Neurocomputing 443 (2021) 75–84

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief papers

URCA-GAN: UpSample Residual Channel-wise Attention Generative Adversarial Network for image-to-image translation ☆

Xuan Nie [a], Haoxuan Ding [b,*], Manhua Qi [c], Yifei Wang [a], Edward K. Wong [d]

[a] School of Software, Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China
[c] RAN Development Department I, ZTE Corporation, Xi'an 710401, China
[d] NYU Tandon School of Engineering, Brooklyn, NY 11201, USA

ARTICLE INFO

Article history:
Received 3 October 2020
Revised 28 December 2020
Accepted 27 February 2021
Available online 10 March 2021
Communicated by Zidong Wang

2010 MSC:
00-01
99-00

Keywords:
Multimodal image-to-image translation
Generative adversarial network (GAN)
Softmax channel-wise attention mechanism

ABSTRACT

Multimodal image-to-image translation is a challenging topic in computer vision. In image-to-image translation, an image is translated from a source domain to different target domains. For many translation tasks, the difference between the source image and the target image is only in the foreground. In this paper, we propose a novel deep-learning-based method for image-to-image translation. Our method, named URCA-GAN, is based on a generative adversarial network and it can generate images of higher quality and diversity than existing methods. We introduce Upsample Residual Channel-wise Attention Blocks (URCABs), based on ResNet and softmax channel-wise attention, to extract features associated with the foreground. The URCABs form a parallel architecture module named Upsample Residual Channel-wise Attention Module (URCAM) to merge features from the URCABs. URCAM is embedded after the decoder in the generator to regulate image generation. Experimental results and quantitative evaluations showed that our model has better performance than current state-of-the-art methods in both quality and diversity. Especially, the LPIPS, PSNR, and SSIM of URCA-GAN on CelebA dataset increase by 1.31%, 1.66%, and 4.74% respectively, the PSNR and SSIM on RaFD dataset increase by 1.35% and 6.71% respectively. In addition, visualization of the features from URCABs demonstrates that our model puts emphasis on the foreground features.

1. Introduction

Image-to-image translation is the task to generate images in a target domain based on images from a source domain by using a mapping. Applications include image colorization [1], super-resolution image generation [2,3], style transfer [4] and others. Researchers have proposed many deep learning methods for different image-to-image translation tasks; for example, changing people's emotion, changing from summer scenery to winter scenery and keepingn the same objects in the scene. The results for image-to-image generation have greatly improved in recent years due to the introduction of Generative Adversarial Networks (GANs) [5] for this task. GAN-based method usually contains an encoder to map the source image into a common latent feature space through convolutions and a decoder to map those latent feature to target domain image through transport convolutions.

We note that for many image-to-image translation tasks, only parts of the image need to be transformed but not the entire image. For example, if we would like to change the emotion of a person from happy to angry, the changes should only be in the facial region. The hair color, hair style and clothing should stay the same. Researchers have proposed many image-to-image translation algorithms to achieve this. One popular method is the CycleGAN [6]. With enforcing the cycle consistency between the source domain and the target domain, the common content in the source image and the target image can be retained during the cross-domain translation. Other methods, such as DualGAN [7] and DiscoGAN [8], also utilize similar principle in their design to maintain the context in the image. To tackle translation among multiple domains, methods such as MUNIT [9] and StarGAN [10] have been proposed. Our method is based on the StarGAN network, which utilizes one generator and one discriminator to achieve multi-domain translation.

☆ This work was supported by the 2020 key research and development plan of Shaanxi Province under Project 2020ZDLSF04-02.

[*] Corresponding author.
E-mail address: haoxuan_ding@mail.nwpu.edu.cn (H. Ding).

https://doi.org/10.1016/j.neucom.2021.02.054
0925-2312/© 2021 Elsevier B.V. All rights reserved.

In the translation process, a good method should be able to focus on the image region where there is difference between the source image and the target image. This is analogous with the visual attention mechanism in the human visual system. Inspired by the human visual system and the successful applications of the attention mechanism in many computer vision algorithms, we propose an UpSample Residual Channel-wise Attention Generative Adversarial Network (URCA-GAN) for image-to-image translation. In the URCA-GAN system, we embed a novel module (UpSample Residual Channel-wise Attention Module, URCAM) after the decoder in the generator for feature filtering through a number of parallel UpSample Residual Channel-wise Attention Blocks (URCABs). URCAB is a neural network block based on the ResNet [11], it utilizes feature residual to control the feature contents in ResNet and extract features of interest for translation by softmax channel-wise attention. The embedded URCAM is jointly trained with the StraGAN network to emphasize features that are most important and discriminative in the channel feature maps.

Our contributions in this paper are as follows:

- We propose URCAB, a novel residual block that utilizes the residual of feature and channel-wise attention to accomplish the improvement of feature filtering implementations, and combine different URCABs into URCAM with a parallel structure.
- We embed the URCAM module into the generator of the StarGAN network and propose the URCA-GAN for image-to-image translation.
- We have shown the different features that can be extracted by the URCABs in URCAM and analyzed the effect through the visualization of feature maps.
- Our experimental results have demonstrated that our method could improve the quality and diversity of synthetic images.

2. Related works

2.1. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [5] provide an effective way to generate images. A typical GAN model consists of two main modules: generator and discriminator. The discriminator is used to distinguish whether the input image is real or synthetic, and the generator would try their best to fool the discriminator and let the discriminator consider the synthetic images from generator as real images. With adversarial training between the generator and the discriminator, the ability of the generator and the discriminator improve over time so that the generator can generate synthetic images as close as possible to the real images. There are many variants of GANs: Deep Convolutional Generative Adversarial Networks (DCGAN) [12] replaces the multilayer perceptron (MLP) with the convolutional neural network (CNN) to enhance the ability of image generation. Conditional Generative Adversarial Networks (cGAN) [13] could control image generation through embedding an extra condition. InfoGAN [14] introduces the Mutual Information Maximization in CGAN to ensure the embedding condition could regulate image generation as much as possible. Wasserstein GAN (WGAN) [15,16] improves the training stability and could avoid mode collapse through the use of a novel loss function.

2.2. Image-to-image translation

Isola et al. proposed the pix2pix algorithm [4], an image-to-image translation unified model based on cGAN [13]. Wang et al. [17] proposed a method to generate high-resolution images by using pix2pix. The pix2pix algorithm combines adversarial loss and L1 loss between the source images and the target images so that the input of this model must be paired datasets. However, the collection of paired datasets is extremely difficult. In order to overcome the insufficiency of paired datasets, many algorithms that use unpaired datasets have been proposed, in order to perform image-to-image translation using unsupervised learning. Zhu et al. proposed the CycleGAN [6], the first model for unsupervised image-to-image translation. CycleGAN first translates the source domain images to target domain images, and then translates the synthetic target domain images to the source domain, which means that after the translation from source domain to target domain and from target domain to source domain the synthetic images could retain the structure and content of the real source domain images. DualGAN [7] and DiscoGAN [8] also utilize the same principle and enforce the robustness of system by modifying the loss functions. Perarnau et al. [18] presented an Inverse cGAN (IcGAN) to modify the conditional domain representation and achieve the domain translation. Li et al. [19] proposed Deep Identity-aware Transfer (DIAT) which first extracts the changed parts by a mask network and translates image parts by an attribute transform network. Li and Tuzel proposed Couple GAN (CoGAN) [20] which learns the joint distribution of two domains in the latent space to achieve unpaired image translation. Liu et al. [21] also utilizes the hypothesis that the source domain and target domain features exist in the same latent feature space and proposes UNIT which combines Variational Auto-Encoder (VAE) [22] and CoGAN [20]. After encoding the source domain image in the latent feature space, the latent features could be decoded into the target domain space. Huang et al.'s MUNIT [9] achieved multi-modal unsupervised image-to-image translation. They consider one of the features to represent image style and another to represent the image content. StarGAN [10] is also a multi-domain image-to-image translation model. It uses target label embedding and auxiliary discriminator for classification, achieving multi-domain translation by using a single GAN model.

2.3. Attention mechanism

Recently, attention mechanism has been successfully introduced into many applications in computer vision and natural language processing, such as image captioning [23–28], text-to-image generation [29–33], visual question answer [34,35], etc. The attention mechanism helps the neural network to focus on the related parts of the input samples in a task and resolves the problem without supervision. Zhang et al. [36] proposed the Residual Channel-wise Attention Network (RCAN) to produce super-resolution images by considering the correlation among different feature maps and adjust them. Fu et al. [37] utilized channel-wise and spatial-wise attention for object segmentation and proposed the DANet that uses a self-attention mechanism to capture the correlation of spatial content and channel feature respectively. Zhang et al. [38] proposes Self-Attention GAN (SAGAN) to combine GAN and the self-attention mechanism. Self-attention could pay attention on the inner relation in the images, which improves the generation by considering both local area captured by the convolution kernel and global area correlation captured by self-attention. Chen et al. [39], Mejjati et al. [40], Yang et al. [41] and Tang et al. [42] added spatial-wise attention in GAN. They all utilize spatial-wise attention map to extract the image foreground during domain translation, and combine the foreground with background from raw real images, which could improve image quality because some parts of the synthetic images came directly from real images. Woo et al. [43] proposed the Convolutional Block Attention Module (CBAM), a module that combines channel-wise attention and spatial-wise attention. CBAM could adaptively refine the feature along two separate dimensions (channel and spatial). Ma et al. [44] utilized CBAM in GANs. Tang et al. [45] proposed the Multi-Channel Attention Selection GAN (Selection-GAN). The selection-GAN is a two-stage GAN. The first stage produces initial coarse results and the second stage refines the results by using multi-channel attention.

3. Our approach

In this section, we present the details of our proposed URCA-GAN. The goal of our method is to produce a mapping from the source domain to the target domain. In the mapping, we only need to modify the content of the foreground at locations where there are differences between the source image and the target image. In practice, the residuals in ResNet represent the changes in features during the training of the neural networks. Inspired by this, we decide to control the residual features in upsample to distinguish the foreground from the background. Fig. 1 shows the overall structure of our network. The generator in StarGAN consists of encoder *En* and decoder *De*. *En* extracts the latent feature of the input images through convolution. In doing so, the spatial dimension will decrease and the channel dimension will increase, separating the mixed features in RGB images. *De* expands the spatial dimension and fuse the features from different feature maps by using transport convolution. With the combination and fusion of features, separate features are merged into lower channel dimension. We embed URCAM at the end of *De*, which means that URCAM regulates features without considering the transformation of spatial dimension.

3.1. Feature pre-processing

The input of URCAM is the output of *De*, which is a series of high-resolution features $F_{De} \in \mathbb{R}^{64\times128\times128}$ containing the main content of the synthetic image:

$$F_{De} = De(En(x|y)) \tag{1}$$

where $x \in P_{data}$ denotes the real source domain images and y denotes the class label of the target domain. Inspired by the method of Selection-GAN [45], pre-processing is performed to the feature maps F_{De} as shown in Fig. 2. In URCAM, we utilize three global average pooling with different kernel sizes for F_{De} and produce three outputs: $F_1 \in \mathbb{R}^{64\times3\times3}, F_2 \in \mathbb{R}^{64\times5\times5}$ and $F_3 \in \mathbb{R}^{64\times7\times7}$, respectively. The reason why the feature maps are pooled to odd spatial dimension is that in the image-to-image translation task, the main contents are always appeared on the center of images. The odd spatial dimension feature maps could make the central pixels represent the areas where need to be translated. And during the experiments, this pro-pressing method improve the evaluative criteria indeed.

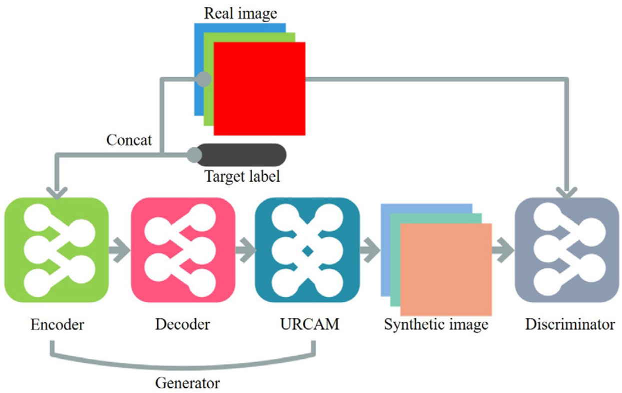

Fig. 1. The structure of our proposed URCA-GAN.

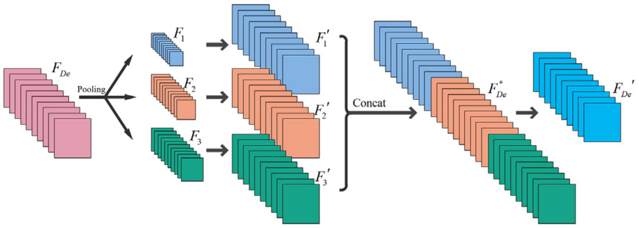

Fig. 2. Feature pre-processing in URCAM.

$$\begin{aligned} F_1 &= GAP_{3\times3}(F_{De}) \\ F_2 &= GAP_{5\times5}(F_{De}) \\ F_3 &= GAP_{7\times7}(F_{De}) \end{aligned} \tag{2}$$

F_1, F_2 and F_3 represent the significance of features with different spatial spans in the feature maps. In order to concatenate these features with different spans, we set an traditional dilated up-sample to restore them to original spatial dimension, receiving $F_1' \in \mathbb{R}^{64\times128\times128}, F_2' \in \mathbb{R}^{64\times128\times128}$ and $F_3' \in \mathbb{R}^{64\times128\times128}$.

$$\begin{aligned} F_1' &= Upsample_{128\times128}(F_1) \\ F_2' &= Upsample_{128\times128}(F_2) \\ F_3' &= Upsample_{128\times128}(F_3) \end{aligned} \tag{3}$$

We concatenate F_1', F_2' and F_3', getting a final feature map F_{De}^*, including the content with different spatial spans in the images:

$$F_{De}^* = concat(F_1', F_2', F_3') \in \mathbb{R}^{192\times128\times128} \tag{4}$$

Then a convolutional layer is added to merge the features and reduce the channel dimension, producing $F_{De}' \in \mathbb{R}^{32\times128\times128}$.

3.2. Upsample Residual Channel-wise Attention Block (URCAB)

Fig. 3 shows the architecture of URCAB. The pre-processed features are fed into URCAB based on the ResNet with bottleneck. In URCAB, we first store the initial features as shortcut, and then feed them as two separate streams. The initial features are first fed into a layer sequence $Layer_{img}$ which consists of a convolutional layer, an instance normalization layer, an ReLU layer, a second convolutional layer and a second instance normalization layer to get the image features. The reason we chose ReLU as the activation function is that ReLU is more effective for image-to-image translation than the Sigmoid or Tanh functions, as demonstrated in our experiments. This sequence extracts the residual features of images F_{img}:

$$F_{img} = Layer_{img}(F_{De}') \tag{5}$$

Meanwhile, the initial features tackled by Sigmoid are fed into another sequence $Layer_{attn}$ with the same structure to predict the attention maps F_{attn}:

$$F_{attn} = Layer_{attn}(Sigmoid(F_{De}')) \tag{6}$$

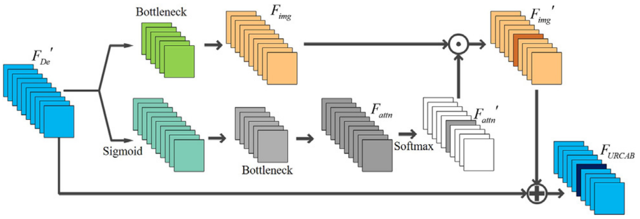

Fig. 3. The architecture of URCAB.

Then we compute softmax channel attention by using softmax along the channel dimension:

$$F'_{attn} = Softmax(F_{attn}) \tag{7}$$

Due to the softmax channel attention, only one attention map in F'_{attn} has a high value, which means this channel attention is reinforced. And we embed channel attention into image features F_{img} through element-wise multiplication:

$$F'_{img} = F_{img} \otimes F'_{attn} \tag{8}$$

Eventually, according to the principle of ResNet, we add the residual features to the shortcut F'_{De}, producing the output feature:

$$F'_{URCAB} = F'_{img} + F'_{De} \tag{9}$$

Due to channel screening by softmax channel-wise attention, URCAB regulates the most distinctive feature in the initial set of features. However, a single URCAB could only control a single feature in the generation process, which means a single URCAB cannot perform well in image-to-image translation. Under this circumstance, an URCAM must contain many URCABs to effectively regulate features.

3.3. Upsample Residual Channel-wise Attention Generative Adversarial Network (URCA-GAN)

Fig. 4 shows the network architecture of URCAM. After pre-processing the features, there are N blocks $\{URCAB_1, URCAB_2, \ldots, URCAB_N\}$ in URCAM. There are two schemes of association of URCAB: parallel architecture and serial architecture. Every URCAB in the parallel architecture of URCAM is separate with each other, which means each URCAB could process the features F'_{De} independently without the influence from other URCABs. In the serial architecture, a URCAB receives features from the previous URCAB and provides the processed features to the next URCAB. We found that the parallel architecture is more effective than the serial architecture and we chose the former in our network through previous experiments.

In the parallel architecture, every URCAB generates a sequence of features with a prominent feature:

$$F_i^{UCRAB} = UCRAB_i(F'_{De}) \quad i = 1 \ldots N \tag{10}$$

With the visualization of feature maps from URCABs, we find that a single URCAB could only enhance one feature in F'_{De} through softmax channel-wise attention, this means that the channel-wise attention could focus on some special features in translation successfully. In order to merge those different enhanced features from URCABs, the output of the parallel architecture URCAM is the average of the features from different URCABs:

$$F_{URCAM} = \frac{1}{N}\sum_{i=1}^{N} F_i^{UCRAB} \tag{11}$$

There is no change of the final channel dimension of features on the output of URCAB. In order to form the three channel RGB image, we add a convolutional layer with the tanh activation function at the end of the network.

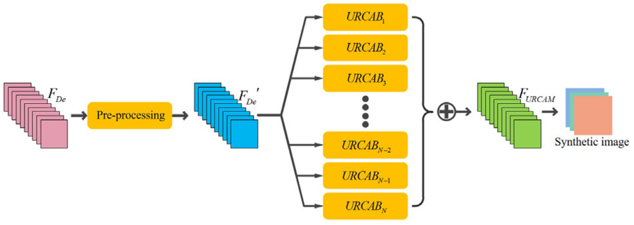

Fig. 4. The network architecture of URCAM.

Since the softmax channel-wise attention do not need any additional restriction, the loss function is the same as the original StarGAN. It includes adversarial loss with gradient penalty L_{adv}, the domain classification loss for real image L^r_{cls}, the domain classification loss for fake image L^f_{cls} and reconstruction loss L_{rec}:

$$\begin{aligned} L_{adv} = {} & \mathbb{E}_x[D_{src}(x)] - \mathbb{E}_{x,y}[D_{src}(G(x,y))] \\ & - \lambda_{gp}\mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1)^2] \end{aligned} \tag{12}$$

where G denotes the generator in StarGAN, D_{src} denotes the discriminator for distinguishing between the real and synthetic images, x denotes the input real image from dataset, y is the target domain label and $\hat{x}$ represents linear interpolation between the real images and synthetic images. λ_{gp} is the weight of the gradient penalty with a value of 10 according to the original StarGAN paper.

$$\begin{aligned} L^r_{cls} &= \mathbb{E}_{x,y'}[-logD_{cls}(y'|x)] \\ L^f_{cls} &= \mathbb{E}_{x,y}[-logD_{cls}(y|G(x,y))] \end{aligned} \tag{13}$$

D_{cls} is the discriminator for domain classification, y' denotes the original label of the input real image x.

$$L_{rec} = \mathbb{E}_{x,y,y'}[\|x - G(G(x,y),y')\|_1] \tag{14}$$

The reconstruction loss is the key of cycle consistency to maintain the content during the unpaired translation.

Finally, the loss functions to train the discriminator and generator are as follows:

$$\begin{aligned} L_D &= -L_{adv} + L^r_{cls} \\ L_G &= L_{adv} + L^f_{cls} + 10L_{rec} \end{aligned} \tag{15}$$

The base network of our model is StarGAN. In order to verify the effectiveness of our module, we did not make any changes to the loss functions and their weights.

4. Experiments

4.1. Datasets

We chose the CelebFaces Attributes (CelebA) [46] and the Radboud Faces Database (RaFD) [47] as datasets. For the CelebA dataset, the training set contains 200,599 face images and the test set contains 2,000 images, randomly selected from the dataset. The initial size of the images are 178 × 218. They were cropped to a size of 178 × 178 and then resized to 128 × 128. We define seven domains in our experiments for multi-domain translation:*black hair, blond hair, brown hair, male, female, young and old*. For the RaFD dataset, the training set contains 1,096 images with eight front view facial expressions and the test set contains 512 images with eight front view facial expressions, 64 images for each expression. The eight facial expressions in the training set and test set are *angry, contemptuous, disgusted, fearful, happy, neutral, sad and surprised*. The images have an initial size of 681 × 1024. They were cropped to a size of 681 × 681 and then resized to 128 × 128.

4.2. Settings

We only embed a module into StarGAN so the settings for training are the same as StarGAN. The network was trained by using Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The input images were flipped randomly in the horizontal direction with probability 0.5. The decay of the learning rate was the same as the strategy used in StarGAN. The model was trained on a single NVIDIA GTX1080 GPU with 8 GB memory. The input batch size is 16. In each training step, we train the discriminator five times and the generator one time.

Empirically, image generation is a procedure to merge the features divided by encoder using the decoder. The approach to merge and compose features has direct effect on the quality of the generated images. A single URCAB in the URCAM can only process a single channel in the feature maps. Under this condition, increasing the number of URCABs allow us to regulate more features in the translation process and better results can be obtained. This has been verified in our experiments. We set the number of URCAB to 8 due to the limitation in GPU memory.

4.3. Experimental results

Fig. 5 illustrates the facial attribute translation results of our model on the CelebA dataset. Fig. 6 illustrates the facial expression translation results of our model on the RaFD dataset. Fig. 7 shows comparison among several popular image-to-image translation models on the CelebA dataset, including CycleGAN [6], StarGAN [10] and ours. The results show that our model can effectively perform multi-domain facial attribute translation and preserve the content in the source domain images. Compared with the results from StarGAN, our model retains more details in the image. For example, for translation to blond hair, the results from StarGAN have a blurred area near the hair tip. In contrary, our model generates clear hair tips that look more realistic. This is due to the ability of our proposed model to select and focus on relevant features during translation.

Fig. 8 shows comparison of results for the RaFD dataset. Experimental results demonstrate that our model could generate images with finer details. Expression changes from our model are more obvious and noticeable when compared with other models. For example, the changes in the corners of the mouth and the external canthus of the eyes in the *sad* expression are very noticeable in our model, and the results from other models show only a slight change. In addition, the mouth in the *surprise* expression from our model has a larger gap than other models. One explanation for this phenomenon is that the channel-wise attention mechanism in our model reinforces the features for these expressions.

4.4. Evaluation

In this section, we evaluate the performance of our model quantitatively. For the CelebA dataset, we utilize the Inception Score (IS) [48,49] and the Mode Score (MS) [50] calculated through a ResNet-18 network which was pre-trained on the AlexNet. These two scores are used to evaluate the quality of the generated images. The Inception Score is the most widely used metric for the evaluation of GANs and the Mode Score is an improvement of the Inception Score, which has the added ability to measure the dissimilarity between the distribution of the real images and the distribution of the generated images. This is achieved by adding KL divergence in the score. Higher IS and MS scores mean the generated images are closer to real images. Meanwhile, we need to evaluate the diversity of models to examine whether mode collapse exists during training. In recent years, the Learned Perceptual Image Patch Similarity (LPIPS) [51] metric is commonly used to evaluate the diversity of generated images. We utilize a pre-trained AlexNet to calculate LPIPS, which is commonly used in the quantitative evaluation of LPIPS. Meanwhile, in order to evaluate the general image quality, we also calculate the PSNR and SSIM of generated images.

Table 1 shows the IS, MS, LPIPS, PSNR and SSIM scores for the evaluated models. We found that the IS score of StarGAN is slightly higher than that of our model. However, the MS score of our model is higher than StarGAN's. These two scores both measure the quality of the generated images but IS only evaluates the quality through samples from generated results, which means IS could not fully utilize the information from real image distribution. MS avoids the limitation of IS and imports the real image distribution in evaluation, gaining more reasonable evaluation results. The MS of our proposed method is higher than that of StarGAN, which means our method could improve the quality of synthetic images effectively. The MS score of URCA-GAN increases by 1.85%. Meanwhile, our method has the highest LPIPS, PSNR, SSIM value, meaning our model has the best diversity and quality on image generation. The LPIPS, PSNR, and SSIM of our method on CelebA dataset increase by $1.31\%, 1.66\%$, and 4.74%, respectively. In addition, due to the limitation of GPU memory, we only embed eight URCABs in URCAM. If we increase the number of URCABs, we could have better image-to-image translation results.

To evaluate the performance on the RaFD dataset, we train a $ResNet-18$ network with 10-fold cross-validation technique to

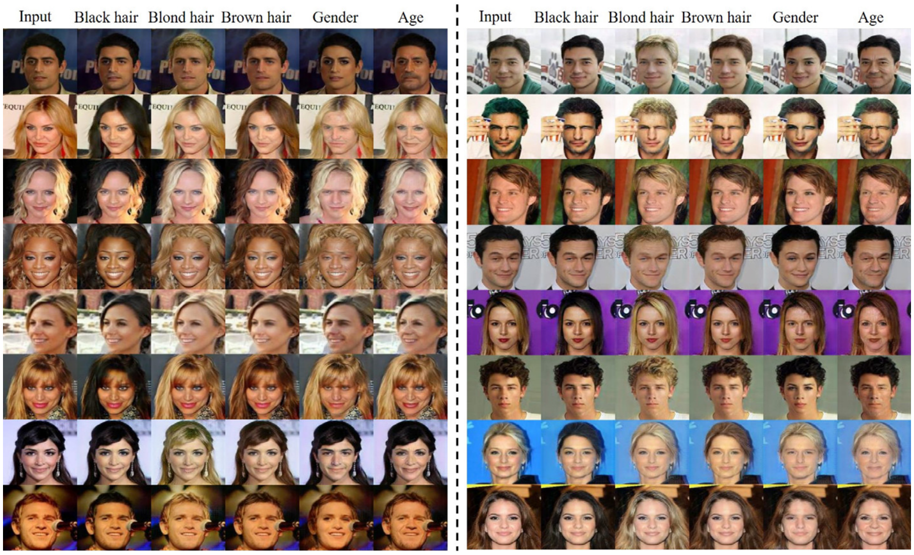

Fig. 5. Experimental results on the CelebA dataset.

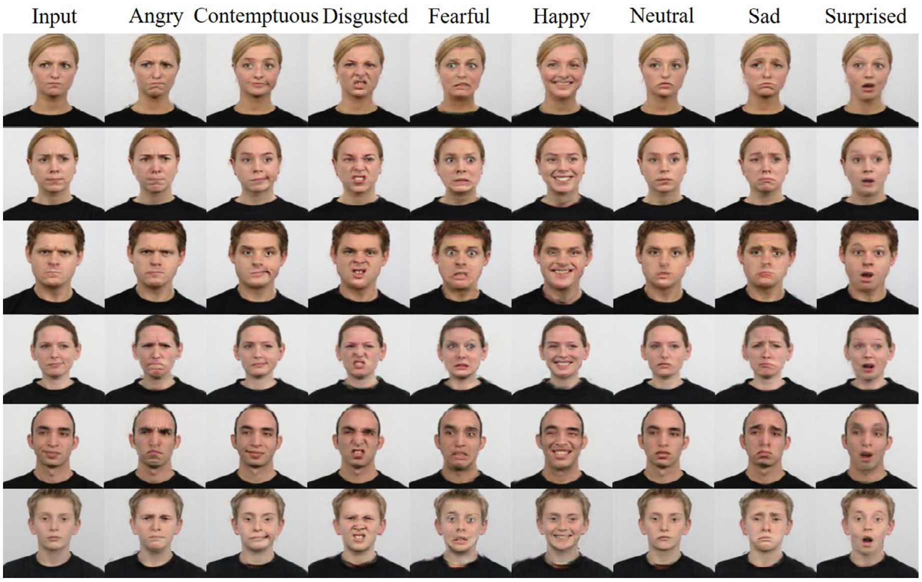

Fig. 6. Experimental results on the RaFD dataset.

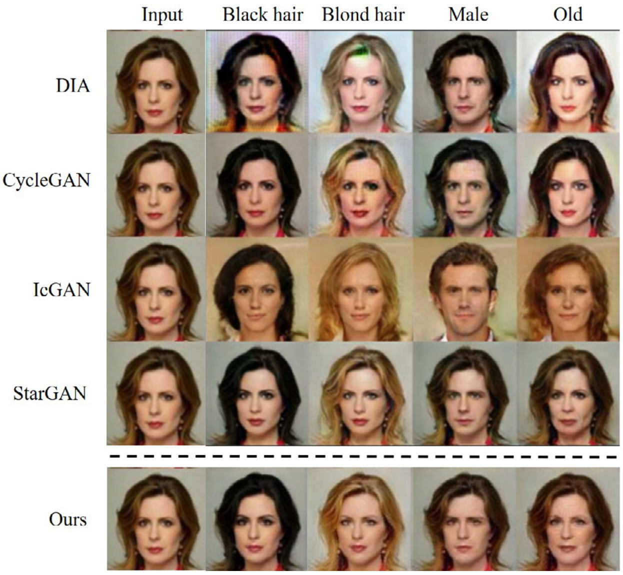

Fig. 7. Comparison of results for the CelebA dataset. Results for CycleGAN and StarGAN came from [10].

classify eight facial expressions (*angry, contemptuous, disgusted, fearful, happy, neutral, sad and surprised*) in the RaFD dataset. We use the network to test the experimental results from different models. If the synthetic images could be classified correctly, we can conclude that the model can generate synthetic images that are close to real images and the model is effective. In addition, we organize a survey in the campus, the participants need to choose the best results among 40 results from five methods in comparison (every model has 8 results with 8 facial expressions). Finally, there are 47 students participate in this investigation, and we count the number of their choice and calculate the selection rate of each model. Table 2 shows the classification accuracy, PSNR, SSIM and survey resluts on the RaFD dataset. We found that our model could increase the classification accuracy of StarGAN by 0.42%. Meanwhile, the general evaluation measurements of image quality, PSNR and SSIM, are also gauged in the experiments. The evaluation results show that our proposed URCA-GAN reaches the top value in the comparison with other models. Especially, the PSNR and SSIM of URCA-GAN on RaFD dataset has increased by 1.35% and 6.71% respectively compared with the baseline (StarGAN).

However, there are also drawbacks of our model. The embedded URCAM have extra parameters and increases the total number of parameters both in training and inference. Table 3 shows the timing analysis among several models. It illustrates that the number of parameters in our proposed method is more than that of StarGAN due to the embedded URCAM. The training time of DIAT and CycleGAN is the longest because these two models are single-modal image-to-image translation method, but the model for single domain translation is small so that there is not a significant increase in the inference time on a single domain. As for our proposed method, the features from decoder are input into the parallel URCABs of URCAM, after all of the URCABs process the features and extract the salient single feature in whole features, the concatenated features are merged and calculated. This procedure in URCAM would cost more time in the generation of images. Meanwhile, the number of parameters of multi-modal methods, StarGAN and IcGAN, are less than our proposed method. Therefore, the training time and inference time would be longer than the StarGAN, which is the drawback of our proposed URCA-GAN.

4.5. Visualization of features extracted by URCABs

In order to analyze URCAM directly, we use visualization techniques to illustrate the features processed by softmax channel-wise attention in URCABs. The softmax channel-wise attention mechanism extracts a single feature map from the set of feature maps in URCAB. There are eight URCABs in our model and every URCAB has one feature map reinforced by softmax channel-wise attention and it represents the regulated feature of URCAB. We visualize the feature maps extracted by the eight URCABs in URCAM. In order to better visualize the features, the feature maps are enhanced by gray scale stretching using the logarithm operation.

Fig. 9 illustrates the visualization of features in the translation on the CelebA dataset. We can see that $URCAB_1$ and $URCAB_6$ mainly extract the background and $URCAB_2$ extracts the shape of the hair. From the comparison between *blond* ↔ *black* and

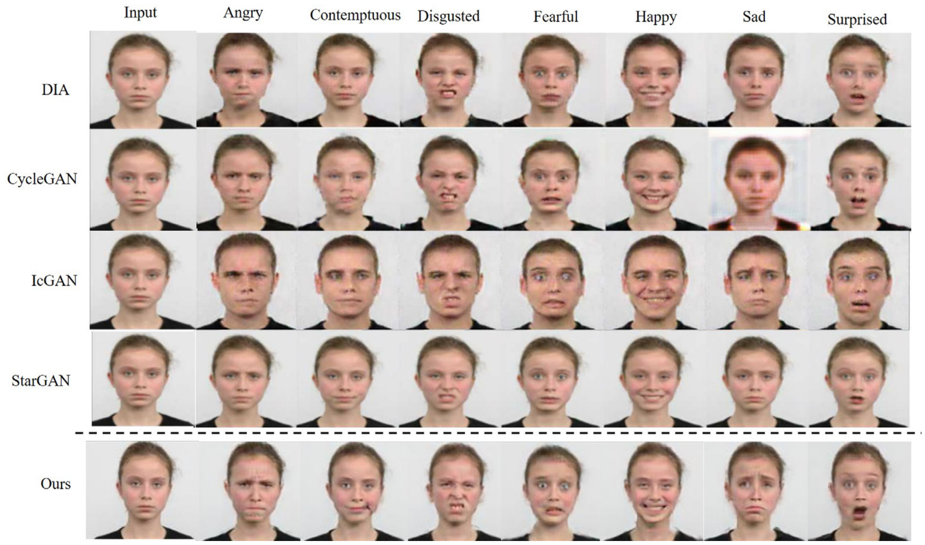

Fig. 8. Comparison of results for the RaFD dataset. The results for CycleGAN and StarGAN came from [10].

Table 1
Quantitative evaluation on the CelebA dataset.

Method	IS	MS	LPIPS	PSNR	SSIM
DIAT	1.137	0.286	0.025	13.41	0.3460
CycleGAN	1.063	0.263	0.012	12.67	0.3575
IcGAN	1.206	0.293	0.028	8.07	0.2965
StarGAN	**2.058**	1.244	0.229	20.54	0.5698
Ours	2.026	**1.267**	**0.232**	**20.88**	**0.5968**

Table 2
Quantitative evaluation on the RaFD dataset.

Method	Classification accuracy/%	PSNR	SSIM	Survey/%
DIAT	92.77	8.66	0.5793	4.3
CycleGAN	91.86	13.37	0.5613	8.5
IcGAN	85.36	10.49	0.4782	0
StarGAN	96.96	14.58	0.6750	31.9
Ours	**97.05**	**18.60**	**0.7203**	**36.2**
real image	99.55	–	–	–
image number	64 × 8	–	–	–

Table 3
The timing analysis of models.

Method	# of parameters	training time	inference time pre domain
DIAT	52.6M × 7	~7 day	0.0436 s
CycleGAN	52.6M × 14	~14 day	0.0436 s
IcGAN	67.8M	~1 day	0.0538 s
StarGAN	53.2M	~1 day	0.0467 s
Ours	69.9M	~2 day	0.0664 s

$blond \leftrightarrow blond(sourcedomain \leftrightarrow sourcedomain)$, we found that the area with a red circle in $URCAB_2$ of $blond \leftrightarrow black$ is smaller than that of $blond \leftrightarrow blond$, and this phenomenon exists in the synthetic images, the black hair is thinner than blond hair in the corresponding marked area. The feature of $URCAB_3$ evidently represents the area where there is a need to change the color, as the $URCAB_3$ is only activated in $blond \leftrightarrow black$ and $blond \leftrightarrow brown$, but not activated in $blond \leftrightarrow blond, female \leftrightarrow male$ and $young \leftrightarrow old$. The ability of $URCAB_4$ is insignificant in our experiments on the CelebA dataset. $URCAB_5$ extracts the key content of the foreground and has control of the content of the whole face. For $female \leftrightarrow male$ and $young \leftrightarrow old$, there are noticeable modifications on the facial profile and individual parts, including the edge of the jaw, nasolabial folds and other parts, to emphasize the special characteristics in the translation of gender and age. The $URCAB_7$ extracts the texture of the hair. From the feature map, we could find many tiny lines that correspond to the texture in the hair. $URCAB_8$ extracts the wrinkles on the face, especially the wrinkles which are marked by the yellow circle in $young \leftrightarrow old$.

Fig. 10 illustrates the visualization of features on the RaFD dataset. The features extracted by URCABs on the RaFD dataset have different meanings and characterizations. $URCAB_1, URCAB_2, URCAB_6$ and $URCAB_8$ learn the generation of background where is unnecessary in the image-to-image translation process. $URCAB_3$ represents the main content for translation, including the face profile, the facial expressions and other features. From the $URCAB_3$, we can find distinct changes on facial expressions, the mouth, the eyes and even the nose for each domain pair. The inactive areas in $URCAB_4$ represent the bright areas on the forehead and jaw of the face due to reflection, so that $URCAB_4$ can capture the shadows on the face. Those inactive areas are marked by red circles in Fig. 10. $URCAB_5$ extracts the edge content on the face.

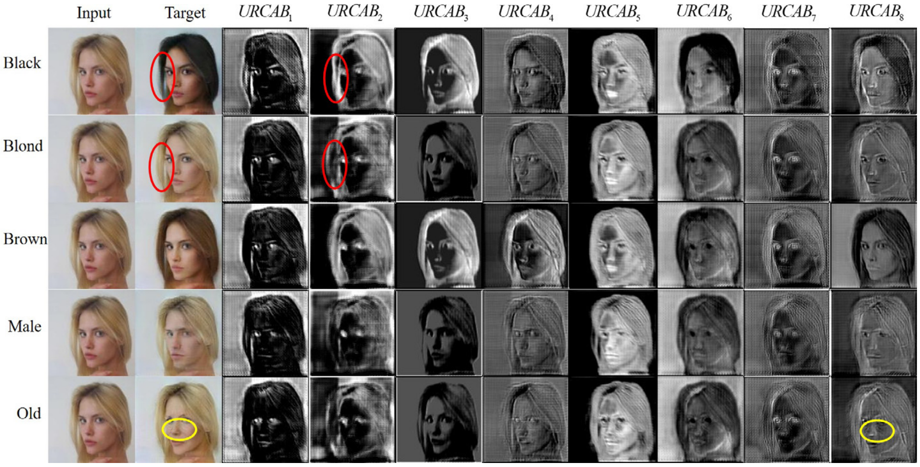

Fig. 9. Features extracted by URCABs on the CelebA dataset. Areas for analysis are marked by red and yellow circles.

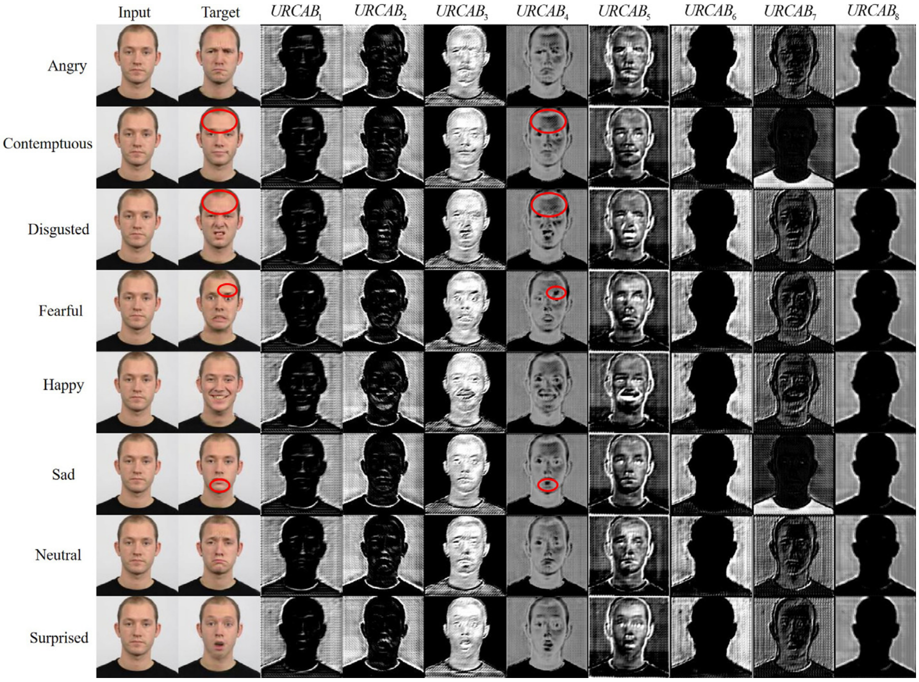

Fig. 10. Features extracted by URCABs on the RaFD dataset. Areas for analysis are marked by red circles.

For example, the lips, nasolabial fold and orbital cavity are always activated in the feature map of $URCAB_5$. $URCAB_7$ did not extract useful features in our experiments.

Every URCAB has regulated different features in the image-to-image translation process. This means URCAB can extract and reinforce single features effectively. Table 4 illustrates the different features extracted by eight URCABs in our experiments on the CelebA and RaFD datasets.

Table 4
Features extracted by URCABs

# of URCAB	CelebA	RaFD
1	background	background
2	shape of hair	background
3	hair color	expressions
4	–	shadows
5	foreground	edges
6	background	background
7	texture of hair	–
8	edges	background

5. Conclusion

In this paper, we have proposed URCA-GAN, a novel deep learning network for the up-sample in image-to-image translation. In URCA-GAN, we embed a module named URCAM which is composed by a number of parallel URCABs, a modified Residual Block with softmax channel-wise attention. A single URCAB is utilized to extract and refine a single feature related to the foreground in the feature maps. The enhanced features from different parallel URCABs are merged to form the output of URCAM. The quantitative evaluation demonstrates that URCA-GAN can improve the quality and diversity of synthetic images. In addition, the visualization of features in URCAM demonstrated that the upsample residual channel-wise attention could extract and regulate the features in image-to-image translation.

CRediT authorship contribution statement

Xuan Nie: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Supervision. **Haoxuan Ding:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft. **Manhua Qi:** Software, Validation, Investigation, Resources. **Yifei Wang:** Validation, Investigation, Resources. **Edward K. Wong:** Writing - review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Cao, Z. Zhou, W. Zhang, Y. Yu, Unsupervised diverse colorization via generative adversarial networks, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I, Vol. 10534 of Lecture Notes in Computer Science, Springer, 2017, pp. 151–166..

[2] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 105–114..

[3] B. Wu, H. Duan, Z. Liu, G. Sun, SRPGAN: perceptual generative adversarial network for single image super resolution, CoRR abs/1712.05927..

[4] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 5967–5976..

[5] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, 2014, pp. 2672–2680..

[6] J. Zhu, T. Park, P. Isola, A.A. Efros, 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, IEEE Computer Society, 2017, pp. 2242–2251..

[7] Z. Yi, H.R. Zhang, P. Tan, M. Gong, Dualgan: Unsupervised dual learning for image-to-image translation, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 2868–2876..

[8] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1857–1865..

[9] X. Huang, M. Liu, S.J. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III, Vol. 11207 of Lecture Notes in Computer Science, Springer, 2018, pp. 179–196..

[10] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, 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, IEEE Computer Society, 2018, pp. 8789–8797..

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 770–778..

[12] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings, 2016..

[13] M. Mirza, S. Osindero, Conditional generative adversarial nets, CoRR abs/1411.1784..

[14] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 2172–2180..

[15] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, CoRR abs/1701.07875..

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems 30 Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5767–5777.

[17] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 8798–8807..

[18] G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Álvarez, Invertible conditional gans for image editing, CoRR abs/1611.06355..

[19] M. Li, W. Zuo, D. Zhang, Deep identity-aware transfer of facial attributes, CoRR abs/1610.05586..

[20] M. Liu, O. Tuzel, Coupled generative adversarial networks, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 469–477..

[21] M. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems 30 Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 700–708.

[22] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings, 2014..

[23] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society, 2015, pp. 3156–3164..

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, Vol. 37 of JMLR Workshop and Conference Proceedings, JMLR.org, 2015, pp. 2048–2057..

[25] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 3242–3250..

[26] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 6298–6306..

[27] J. Lu, J. Yang, D. Batra, D. Parikh, Neural baby talk, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 7219–7228..

[28] Y. Mori, H. Fukui, T. Hirakawa, J. Nishiyama, T. Yamashita, H. Fujiyoshi, Attention neural baby talk: Captioning of risk factors while driving, in: 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, Auckland, New Zealand, October 27–30, 2019, IEEE, 2019, pp. 4317–4322.

[29] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, Vol. 48 of JMLR Workshop and Conference Proceedings, JMLR.org, 2016, pp. 1060–1069..

[30] H. Zhang, T. Xu, H. Li, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 5908–5916..

[31] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, 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, IEEE Computer Society, 2018, pp. 1316–1324..

[32] Y. Cheng, Z. Gan, Y. Li, J. Liu, J. Gao, Sequential attention GAN for interactive image editing via dialogue, CoRR abs/1812.08352..

[33] B. Li, X. Qi, T. Lukasiewicz, P.H.S. Torr, Controllable text-to-image generation, in: Advances in Neural Information Processing Systems 32 Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 2019, pp. 2063–2073.

[34] Z. Yang, X. He, J. Gao, L. Deng, A.J. Smola, Stacked attention networks for image question answering, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 21–29..

[35] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 6077–6086..

[36] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Computer Vision – ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Vol. 11211 of Lecture Notes in Computer Science, Springer, 2018, pp. 294–310..

[37] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE, 2019, pp. 3146–3154..

[38] H. Zhang, I.J. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks, in: Proceedings of the 36th International Conference on

Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 7354–7363..

[39] X. Chen, C. Xu, X. Yang, D. Tao, Attention-gan for object transfiguration in wild images, in: Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part II, Vol. 11206 of Lecture Notes in Computer Science, Springer, 2018, pp. 167–184..

[40] Y.A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, K.I. Kim, Unsupervised attention-guided image-to-image translation, in: Advances in Neural Information Processing Systems 31 Annual Conference on Neural Information Processing Systems 2018, Montréal, Canada, 2018, pp. 3697–3707.

[41] C. Yang, T. Kim, R. Wang, H. Peng, C.J. Kuo, Show, attend, and translate: Unsupervised image translation with self-regularization and attention, IEEE Transactions on Image Processing 28 (10) (2019) 4845–4856.

[42] H. Tang, D. Xu, N. Sebe, Y. Yan, Attention-guided generative adversarial networks for unsupervised image-to-image translation, in: International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14–19, 2019, IEEE, 2019, pp. 1–8..

[43] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Computer Vision – ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Vol. 11211 of Lecture Notes in Computer Science, Springer, 2018, pp. 3–19..

[44] B. Ma, X. Wang, H. Zhang, F. Li, J. Dan, CBAM-GAN: generative adversarial networks based on convolutional block attention module, in: Artificial Intelligence and Security - 5th International Conference, ICAIS 2019, New York, NY, USA, July 26–28, 2019, Proceedings, Part I, Vol. 11632 of Lecture Notes in Computer Science, Springer, 2019, pp. 227–236..

[45] H. Tang, D. Xu, N. Sebe, Y. Wang, J.J. Corso, Y. Yan, Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE, 2019, pp. 2417–2426..

[46] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[47] O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus, A.F.M.V. Knippenberg, Presentation and validation of the radboud face database, Cognition & Emotion 24 (8) (2010) 1377–1388.

[48] T. Salimans, I.J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 2226–2234..

[49] S.T. Barratt, R. Sharma, A note on the inception score, CoRR abs/1801.01973..

[50] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, K.Q. Weinberger, An empirical study on evaluation metrics of generative adversarial networks, CoRR abs/1806.07755..

[51] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 586–595..

Xuan Nie was born in 1976, He is an associate professor with School of Software, Northwestern Polytechnical University, Xi'an City, China. He received the B.S. degree, the M.S. degree and the Ph.D. in Automatic Control, Pattern Recognition and Computer Application Technology from Northwestern Polytechnical University of China, Xi'an City, China, in 1998, 2001, and 2005 respectively. He joined the School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an, China, in 2006, as a lecture and has been a Associate Professor of Since 2010. He was a visiting professor in Hong Kong Poly-technical University in 2010 and in University of Michigan, USA during 2013 respectively. His main research interest covers Machine Learning, Computer Vision, Image Processing, and their applications. He has authored and coauthored over 30 journal and conference papers, three monographs and co-invented patents. Dr. Nie was a reviewer for the IEEE Internet of Things Journal. He is a recipient of Science and Technology Achievement Award of Xi'an City 2015.

Haoxuan Ding was born in Xi'an City, China in 1995. He received the B.S. degree in Flight Vehicle Propulsion Engineering from Northwestern Polytechnical University, China, in 2018. He has been pursuing the M.S. degree with Northwestern Polytechnical University, Xi'an, China, since 2018. His current research interests include Generative Adversarial Network, object detection and their industrial applications.

Manhua Qi was born in Xi'an City, China in 1994. She received the M.S. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, in 2020. She currently works in ZTE Corporation as a 5G Software Engineer.

Yifei Wang was born in 1995. He received the B.S. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, in 2017, where he is currently pursuing the M.S. degree. His current research interests include image super-resolution and object detection.

Edward K. Wong received his B. E. degree from the State University of New York at Stony Brook, his Sc. M. degree from Brown University and his Ph. D. degree from Purdue University, all in Electrical Engineering. He is currently associate professor in the Department of Computer Science and Engineering at the NYU Tandon School of Engineering, Brooklyn, NY. His research interests lie in the areas of computer vision, multimedia computing, medical image processing, and digital forensics, and he has published extensively in these areas. He has worked on many research projects funded by federal and state agencies, as well as private industry. He has served as an associate editor for the journal Information Sciences and the International Journal of Multimedia Intelligence and Security, and is currently an associate editor for the journal Springer LNCS Transactions on Data Hiding and Multimedia Security. Dr. Wong has also served on the organizing committee and technical program committee of numerous IEEE, ACM, and other international conferences.