$, i.e., $G : \{x, z\} \rightarrow \{y_t|t \in \Omega\}$. The network is trained in an adversarial fashion. The generator G tries to generate data samples as "real" as possible such that, $\forall t \in \Omega$, the adversarially trained discrimintor D_t cannot tell if a sample is generated by G or from the data domain of task t.

Similar to the classic GAN, the objective function of a collaGAN can be given as follows:

$$\begin{aligned}\mathcal{L}^{\Omega}_{adv} = \min_{G} \max_{D_\Omega} \sum_{t \in \Omega} \mathbb{E}_{x, y_t \sim p_{data}(x, y_t)}[\log D_t(x, y_t)] \\ + \lambda^t_{adv} \mathbb{E}_{x \sim p_{data}(x), z \sim p_z(z)}[1 - \log D_t(x, G^t(x, z))],\end{aligned} \tag{1}$$

Here, $D_\Omega = \{D_t|t \in \Omega\}$. p_{data} and p_z denote the data and noise distribution, respectively. $G(x, z)$ is the channel-wisely stacked generator outputs of the tasks

with $G^t(x, z)$ denoting the ouput generated for task t. λ^t_{adv} balances the importance of the generator loss for task t. D_Ω and G play a minmax game where D_Ω tries to maximize the objective and G tries to minimize it.

In this work, as shown in Figure 2, $\Omega = \{i, s, d\}$, denoting three tasks of *inpainting, segmentation, and (landmark) detection*. x is the occluded image, y_i is the original face, y_s is the segmentation mask and y_d is the detection heatmap (See Section 4.1). $D_\Omega = \{D_i, D_s, D_d\}$ where D_i, D_s and D_d are the discriminators for the inpainting, segmentation and detection tasks, respectively. $G(x, z) = [G^i(x, z), G^s(x, z), G^d(x, z)]$ are the generator outputs for the three tasks.

Note that instead of only applying adversarial loss to the inpainted image, we have dedicated discriminators for each of the tasks. Our design follows the observation that GANs can also be helpful in some non-generative tasks [23]. In our case, GANs are used to keep the long-range spatial label contiguity of the detection heatmap and the segmentation mask. We have also experimented with an optional setting that feeds the outputs of all the tasks into a single discriminator and discover that the single discriminator selectively ignores the inpainting result as the outputs from other tasks are much easier for judging the realness. Besides, the collaGAN is conditioned on multiple discriminators, one per task. Such a design, on one hand, ensures the perceptual quality of the generated image and, on the other hand, keeps the spatial consistency between the input image and generated image. This choice was also shown to be effective in [12].

3.2 Reconstruction Loss

It has been found by previous approaches [12,25] that mixing the adversarial loss with a reconstruction loss is beneficial to generative image-to-image models. A reconstruction loss gives pixel-level measurement of the errors which is a direct regularization between the output and the ground truth. It can capture the overall structure of the target object but is usually unable to give a sharp output. An adversarial loss, on the other hand, can produce perceptually better result but the output may not be structurally consistent with input. Therefore, we combine these two to achieve both realistic and coherent output.

In our case, the reconstruction loss is computed for the three tasks as denoted in Figure 2. For the inpainting task, the reconstruction loss $\mathcal{L}^i_{rec}$ measures the L1 distance between the inpainted image and the unoccluded image. Here, we use the L1 loss instead of the L2 loss due to the observation that the L2 loss tends to give slightly blurry outputs. For the segmentation mask, we first convert the label mask to a multi-channel map with each channel denoting the binary mask of a label. Then, we apply the L2 loss $\mathcal{L}^s_{rec}$ to the multi-channel map to measure the difference between the network output and the ground truth. We use the L2 loss against the typical cross-entropy loss simply for the ease of implementation as the generator (See section 3.4) produces outputs with values between $(-1, 1)$ which favors a regression loss. In our experiments, we find the L2 loss gives good enough segmentation map for this study. For the detection heatmap, we also

(a) (b) (c) (d)

Fig. 3: Example output of the network trained with the inpainting concentrated scheme. From (a) to (d): (a) original image, (b) masked image, (c) output from the generator, and (d) the final inpainting result by combining (b) and (c).

use the L2 loss $\mathcal{L}^d_{rec}$ as the regularizer adapting the choices from [27,26]. Let $\Omega = \{i, s, d\}$, the total reconstruction loss can be written as

$$\mathcal{L}^{\Omega}_{rec} = \sum_{t \in \Omega} \lambda^t_{rec} \mathcal{L}^t_{rec}, \tag{2}$$

where $\mathcal{L}^t_{rec}$ denotes the reconstruction loss of task t and λ^t_{rec} denotes the importance of each loss. The final objective function is then given by

$$\mathcal{L}^{\Omega} = \mathcal{L}^{\Omega}_{adv} + \mathcal{L}^{\Omega}_{res}. \tag{3}$$

3.3 Inpainting Concentrated Generation

已知和未知区域都要重建

In previous work [20], image inpainting is performed through an autoencoder and the unoccluded region is reconstructed along with the occluded part. Thus, the network spends significant portion of its computing power on autoencoding the already available information while the inpainting itself is not fully addressed. A direct approach [25] to this problem is having the generator only outputs the content within the mask. However, this only works for the case where the masks are rectangular. When random shaped regions are occluded, this approach fails as convolutional neural networks cannot generate non-rectangular images. To address this problem for arbitrary shaped masks, we propose an inpainting concentrated scheme. The scheme consists of two parts: an adversarial part and a reconstruction part. For the adversarial part, we modify the adversarial loss for the inpainting such that the discriminator, instead of judging the realness of the generated image, concentrates on finding the incoherence between the inpainted region and the context. Formally, let M be a binary mask where pixels in the occluded region are 0 and anywhere else are 1. The new adversarial loss for inpainting can be written as

卷积神经网络的局限在于不能处理非矩形区域

$$\begin{aligned} \mathcal{L}^i_{adv} &= \mathbb{E}_{x,y_i \sim p_{data}(x,y_i)}[\log D_i(x, y_i)] \\ &+ \lambda_i \mathbb{E}_{x \sim p_{data}(x), z \sim p_z(z)}[(1 - \log D_i(x, \hat{y}_i))], \end{aligned} \tag{4}$$

where

$$\hat{y}_i = G^i(x, z) \odot (1 - M) + x \odot M. \tag{5}$$

对抗loss 用来判断产生的区域和上下文是否和谐，微调adv loss 和reconstruction loss 使算法集中于判断生成的内容是否合理，但是没有解决非矩形mask 的问题

As demonstrated in Figure 2, $\hat{y}_i$ is nothing but the inpainting output from the generator with the unoccluded region replaced by the ground truth. Such a replacement guarantees that the discriminator does not need to worry about the unrealness of the context and thus the inpainted region is concentrated. For the reconstruction part, we introduce an inpainting concentrated reconstruction loss that only computes the L1 distances within the occluded region. That is,

$$\mathcal{L}_{rec}^{i} = \|y_i \odot (1 - M) - G^i(x, z) \odot (1 - M)\|_1. \tag{6}$$

With this scheme, we make sure that no errors outside the occluded region will be backpropagated to the generator, and therefore the unnecessary autoencoding is not learned. Figure 3(c) shows an example output of the network trained with the proposed scheme. The network produces sharp and realistic results inside the occluded region and produces inferior results for the context region that contributes little to the inpainting. Figure 3(d) is obtained using Equation (5). The inpainted region coherently fits with the context and the image overall looks realistic.

3.4 Network Architecture

For the generator and discriminator, we follow the architecture choices in [28] for stable deep convolutional GANs. Both the generator and discriminator take an input image of size 128×128. For the generator, its encoder has 7 convolution layers with each layer followed by a batch normalization layer [11] and a LeakyReLU [7] layer. The decoder has a symmetric structure with the encoder, except that it uses transposed convolution and ReLU [15]. All the convolutional layers have a 4×4 kernel size with a stride of 2. The output layer of the generator is a tanh function. We also adapt the design suggestion from [12] by adding skip connections between encoder and decoder. Skip connections shuttle low level features directly to decoder without passing through the "bottleneck layers". Such a circumvention is critical to some tasks such as semantic segmentation [29] (which is also included in the collaborative training) and we also find this helpful in improving the coherence between the context and the inpainted region. The discriminator has a similar structure to the encoder, except that it only has 5 convolutional layers.

4 Experimental Results

4.1 Datasets

The dataset used in our experiment is the CelebA [22] dataset. It has $202,599$ face images and we use the official split for training, validating and testing. Unlike the state-of-the-art works [34,20], we do not align the faces according to the eyes. In fact, we find such alignment makes the inpainting much easier for the models as they do not need to semantically learn too much about the locations of the eyes and other face parts. Hence, to avoid overfitting and achieve

better generalization, when cropping the faces, we only guarantee that the eyes, nose and mouth are included and *no alignment is performed.* We also augment the dataset by random shift, scaling, rotation and flipping to further ensure the diversity of the faces during training.

Along with each face image, the CelebA dataset readily provides the locations of 5 face landmarks (the two eyes, nose and the two corners of mouth). We create heatmaps from the landmarks according to the method denoted in [27,26]. The generated heatmaps will be used during collaborative training. For a fair comparison, we obtain segmentation masks for each of the faces using the parsing network provided by [20]. The network is trained based on the Helen [17] dataset and achieves a close to state-of-the-art performance. Hence, it is considered sufficient for this study, which is supported by our experimental results. Also, using computer-generated annotations alleviates the burden of laborious manual annotation effort.

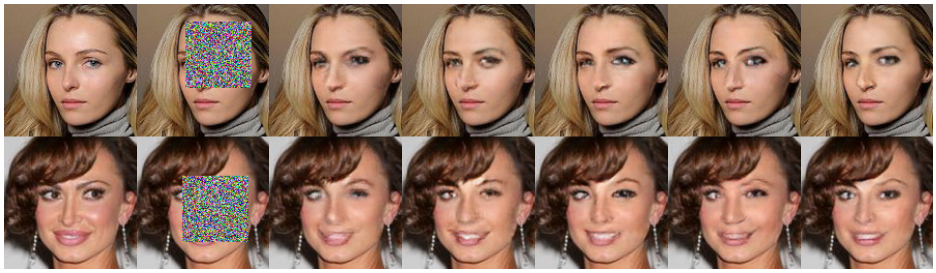
All images in the experiments are resized to 128×128. We use this size choice for a fair comparison with other approaches. It is straighforward for the proposed method to use a large image size. Following the mask generation strategies in previous works, we apply three different masks to the resized images: 1) random block mask with a 64×64 block [20]; 2) random pattern mask [25] with roughly 25% of the pixels missing; 3) random noise mask with *80% of the pixels missing* [34]. For 1) and 3) the masked region is filled with random noise. For 2), the masked region is filled with zeros as the mask itself is already noisy.

4.2 Models

To demonstrate the effectiveness of the proposed method, the performances under different model settings are investigated. We denote M_Ω as the model trained using L^Ω and M^*_Ω as the model trained in addition with the inpainting concentrated scheme. Model settings are changed by varying Ω and switching between M_Ω and M^*_Ω. All the investigated models are trained for 20 epochs. For the optimization, we use the Adam [13] optimizer with a learning rate of 0.01. For the weights of different losses, we emperically find the following settings work well:

- Models with one task: $\lambda^i_{adv} = 1.0$, $\lambda^s_{adv} = 1.0$, $\lambda^d_{adv} = 1.0$, $\lambda^i_{res} = 100$, $\lambda^s_{res} = 1000$, $\lambda^d_{res} = 1000$;
- Models with two tasks: $\lambda^i_{adv} = 0.8$, $\lambda^s_{adv} = 0.2$, $\lambda^d_{adv} = 0.2$, $\lambda^i_{res} = 100$, $\lambda^s_{res} = 200$, $\lambda^d_{res} = 200$;
- Models with three tasks: $\lambda^i_{adv} = 0.8$, $\lambda^s_{adv} = 0.1$, $\lambda^d_{adv} = 0.1$, $\lambda^i_{res} = 100$, $\lambda^s_{res} = 200$, $\lambda^d_{res} = 200$.

For models with one or two tasks, the parameters will be used only when they are applicable. For example, model $\mathrm{M_i}$ will only have $\lambda^i_{adv} = 1.0$ and $\lambda^i_{res} = 100$. Other parameters are not used as they are for $\mathrm{M_s}$ and $\mathrm{M_d}$. Note for this study we are not interested in the best parameter settings for each model. Hence, the parameters are chosen when they work reasonably well. For the λ^i_{adv}, we



(a) original (b) masked (c) M_i (d) $M_{i,d}$ (e) $M_{i,s}$ (f) $M_{i,s,d}$ (g) $M^*_{i,s,d}$

Fig. 4: Qualitative face completion comparison of our models with different settings and varying numbers of tasks.

generally find 0.8 works better in a multi-task scenario and other adversarial loss paramters are chosen such that they sum up to 1. For the reconstruction loss parameters, we find minor performance differences when they are in a reasonable range. In general, setting λ^i_{res} around 100 and λ^s_{res} and λ^d_{res} around 200 gives good performance when several tasks are presented. Setting values close to these numbers give similar performances and the performances are degraded when the values are too large or too small.

4.3 Face Completion

Qualitative Comparison During the experiments, we find in general the models trained with more collaborative tasks produce more realistic results. Figure 4 shows some example outputs of our models. The first row demonstrates that, for an easy inpainting task where context coherence is not critical, the models can all produce relatively good looking faces and the collaboratively trained models from (d) to (g) tend to emphasis more on landmarks (eyes in this case) to give even better inpainting outputs. In the second row, we show a challenging case that only part of the woman's left eye is present and the entire right eye is missing. This inpainting task requires the model to both complete the left eye and reconstruct the entire right eye such that the two eyes together looking realistic. Without collaborative inpainting, the M_i model does not consider the coherence between the two eyes and inpaints the two eyes independently. The other models try to more or less balance the two eyes so that they have the same shape, size and color. The two models trained with three tasks in the last two columns give relatively better inpainting of the eyes. Overall, the $M^*_{i,s,d}$ model trained using the inpainting concentrated scheme gives the most realistic output.

We then compare the proposed method with other state-of-the-art models: CE [25], SII [34] and GFC [20]. Since we use a slightly different face cropping and data augmentation strategy, we retrain those models for a fair comparison. All retrainings are based on their officially released code with the training parameters unchanged. For SII [34], all the experiments are performed using 64×64 images as DCGAN [28] only works well on images with lower resolution. For all the models,

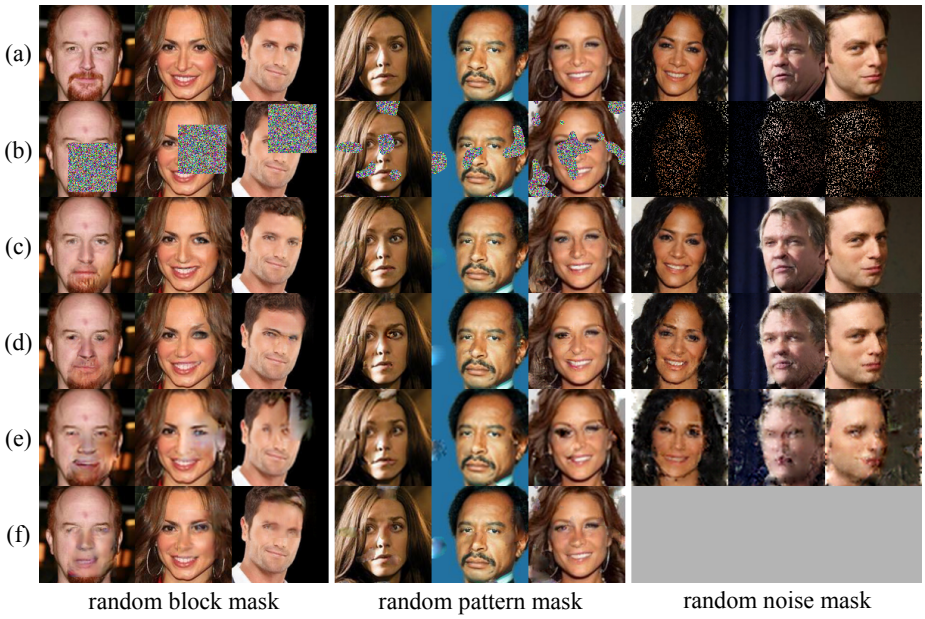


Fig. 5: Qualitative face completion comparison between the proposed method and the state-of-the-art methods. Different masks are applied. From (a) to (f): (a) original image, (b) masked image, and results of (c) the proposed $M^{*}_{i,s,d}$ model, (d) CE [25], (e) SII [34] and (f) GFC [20].

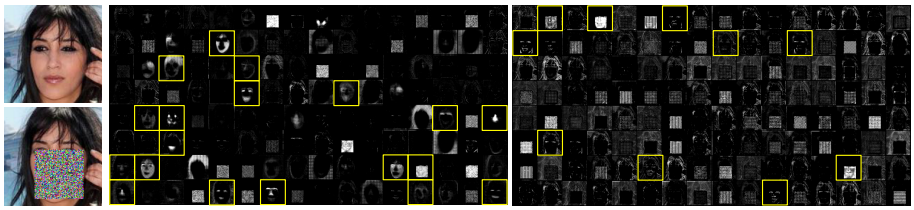
poisson blending is performed. The comparison results are given in Figure 5. To demonstrate the generalizability of the models on various shaped masks, the random pattern masked images in columns 4-6 are inpainted using the models trained with random block masks. For the random noise mask case, since GFC [20] requires square masks for the local loss, it cannot be used to complete random noise masked images. Hence, the inpainting results for GFC [20] in columns 7-9 are omitted. Also, for CE [25] and our method, we train a new model for the random noise mask case due to the uniqueness of the mask.

From Figure 5, we observe that our model consistently gives better inpainting results than the state-of-the-art methods. In general, SII [34] gives the worst results in all the cases. Due to the unaligned faces and the data augmentation we performed during training, the face scene becomes more complex. Thus, the DCGAN model used in SII [34] can not learn the face data distribution very well, which as a result yieds inferior inpainting performance during the inference step. The CE [25] and GFC [20] models in general produce reasonably good results, especially when the faces are aligned. However, as they either train the model without using additional structural constrains or has an indirect measurement of structural inconsistency, their synthesized images are less structurally realistic than ours.

Figure 6 shows the feature maps extracted from the last common layer in the generator G of the $M_{i,s,d}$ and M_i models. The $M_{i,s,d}$ model is more predictive in the masked region and outputs better face related features. It is interesting to notice that the $M_{i,s,d}$ model tends to treat face parts independently. Some maps contain features only about nose, mouse or eyes. While the M_i model mostly outputs features about the whole face. This indicates that the $M_{i,s,d}$ model is more discriminative about the face than the M_i model and it learns to distinguish (and generate) each of the face part due to the training with other tasks.

Quantitative Comparison In addition to the visual comparison, we also quantitatively compare our models with the state-of-the-art models to statistically understand their inpainting performances. All the evaluated models are trained with random block masks. We use two classic metrics, PSNR and SSIM, to evaluate the similarity between the inpainted image and the ground truth. PSNR gives the similarity score at pixel-level while SSIM evaluates the similarity at perceptual level. The ground truth image used in this evaluation is the original image before the occlusion. However, since given an occluded image, there could be multiple inpainting results that are perceptually correct, we realize that neither PSNR or SSIM offers a perfect evaluation. But due to the symmetric nature and the relatively simple structure of human faces, as long as the occlusion does not hide significant portion of the face, all the possible inpainting results should be similar which gives us an opportunity to roughly evaluate the models' performance. Table 1 shows the evaluation results at 6 different mask locations. We adapt the mask location choices from [20] by masking out left eye (O1), right eye (O2), upper face (O3), left face (O4), right face (O5), and lower face (O6), respectively. We intentionally select relatively smaller mask sizes (on average 40×48) to limit the variation of all possible inpaintings. We observe from Table 1 that our base model M_i has already obtained comparable or slightly better performance than the state-of-the-art models which demonstrates the effectiveness of introducing skip connections into collaGAN. In general, models trained with more tasks give better numbers in both PSNR and SSIM. This shows that the knowledge among tasks is indeed collaborative during training. We also find that $M^*_{i,s,d}$ performs slightly better than $M_{i,s,d}$ which means the inpainting concentrated scheme is really helpful in getting better inpainting results. Overall, we spot a significant performance jump with the proposed method when compared with the state-of-the-art models.

To further understand the collaborative nature of the proposed method, we also investigate if the inpainting can help other tasks as well. Specifically, we want to know that, through the learning of inpainting, if the model can predict better of the semantical labels and landmarks in the occluded region. We use the Dice coefficient and localization error to evaluate the performance of semantic segmentation and landmark detection, respectively. The Dice coefficient measures the similarity of two segmentation masks and the localization error computes the Euclidean distances between two landmarks. Table 2 gives the Dice coefficient of different models trained with the segmentation task. The performances of the facial semantic labels are shown. It is clear that *the collab-*



(a) (b) (c)

Fig. 6: Feature maps from the last common layer in the generator G. Face related features are marked with yelow squares. Features from early layers through the skip connection are omitted. (a) The original and masked images, (b) Feature maps of the $M_{i,s,d}$ model, and (c) Feature maps of the M_i model.

Table 1: Quantitative face completion comparison of different models evaluated at 6 different mask locations: left eye (O1), right eye (O2), upper face (O3), left face (O4), right face (O5), and lower face (O6). The numbers in each cell are SSIM (%)/PSNR (dB), the higher the better.

	O1	**O2**	**O3**	**O4**	**O5**	**O6**
CE [25]	90.5/26.74	90.6/27.01	93.8/27.90	95.8/30.37	96.0/30.65	90.0/27.11
SII [34]	87.5/23.93	87.7/24.12	93.0/26.61	95.6/28.94	95.9/29.38	87.8/24.84
GFC [20]	90.6/27.10	90.9/27.34	94.0/28.68	96.3/31.18	96.3/31.11	90.0/27.13
$\mathbf{M_i}$	90.8/27.23	91.1/27.42	94.0/28.53	96.0/30.94	96.1/31.06	90.1/27.15
$\mathbf{M_{i,d}}$	91.7/27.57	91.6/27.81	94.5/28.76	96.4/31.33	96.5/31.36	90.7/27.37
$\mathbf{M_{i,s}}$	91.4/27.59	91.5/27.65	94.3/28.66	96.4/31.22	96.4/31.28	90.7/27.55
$\mathbf{M_{i,s,d}}$	91.7/27.66	91.8/27.87	94.5/28.77	96.5/31.31	96.6/31.39	90.8/27.57
$\mathbf{M^*_{i,s,d}}$	**92.4/27.76**	**92.6/27.96**	**95.2/28.79**	**97.2/31.44**	**97.2/31.50**	**91.7/27.81**

oratively trained models perform better than single task models and the models trained with three tasks better than with two tasks. The $M_{i,s,d}$ and $M^*_{i,s,d}$ models have very close performance. This is reasonable as the inpainting concentrated scheme is designed solely for inpainting and does not introduce additional information to other tasks during training. Note that, for the segmentation task, the performance difference of the models is not significant. We speculate that this is because the "ground truth" is computer-generated from the parsing network which sometimes may not give perfect prediction. For the landmark detection task with accurately labeled landmark locations provided by the CelebA dataset, the performance boost of introducing other tasks is more significant as shown in Table 3. For example, the average landmark error is reduced from 2.38 (the M_d model) to 1.72 (the $M^*_{i,s,d}$ model), a 27.7% decrease.

5 Conclusion

We present a novel collaborative GAN framework for face completion. The experimental results suggest that training multiple related tasks together within

Table 2: Semantic segmentation performance of different models. The numbers are given as Dice coeffient (%). Higher numbers are better.

	$\mathbf{M_s}$	$\mathbf{M_{i,s}}$	$\mathbf{M_{i,s,d}}$	$\mathbf{M^*_{i,s,d}}$
Face	93.7	94.1	**94.2**	94.1
Left eyebrow	74.2	74.6	75.0	**75.2**
Right eyebrow	72.3	72.8	**73.8**	73.5
Left eye	70.7	71.9	72.2	**72.7**
Right eye	70.0	70.3	**71.2**	**71.2**
Nose	90.5	90.9	90.9	**91.0**
Upper lip	68.1	67.2	**68.3**	67.9
Teeth	64.2	66.5	**66.8**	66.7
Lower lip	81.4	82.1	**82.8**	82.4
Average	76.1	76.7	**77.2**	**77.2**

Table 3: Landmark detection performance of different models. The numbers are given as localization errors in pixels. Lower numbers are better.

	$\mathbf{M_d}$	$\mathbf{M_{i,d}}$	$\mathbf{M_{i,s,d}}$	$\mathbf{M^*_{i,s,d}}$
Left eye	2.12	1.73	1.61	**1.60**
Right eye	2.29	1.71	**1.59**	1.62
Nose	2.53	2.15	**1.92**	1.93
Left mouth	2.39	1.80	1.74	**1.73**
Right mouth	2.57	1.86	1.77	**1.72**
Average	2.38	1.85	1.73	**1.72**

the proposed framework is beneficial. By infusing more knowledge, the generative model learns better about the inpainting through the knowledge sharing of the segmentation and detection tasks, whose performances are boosted vice versa. We have also found that optimizing directly toward inpainting other than autoencoding produces better inpainting results in an image-to-image network. Finally, we have demonstrated that the proposed method can give superior inpainting performance than the state-of-the-art methods.

Disclaimer: This feature is based on research, and is not commercially available. Due to regulatory reasons its future availability cannot be guaranteed.

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24–1 (2009)
2. Caruana, R.: Multitask learning. In: Learning to learn, pp. 95–133. Springer (1998)
3. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
4. Girshick, R.: Fast r-cnn. arXiv preprint arXiv:1504.08083 (2015)
5. Hays, J., Efros, A.A.: Scene completion using millions of photographs. In: ACM Transactions on Graphics (TOG). vol. 26, p. 4. ACM (2007)
6. He, K., Sun, J.: Statistics of patch offsets for image completion. In: Computer Vision–ECCV 2012, pp. 16–29. Springer (2012)
7. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)

8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006)
9. Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Image completion using planar structure guidance. ACM Transactions on Graphics (TOG) **33**(4), 129 (2014)
10. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (TOG) **36**(4), 107 (2017)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
16. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)
17. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision. pp. 679–692. Springer (2012)
18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016)
19. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: IEEE CVPR (2017)
20. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. arXiv preprint arXiv:1704.05838 (2017)
21. Liu, P., Qiu, X., Huang, X.: Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742 (2017)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (12 2015)
23. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
25. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544 (2016)
26. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 230–238. Springer (2016)
27. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1913–1921 (2015)
28. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
30. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 568–576. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf
31. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103. ACM (2008)
32. Whyte, O., Sivic, J., Zisserman, A.: Get out of my picture! internet-based inpainting. In: BMVC. vol. 2, p. 5 (2009)
33. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, p. 3 (2017)
34. Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses. arXiv preprint arXiv:1607.07539 (2016)
35. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. pp. 94–108. Springer (2014)