CooGAN: A Memory-Efficient Framework for High-Resolution Facial Attribute Editing

Xuanhong Chen[1,2★], Bingbing Ni[1,2†], Naiyuan Liu[1‡], Ziang Liu[1], Yiliu Jiang[1], Loc Truong[1], and Qi Tian[3]

[1] Shanghai Jiao Tong University, China
{chen19910528,nibingbing,acemenethil,jiangyiliu}@sjtu.edu.cn,
{liunaiyuan27,ttanloc}@gmail.com
[2] Huawei HiSilicon, China
{chenxuanhong,nibingbing}@hisilicon.com
[3] Huawei, China
tian.qi1@huawei.com

Abstract. In contrast to great success of memory-consuming face editing methods at a low resolution, to manipulate high-resolution (HR) facial images, *i.e.*, typically larger than 768^2 pixels, with very limited memory is still challenging. This is due to the reasons of 1) intractable huge demand of memory; 2) inefficient multi-scale features fusion. To address these issues, we propose a NOVEL pixel translation framework called *Cooperative GAN*(CooGAN) for HR facial image editing. This framework features a local path for fine-grained local facial patch generation (*i.e.*, patch-level HR, LOW memory) and a global path for global low-resolution (LR) facial structure monitoring (*i.e.*, image-level LR, LOW memory), which largely reduce memory requirements. Both paths work in a cooperative manner under a local-to-global consistency objective (*i.e.*, for smooth stitching). In addition, we propose a lighter selective transfer unit for more efficient multi-scale features fusion, yielding higher fidelity facial attributes manipulation. Extensive experiments on CelebA-HQ well demonstrate the memory efficiency as well as the high image generation quality of the proposed framework.

Keywords: Generative Adversarial Networks, Conditional GANs, Face Attributes Editing

1 Introduction

Recent development in the field of deep learning has led to a rising interest in Facial Attribute Editing (FAE) [8,33,26,18,30,4,5,20,25], *e.g.*, manipulating structural, emotional, semantic attributes (*e.g.*, gender, hair color, age etc.) of a given

---

★ Work done during an internship at Huawei HiSilicon.
† Corresponding author: Bingbing Ni.
‡ Contributed to the work while he was a research assistant at Shanghai Jiao Tong University.
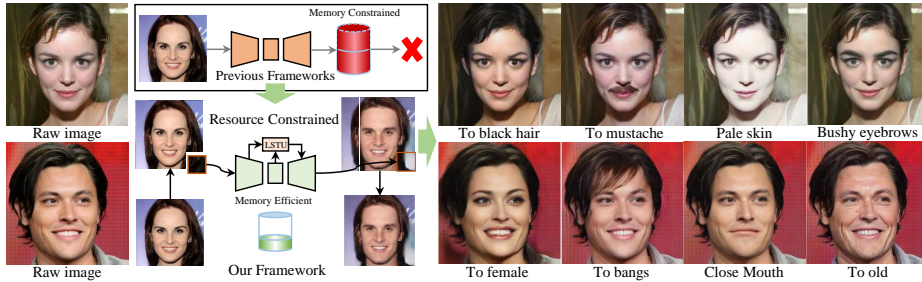
Fig. 1. For the purpose of translating the high-resolution images in the memory-constrained platform, we develop a memory-efficient image translation framework. In the figure, we show some high resolution (768×768) editing results and its manipulation labels, more results can be found in the suppl. From the results, it can be seen that our framework can effectively edit both regional and global attributes.

facial image. These functionalities are highly demanded in emerging mobile applications. However, most of recent facial editing approaches [20,18,26] are based on deep image-to-image translation model, which can ONLY deal with low resolution (LR) facial images due to excessive deep model (*e.g.*, GAN [31,20]) size. Concretely, as the image size increases, the memory consumption of deep models also increases dramatically, even surpasses the limitation of device. Such limitation forbids many promising applications on mobile platforms (*e.g.*, ZAO [1] processing images in the cloud service). Although upgrading hardware can alleviate these issues, it is a non-cheap solution and still hard to process higher resolution images (*e.g.*, 4K). In this work, we aim at developing a memory-efficient deep framework to effectively handle high resolution facial attribute editing on resource-constrained devices.

There are two major challenges of deep model based HR facial attribute editing: 1. *Constrained Computational and Memory Resource.* In some mobile scenarios (*e.g.*, smartphone, AR/VR glasses) with only limited computational and memory resources, it is infeasible to use popular image editing models [18,30] which require sophisticated networks. To address this issue, methods based on model pruning and operator simplifying [35,13,28,34,23] have been proposed to reduce the inference computational complexity. However, the metric-based way to reduce the model size will do harm to model perceptual representation ability, the output facial image quality is usually largely sacrificed. 2. *Inefficient Multi-scale Features Fusion.* In order to achieve high-level semantic manipulation while maintaining local details during image generation, multi-scale features fusion is widely adopted. It is a common practice to utilize skip connection, *e.g.*, U-Net [27]. However, fixed schemes, such as skip connection, usually result in infeasible or even self-contradicting fusion output (*e.g.*, during style transfer, content is well-preserved but failed to change the image style), and flexible schemes such as GRU [9] lead to additional computational burden (*e.g.*, applying GRU-

based unit [20] directly for multi-scale features fusion can achieve excellent fusion effects, but will increase network parameters by more than four times).

In this work, a novel image translation framework for high resolution facial attribute editing, dubbed *CooGAN*, is proposed to explicitly address above issues. It adopts a divide-and-combine strategy to break the **intractable** whole HR image generation task down to several sub-tasks for reducing memory cost greatly. More concretely, the pipeline of our framework consists of a series of local HR patch generation sub-tasks and a global LR image generation sub-task. To handle these two types of sub-tasks, our framework is also composed of two paths, *i.e.*, local network path and global network path. Namely, the local sub-task is to generate an HR patch with edited attributes and fine-grained details. And a global sub-task is to generate an LR whole facial snapshot with structural coordinates to guide the local workers properly recognize the correct patch semantics. As only tiny size patch (*e.g.*, 64×64) generation sub-task is involved in the pipeline, the proposed framework avoids processing large size feature maps. As a result, this framework is very light-weighted and suited for resource constrained scenarios. In addition, a local-to-global consistency objective is proposed to enforce the cooperation of sub modules, which guarantees between-patch contextual consistency and appearance smoothness in fine-grained scale.

Moreover, we design a variant of SRU [19], *Light Selective Transfer Unit (LSTU)*, for multi-scale features fusion. The GRU-based STU [20] has similar functions, but needs two states (one state obtained from encoder, another from the higher level hidden state) to inference the selected skip feature. As a result, it has to face a heavy computing burden and is not friendly to GPU acceleration. Unlike STU, our SRU-based LSTU just need a single hidden state to get the gating signal, which greatly reduces the complexity of the unit. Actually, the LSTU has only half as many parameters as STU and almost the same multi-scale features fusion effect, achieving a good balance between model efficiency and output image fidelity. Under this design, the framework is able to selectively and efficiently transfer the shallow semantics from the encoder to decoder, enabling more effective multi-scale features fusion with constrained memory consumption.

We extensively experiment the proposed framework in terms of both qualitative and quantitative evaluations on facial attribute editing tasks. It is demonstrated that our framework can process 768×768 images well with less than $1GB$ GPU memory rather than over $6.3GB$ (the previous state-of-the-art [20] level). Furthermore, the proposed model has the capability to achieve 84.8% facial attribute editing accuracy on CelebA-HQ.

2 Related Works

Image-to-image translation. Image-to-image translation aims at learning cross-domain mapping in supervised or unsupervised settings [20]. In the early stage, methods like pix2pix use image-conditional GANs to train paired images in the supervised manner. Later, unpaired image-to-image translation frameworks, which work in the unsupervised manner, are also proposed in [36,21].

For instance, CycleGAN [36] learns the mapping between one image distribution and another and introduces a cycle consistency loss to preserve key features during translation. However, all those mentioned frameworks can only perform coarse-grain translation from one image distribution to another. They cannot modify specific attributes of an image, which is more important in real world applications. To address such problem, StarGAN [8] is proposed to learn multiple domain translation using only a single model.

Facial attribute editing is a typical task of image-to-image translation that focuses on editing specific attributes within image. Early methods [15,29,6] are proposed to learn only a single attribute editing model like expression transfer [29], age progression [15], and so on. To achieve arbitrary attribute editing, IcGAN [25] uses a conditional GAN to process encoded features. To strengthen the relations between high-feature space and attributes, Lample G. et al. [17] applies adversarial constraints. StarGAN [8] and AttGAN [12] use target attribute vector input and achieve great success. Based on them, the STGAN [20] is further proposed for simultaneously enhancing the attribute manipulation ability and image quality. Especially, Ran Yi et al. [32] propose a hierarchical facial editing method to modify the face via facial features patches. In our work, we analyze the limitation of StarGAN, AttGAN, [32] and STGAN and then further develop a patch-based model to make facial attribute editing suitable for higher resolution images.

3   Methodology

3.1   Overview

The proposed CooGAN framework for conditional facial generation presents two innovative modules, the global module and the local module. The global module is designed to generate LR translated facial image, and the local module aims at generating HR facial image patches and stitching them together. A cooperation mechanism is introduced to make these two modules work together, so that the global module provides the local module with a global-to-local spatial consistency constraint. In addition, to guarantee the performance and edit-ability of the generated images, we propose a well-designed unit, LSTU, to filter the features from latent space and infuse them with detail information inside the naive skip connection.

3.2   The Cascaded Global-to-Local Face Translation Architecture

The *CooGAN* consists of two interdependent generation modules. We depict the framework architecture in Fig. 2.

Global module. The global module is designed for generating the translated snapshot X_t' which carries the whole image spatial coordinate information. As is shown in Fig. 2, this module contains two components: the global-aware generator $\mathcal{G}_g$ and the global-aware discriminator $\mathcal{D}_g$. $\mathcal{G}_g$ has a conventional U-Net [27] structure strengthening by LSTU. It takes in X_i', the down-sampled
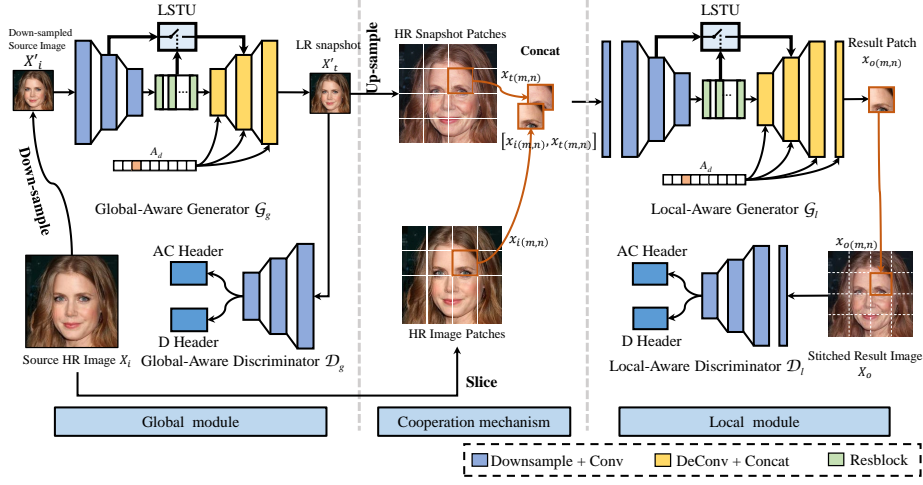
Fig. 2. An overview of our framework. Our framework has two modules–the global image translation module on the left side and the local patch refinement module on the right side. Either module contains one generator and one discriminator, as indicated in the graph. The two modules cooperate with each other through the cooperation mechanism in the middle. All specific operations are elaborated in the graph.

image of X_i, and generates the LR snapshot X'_t. Following the ACGAN-like [24] fashion, the discriminator $\mathcal{D}_g$ has two headers sharing the same extracted features. Note that the main purpose of global module is to guarantee the global semantic consistency of the final result.

Local module. The local module is designed for processing HR image patches under the guidance of LR translated patches with limited computational resources. As is shown in Fig. 2, the local module contains two components: the local-aware generator $\mathcal{G}_l$ and the local-aware discriminator $\mathcal{D}_l$. We concatenate the original HR image patch and up-sampled snapshot patch, *i.e.*, $[x_i(m, n), x_t(m, n)]$, where (m, n) represents the coordinate of the patch in the whole image. Then the local-aware generator $\mathcal{G}_l$ processes the concatenated data sequentially to generate translated HR patches. The generated patches can be stitched together *without overlap* to synthesize the final output X_o. To avoid the inconsistency between generated patches, the local-aware discriminator $\mathcal{D}_l$ is introduced to make the final stitched output smooth and seamless. Note that the discriminator is just introduced temporarily for training our model, so it will not add any memory cost during the inference stage.

Cooperation mechanism. In the previous sections, we discuss the functions of the local module and the global module. The global module requires more computational resources while the size of image increases, which seriously limits its application for generating HR images. The local module decomposes the full-size image into patches and translates the patches into HR ones, which is suitable for processing HR image without adding any memory cost. But the local

module alone cannot correctly perceive the patch semantics and is inapplicable for editing local attributes, as can be seen in Fig. 6. To this end, we introduce a valid cooperation mechanism to encourage the two generators to cooperate well for achieving satisfying generation performance. As shown in the middle of Fig. 2, the global snapshot X'_t is up-sampled to the same size as HR image X_i. Then, we decompose X'_t and X_i into a series of patches and concatenate the patches with the same coordinate. The global patch carries the global spatial coordinate information while the HR image contains detailed textural information. By this way, we combine the global spatial information and detailed textural information together. Namely, the global spatial information is used to make the generated local patch smoother and globally consistent, and the detailed textural information is used to preserve the quality of the generation.

3.3   Objective Function

Our framework contains two sets of adversarial training in global and local module, respectively. For the sake of clarity, we first describe the inference stage of either module. In the global module, we get the snapshot X'_t by the generator $\mathcal{G}_g$ and then feed X'_t into the discriminator $\mathcal{D}_g$. Following [20], we use A_d to facilitate the adversarial training, where A_d denotes the difference attribute vector between the target A_t and the source A_i. The training process can be defined as:

$$X'_t = \mathcal{G}_g \left(X'_i, A_d \right). \tag{1}$$

In the local module, $x_{i(i,j)}$ denotes a patch of HR image X_i and $x_{t(i,j)}$ denotes a patch of up-sampled snapshot image X_t. We use $[\cdot, \cdot]$ to represent the channel-wise patch concatenation,

$$x_{o(i,j)} = \mathcal{G}_l \left(\left[x_{i(i,j)}, x_{t(i,j)} \right], A_d \right). \tag{2}$$

Image reconstruction loss. We apply the reconstruction loss in our framework to improve image quality and avoid as many editing miscues as possible. This loss is achieved in a self-supervision manner, *i.e.*, we feed the framework with zero condition input, which can be depicted as:

$$\mathcal{L}_{rec} = ||x - \mathcal{G}(x, 0)||_1. \tag{3}$$

Adversarial loss. For FAE task, there is no ground-truth for training. Adversarial loss is usually applied due to its domain-to-domain translation nature. To alleviate the instability of adversarial loss in training process, gradient penalty [10] is introduced:

$$\max_{\mathcal{D}_{adv}} \mathcal{L}_{\mathcal{D}_{adv}} = \mathbb{E}_x \left[\mathcal{D}_{adv}(x) \right] - \mathbb{E}_{\hat{y}} \left[\mathcal{D}_{adv}(\hat{y}) \right] + \\ \lambda \mathbb{E}_{\hat{x}} \left[(\|\nabla_{\hat{x}} \mathcal{D}_{adv}(\hat{x})\|_2 - 1)^2 \right], \tag{4}$$

$$\max_{\mathcal{G}} \mathcal{L}_{\mathcal{G}_{adv}} = \mathbb{E}_{x, A_d} \left[\mathcal{D}_{adv}(\mathcal{G}(x, A_d)) \right], \tag{5}$$

where $\mathcal{D}_{adv}$ denotes the adversarial header of the discriminator. x is the input real image, $\hat{y}$ is the generated image and $\hat{x}$ is the sampled point along straight lines between the real image distribution and generated image distribution.

Attribute editing loss. As discussed above, attribute editing has no ground truth. Following the ideology of ACGAN [24], we add a classification header sharing features with adversarial header. Such a design effectively realizes adversarial training along with attribute learning. The attribute editing loss below supervises the two modules of our framework individually,

$$\mathcal{L}_{\mathcal{D}_a} = -\sum_{i=1}^{N_c} \left[a_t^{(i)} \log \mathcal{D}_a^{(i)}(x) + (1 - a_t^{(i)}) \log\left(1 - \mathcal{D}_a^{(i)}(x)\right) \right], \tag{6}$$

where $a_t^{(i)}$ represents the i-th target attribute and $\mathcal{D}_a$ represents the classification header. We use the combination of the above three types of loss to train our model.

3.4 Light Selective Transfer Unit

As mentioned above, we proposed a high-resolution facial image editing framework which can process HR images with limited computational resources. In order to further improve the generation performance, we improve the manner of multi-scale features fusion.

The most popular method for multi-scale features fusion in image-to-image translation is the *Skip Connection* structure. It helps the network balance the contradiction between the pursuit of larger receptive field and loss of more details. One of its classic deployments is U-Net. However, there is a fatal drawback of the original skip connection: it will degrade the function of deeper parts and further damage the effectiveness of condition injection. From Table 1, it is obvious that PSNR increases but attribute editing accuracy decreases when the skip connection number multiplies. A detailed graph showing the editing accuracy of each specific attribute is given in suppl. STGAN [20] tries to alleviate the problem with the STU, a variant of GRU [7,9], which uses the latent feature to control the information transfer in the skip connection through the unit. This feature carries the conditional information added through the concatenation. Unfortunately, such a unit omits the spatial and temporal complexity and it

SC type	SC0	SC1	SC2	SC3	SC4
PSNR	20.2	24.29	28.0	29.9	34.8
SSIM	0.643	0.777	0.877	0.916	0.989
Average Acc	89.5%	82.4%	75.1%	69.1%	63.9%

Table 1. PSNR/SSIM performance and average attribute editing accuracy of models with different Skip Connection (SC) numbers, SCi represents model with i skip connections.

$$\hat{h}^{l+1} = \mathbf{W}_T * [h^{l+1}, A_d]$$

$$\tilde{h}^l = \mathbf{W}_{1\times1} * \hat{h}^{l+1}$$

$$f^l = \sigma(\mathbf{W}_f * \hat{h}^{l+1})$$

$$r^l = \sigma(\mathbf{W}_r * \hat{h}^{l+1})$$

$$\hat{h}^l = (1 - f^l) \circ \tilde{h}^l + f^l \circ x^l$$

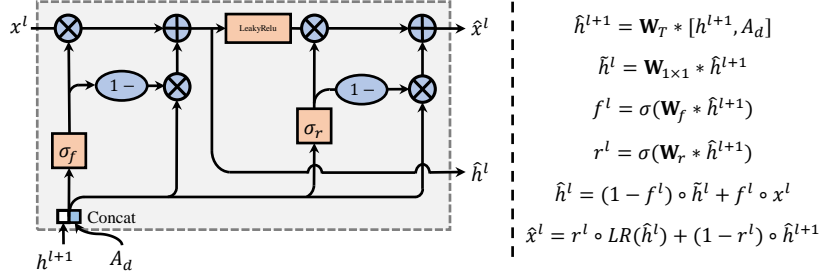$$\hat{x}^l = r^l \circ LR(\hat{h}^l) + (1 - r^l) \circ \hat{h}^{l+1}$$

Fig. 3. The structure of proposed LSTU. The design of LSTU is inspired by SRU, which makes LSTU more lightweight than STU and suitable for GPU parallel acceleration. The right side shows the mathematical expression of the inference process of our LSTU. LR is short for LeakyRelu.

is a time-consuming process for the underlying feature to bubble up from the bottleneck to drive the STU of each layer.

To explicitly address the mentioned problem, we present our framework to employ *Light Selective Transfer Unit (LSTU)* to efficiently and selectively transfer the encoder feature. LSTU is an SRU-based unit with totally different information flow. Compare to STU, our LSTU discards the dependence on the two states when calculating the gating signal, which greatly reduces our parameters but the unit is still efficient. The detailed structure of LSTU is shown in Fig. 3. Without loss of generality, we choose the l-th LSTU as an analysis example. The l-th layer feature coming from the encoder side denotes as x^l. h^{l+1} denotes the feature in the adjacent deeper layer. It contains the filtered latent state information of that layer . h^{l+1} is firstly concatenated with attribute difference A_d to obtain up-sampled hidden state $\hat{h}^{l+1}$. Then $\hat{h}^{l+1}$ is used to independently calculate the masks f^l, r^l for the forget-gate and reset-gate. $\mathbf{W_T}$, $\mathbf{W_{1\times1}}$, $\mathbf{W_f}$ and $\mathbf{W_r}$ represent parameter matrix of transpose convolution, linear transform, forget gate and update gate. The further process is similar to SRU. The equation of gates is shown on the right side of Fig. 3.

4   Experiments

In this section, we take the following steps to evaluate the performance of proposed CooGAN: firstly, we discuss the facial attribute editing performance and the memory usage of our overall framework. Then, we verify the effectiveness of the LSTU in multi-scale features fusion.

4.1   Implement Details

Dataset. We train our model on the CelebA [22] dataset (for the global module) and CelebA-HQ dataset (for the local module). CelebA is a collection of

more than 200,000 human facial images and labels of 40 attributes for each of the images. CelebA-HQ is an artificial high resolution dataset translated from CelebA utilizing the method from [14]. Considering the distinguishability of attributes, in the experiments we mainly use 13 attributes including *Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Male, Mouth Slightly Open, Mustache, No Beard, Pale Skin and Young.*

Optimizer. The optimizer used in our training process is Adam [16] with hyper-parameter $\beta = (0.5, 0.99)$ and learning rate 0.0002. And we apply the learning rate decay strategy with the decay rate 0.1 and decay epoch 100.

Quantitative metric. We design and train an attribute-classifier to quantify attribute editing effects in form of attribute classes accuracy. It takes form of the classifying part in our discriminator and achieves a 96.0% average accuracy on CelebA. And for the texture sharpness of reconstruction results, Peak Signal to Noise Ratio(PSNR) and Structural Similarity(SSIM) are used. They are two common metrics to measure the generated image fidelity. In addition, we use the floating-point operations per second (FLOPS) and model parameter size to evaluate the memory efficiency of networks.

4.2 Experimental analysis of Framework.

Experiment setup. Our framework consists of two modules: global module and local module. In the following experiments, the input image size of the global module is set as 256×256 and the input patch size of local module is 128×128. The final high resolution image size is 768×768. We use the two most effective previous methods (AttGAN, STGAN) and our global module as comparison objects. Since these models are not originally designed for HR image processing, we make proper adjustments of increasing their layers (AttGAN-HR: 7 layers, STGAN-HR: 7 layers, LSTU-net-HR: 7 layers) to expand their perceptual fields. We expand the models based on their official codes and train them with their original loss functions. Those adjusted approaches can be denoted as: AttGAN-HR (AttGAN), STGAN-HR (STGAN) and LSTU-net-HR(global module). The complete model structure setting of our framework, AttGAN-HR and STGAN-HR can be found in suppl. Models are evaluated in the CelebA-HQ dataset.

Quantitative evaluation. We report the attribute editing accuracy of our framework and other methods in Fig. 4. Our framework can achieve 84.8% in classification accuracy, 31.3 in PSNR and 0.923 in SSIM. In terms of memory usage, the largest memory consumption occurs in the global module for the reason that the local module only processes image patches with considerably small

Method	StarGAN	AttGAN	STGAN	Ours
PSNR/SSIM	22.86/0.848	24.1/0.821	31.6/0.934	**32.1/0.938**

Table 2. Reconstruction PSNR and SSIM (CelebA 128×128) of different facial attribute editing frameworks.
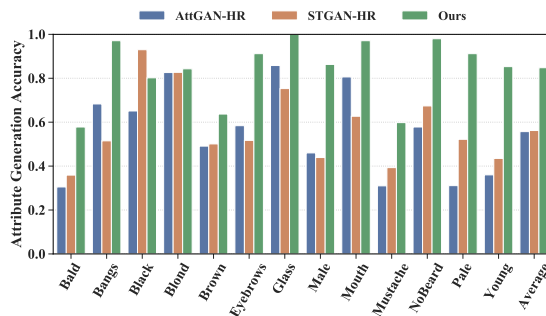
Fig. 4. Our CooGAN against other approaches in attribute translation accuracy. Our framework transcend others in most attributes.

sizes. As is shown in Table 3, while processing 768×768 images the maximum consumption of our framework is below all compared models. This is on the grounds that the peak consumption in our framework happens when the global module processes 256×256 images, while other models process 768×768 images directly. Besides, our framework reduces large memory consumption at the cost of only a little inference time increase, as there is *no overlap* when our framework processes patches, i.e., our advantage. Therefore, the prominent image processing efficiency of our framework is verified.

Qualitative analysis. The visual effect comparison are shown in Fig. 5. In general, our CooGAN can effectively generate high-resolution images with smooth patch boundaries. As is shown, images generated by our CooGAN possess more prominent characteristics as well as better textural quality. For local attributes like bald and mustache, our framework performs much more accurate and significant modification. And in global attributes like gender and age, our CooGAN translates the image more thoroughly and successfully. It is worth mentioning that HR translated images have better effect in some attributes like age and gender. This is because these attribute can not be properly manifested in LR images, *e.g.*, the wrinkles of aging are obvious only in high resolution.

User study. To further validate the subjective performance of our framework, it is important to conduct survey on a crowdsourcing platform to acquire people's general opinions of our HR edited images and compare our CooGAN

Method	AttGAN-HR	STGAN-HR	Ours
GPU Memory	3692MB	6306MB	**985**MB
Inference time	0.0414s	0.0502s	**0.0645**s

Table 3. GPU memory consumption and average inference time (200 images), measured on a 2080ti, for different framework with batch size 1, input images at the 768×768 pixels.

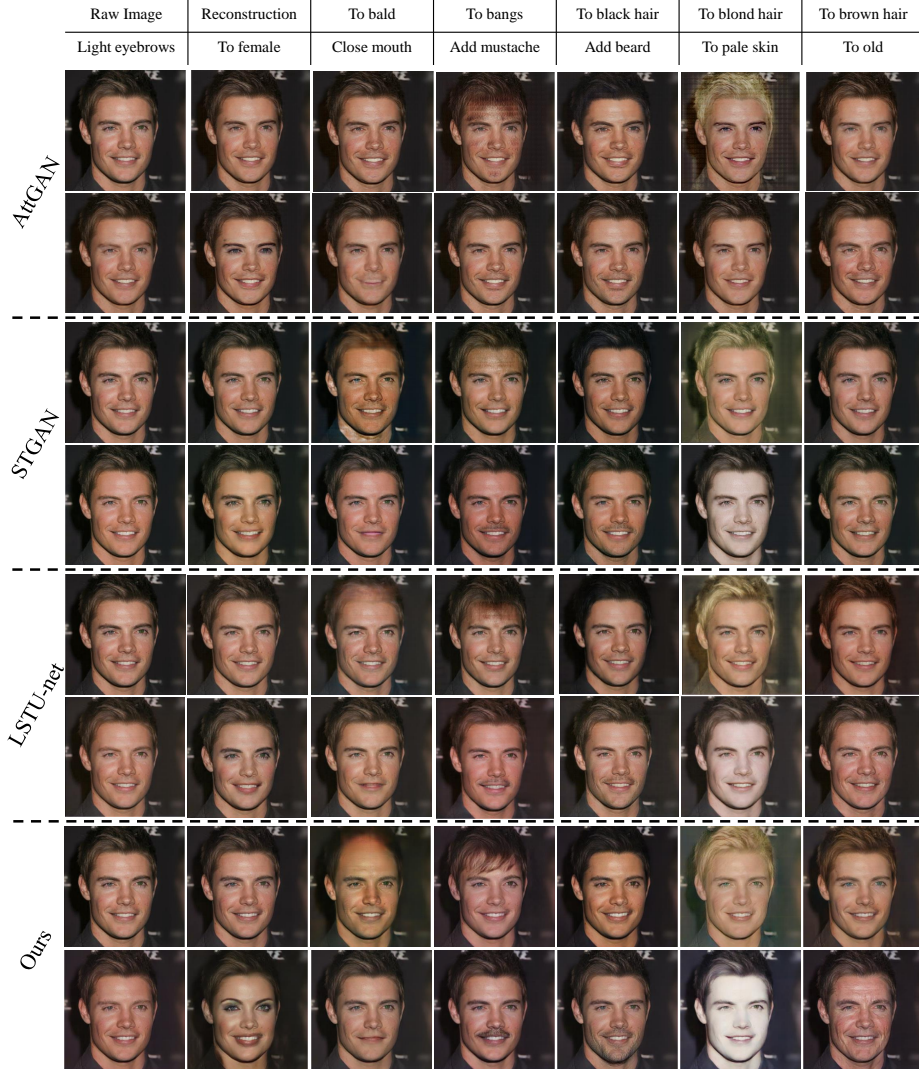| Raw Image	Reconstruction	To bald	To bangs	To black hair	To blond hair	To brown hair
Light eyebrows	To female	Close mouth	Add mustache	Add beard	To pale skin	To old

Fig. 5. High resolution (768 × 768) facial attribute editing results of AttGAN-HR, STGAN-HR, LSTU-net-HR and our framework. The textural details can be observed clearly through zooming in. More HR editing results can be found in suppl.

with the modified HR versions of previous state-of-the-art models including STGAN-HR and AttGAN-HR. In the survey, 50 questions of 14 attributes/quality are involved, *i.e.*, 13 previous mentioned attributes and the smoothness of the image. And in each question, a randomly selected image and its edited versions by the three models are presented in random sequences. And people will
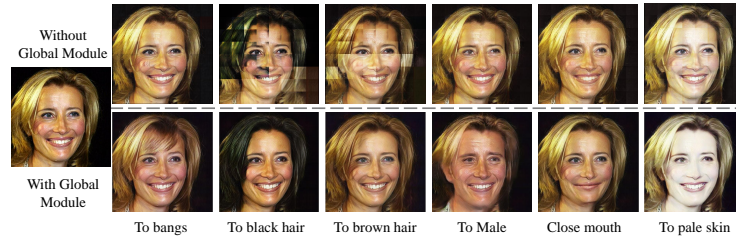
Fig. 6. 768×768 results generated by our framework without and with global module.
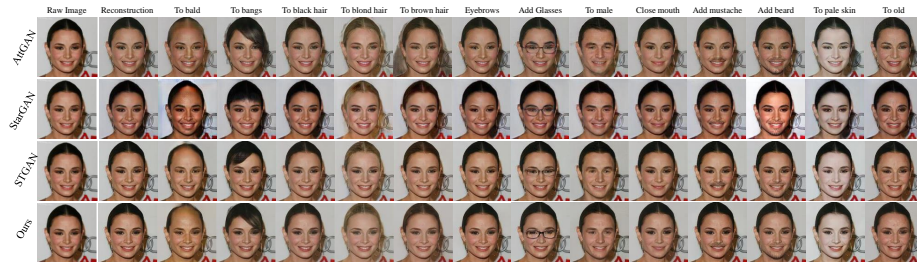

Fig. 7. Low resolution (128×128) facial attribute editing results of StarGAN, AttGAN, STGAN and our LSTU-Net.

determine whether an edited image has correct and realistic attributes and is generally smooth and consistent. The user study results are given in Table 4. In comparison, our CooGAN shows its excellent performance in attributes editing. Because our framework generates image by patches, it is slightly inferior to the other two methods in terms of smoothness.

Ablation study for Cooperative mechanism. In this part, our focus is to study the importance and effectiveness of global module in the framework. We train our framework with and without the global module. The results show that without global module the attribute editing will be impractical, as is shown in Fig. 6. On the one hand, without the spatial information given by global

Method	Bald	Bangs	Black	Blond	Brown	Eyebrows	Glasses
AttGAN-HR	25.7%	66.2%	61.2%	89.4%	47.9%	51.7%	86.7%
STGAN-HR	38.8%	50.5%	78.2%	92.8%	53.5%	65.4%	71.5%
Ours	**61.5%**	**93.3%**	**90.6%**	**94.9%**	**73.6%**	**84.9%**	**94.8%**

Method	Male	Mouth	Mustache	No Beard	Pale	Young	Smooth
AttGAN-HR	36.8%	83.2%	29.8%	51.4%	34.5%	41.4%	99.8%
STGAN-HR	51.3%	67.7%	49.1%	67.8%	61.4%	49.9%	99.9%
Ours	**87.9%**	**91.7%**	**61.9%**	**93.8%**	**98.1%**	**64.7%**	91.5%

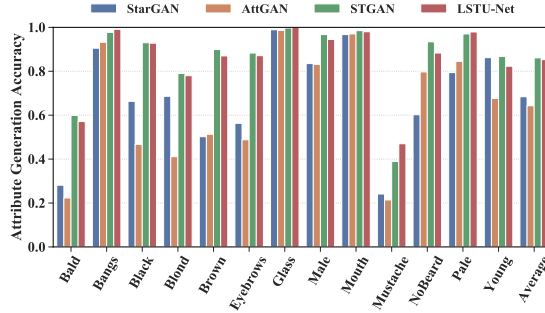**Table 4.** User study results of our framework, STGAN-HR and AttGAN-HR.

Fig. 8. Our LSTU-Net against other approaches in attribute translation accuracy.

snapshot, the local module can not well recognize local attributes like bangs regions and mouth. On the other hand, the shrinkage of perceptive field causes loss of global semantics, rendering certain features (like gender characteristics), difficult for the framework to discern. Thus, we confirm the irreplaceable role of global module in our framework.

4.3   Ablation study for LSTU

LSTU is a vital component of our framework. We will verify the feasibility and effectiveness of LSTU in this section.

Experiment setup. Our LSTU is a lightweight and equally effective replacement unit for STU [20] (previous state-of-the-art multi-scale features fusion unit proposed in STGAN). To validate its effect and efficiency, we replicate the network structure of STGAN and replace its STU with our LSTU. We name this new network LSTU-Net to be distinguished from the original STGAN. Besides, we also compare LSTU-Net with StarGAN and AttGAN since they are the most relative and effective facial attribute editing methods that can be found. However, these compared models are not designed for HR image generation tasks. They are mostly purposed to process 128×128 images. To be fair, all models are tested on 128×128 CelebA dataset.

In addition, the official codes and pretrained models of those three compared models are used in our experiments for the sake of impartiality.

Quantitative evaluation. The attribute editing accuracy and image fidelity comparisons are shown in Fig. 8 and Table 2. For attribute accuracy, our LSTU-Net outperforms StarGAN, AttGAN and ties with STGAN. This circumstance remains the same for image fidelity manifested through PSNR and SSIM. These results show that LSTU has the same level of effect as STU, whether in attribute editing accuracy or image fidelity. The complexity and computational cost comparisons of different units are demonstrated in Table 5. Compared to STU, our LSTU reduces parameter size by half and FLOPS by over 60 %. These comparisons prove the effectiveness and efficiency of implementing LSTU.

channel	16 in-channel		32 in-channel		64 in-channel	
metric	param	FLOPS	param	FLOPS	param	FLOPS
STU	1.63M	5.06G	6.40M	19.99G	25.37M	79.50G
LSTU	0.81M	2.02G	3.20M	7.85G	12.68M	30.80G

Table 5. The total number of parameters and FLOPS in the skip connections within a five-layer network with different input channels for the LSTU and STU. G denotes GFLOPS.

Qualitative analysis. The visual effect comparison of 128×128 images is given in Fig. 7. The editing result of our LSTU is better than StarGAN, AttnGAN, and achieves similar visual effect with STGAN. In almost all attributes, our model not only performs thorough and accurate translation, but also suffer the least from textural artifacts. In contrast, the results of StarGAN have distinct color difference due to the lack of skip connection. AttGAN's results contain fake artificial textures and half-changed features in attributes like bald and hair color. Such inconsistency is caused by the direct use of skip connection. STGAN is more successful than AttGAN and StarGAN. And its edited features and ours are equally effective. These comparison results present the state-of-the-art level attribute editing ability and tremendous efficiency advantage of our LSTU. More results generated by the LSTU-Net are given in the suppl.

5   Conclusion

In this paper, we study the problem of high-resolution facial attribute editing in resource-constrained conditions. By proposing the patch-based local-global framework *CooGAN* along with the multi-scale features fusion method LSTU, we attained the high-resolution facial attribute editing with constrained GPU memory. We use an up-to-down approach for patch processing to preserve global consistency and most importantly, lower the computational resource requirement. And the LSTU retains the attribute variety in skip connection with fewer parameters and lower computation cost. Experiments on facial attribute editing exhibit the superior performance of our framework in quality, attribute accuracy and efficiency to state-of-arts in the scope of facial attribute editing with constrained computational resources. Theoretically, our framework has no image size limitation owing to its patch processing method. We believe it has promising prospect not only in resource-constrained situations, but also in extremely high-resolution image processing tasks.

Acknowledge

References

1. Zao. `https://baike.baidu.com/item/ZAO/23721314`
2. 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society (2018)
3. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE (2018)
4. Awiszus, M., Ackermann, H., Rosenhahn, B.: Learning disentangled representations via independent subspaces. CoRR **abs/1908.08989** (2019)
5. Cao, J., Huang, H., Li, Y., Liu, J., He, R., Sun, Z.: Biphasic learning of gans for high-resolution image-to-image translation. CoRR **abs/1904.06624** (2019)
6. Chen, Y., Lin, H., Shu, M., Li, R., Tao, X., Shen, X., Ye, Y., Jia, J.: Facelet-bank for fast portrait manipulation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 [2], pp. 3541–3549. https://doi.org/10.1109/CVPR.2018.00373
7. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIG-DAT, a Special Interest Group of the ACL. pp. 1724–1734. ACL (2014), `https://www.aclweb.org/anthology/volumes/D14-1/`
8. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 [2], pp. 8789–8797. https://doi.org/10.1109/CVPR.2018.00916
9. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014)
10. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. CoRR **abs/1704.00028** (2017)
11. Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.): Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA (2017)
12. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Attgan: Facial attribute editing by only changing what you want. IEEE Trans. Image Processing **28**(11), 5464–5478 (2019). https://doi.org/10.1109/TIP.2019.2916751
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), `https://openreview.net/group?id=ICLR.cc/2018/Conference`
15. Kemelmacher-Shlizerman, I., Suwajanakorn, S., Seitz, S.M.: Illumination-aware age progression. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 3334–3341. IEEE Computer Society (2014). https://doi.org/10.1109/CVPR.2014.426, `https://ieeexplore.ieee.org/xpl/conhome/6909096/proceeding`

16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), `https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:accepted-main.html`

17. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. In: Guyon et al. [11], pp. 5967–5976

18. Lee, C., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. CoRR **abs/1907.11922** (2019)

19. Lei, T., Zhang, Y., Wang, S.I., Dai, H., Artzi, Y.: Simple recurrent units for highly parallelizable recurrence. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 4470–4481. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1477

20. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: STGAN: A unified selective transfer network for arbitrary image attribute editing. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 [3], pp. 3673–3682

21. Liu, M., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Guyon et al. [11], pp. 700–708

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) (December 2015)

23. Ma, N., Zhang, X., Zheng, H., Sun, J.: Shufflenet V2: practical guidelines for efficient CNN architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. Lecture Notes in Computer Science, vol. 11218, pp. 122–138. Springer (2018). https://doi.org/10.1007/978-3-030-01264-9_8, `https://doi.org/10.1007/978-3-030-01264-9`

24. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 2642–2651. PMLR (2017), `http://proceedings.mlr.press/v70/`

25. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. CoRR **abs/1611.06355** (2016)

26. Qian, S., Lin, K., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. CoRR **abs/1908.07191** (2019)

27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28, `https://doi.org/10.1007/978-3-319-24574-4`

28. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR **abs/1801.04381** (2018)

29. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183:1–183:14 (2015). https://doi.org/10.1145/2816795.2818056

30. Xu, S., Huang, H., Hu, S., Liu, W.: Faceshapegene: A disentangled shape representation for flexible face image editing. CoRR **abs/1905.01920** (2019)

31. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 [2], pp. 1316–1324. https://doi.org/10.1109/CVPR.2018.00143

32. Yi, R., Liu, Y., Lai, Y., Rosin, P.L.: Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 10743–10752. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01100

33. Yin, W., Liu, Z., Loy, C.C.: Instance-level facial attributes transfer with geometry-aware flow. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 9111–9118. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33019111, `https://www.aaai.org/Library/AAAI/aaai19contents.php`

34. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 [2], pp. 6848–6856

35. Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., Tian, Q.: Variational convolutional neural network pruning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 [3], pp. 2780–2789

36. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2242–2251. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.244, `https://ieeexplore.ieee.org/xpl/conhome/8234942/proceeding`