A GAN based method for multiple prohibited items synthesis of X-ray security image*

LI Da-shuang (李大双)[1], **HU Xiao-bing** (胡小兵)[1]**, **ZHANG Hai-gang** (张海刚)[2], and **YANG Jin-feng** (杨金锋)[2]

1. Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China

2. Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic, Shenzhen 518055, China

(Received 25 February 2020; Revised 22 April 2020)

Detecting prohibited item based on convolutional neural networks (CNNs) is of great significance to ensure public safety. However, the natural occurrence of such prohibited items is a small-probability event, collecting enough datasets to support CNN training is a big challenge. In this paper, we propose a new method for synthesizing X-ray security image with multiple prohibited items from semantic label images basing on Generative Adversarial Networks (GANs). Theoretically, we can use it to synthesize as many X-ray images as needed. A new generator architecture with Res2Net is presented, which is more effective in learning multi-scale features of different prohibited items images. This method is extended by establishing the semantic label library which contains 14 000 images. So we totally synthesize 14 000 X-ray security images. The experimental results show the super performance (Fréchet Inception Distance (FID) score of 30.55). And we achieve 0.825 of mean average precision (mAP) with Single Shot MultiBox Detector (SSD) for object detection, demonstrating the effectiveness of our approach.

Document code: A **Article ID:** 1673-1905(2021)02-0112-6

DOI https://doi.org/10.1007/s11801-021-0032-7

In order to ensure the safety of public transportation, X-ray safety baggage inspection system is widely used at railway stations, subway stations, airports, etc. Especially in the field of civil aviation, airport security is the last line of defense to ensure the safe operation of aviation. However, in practice, the efficiency and reliability of manual detection are undesirable. The prohibited items in the baggage are often squeezed and overlapped each other, so the detection of them takes a lot of time, which greatly reduces the travel efficiency of passengers. Especially in rush hours, security inspectors have tremendous work intensity, which increases the possibility of missing detection. Therefore, it is of great significance to establish a fast and accurate automatic detection system for X-ray security image of prohibited items in security check. Previous approaches are primarily based on the bag of visual words model (BoVW)[1]. More recently, convolutional neural networks have been shown to be more effective in detecting security X-ray images of prohibited items than BoVW. In Ref.[2], two CNN architectures, Faster R-CNN and RetinaNet, are used to detect prohibited items. The performance of this detection method depends on a large number of X-ray security images containing prohibited items. There are two public X-ray image datasets, GDXray[3] and SIXray[4]. However, the availability of GDXray is limited because of the gray-scale images included. SIXray has more than one million realistic, colored X-ray security images, but only 8 929 of them contained prohibited items, which are not enough to meet the training requirements. In addition, collecting images of prohibited items manually is very difficult, because carrying prohibited items on daily trips is a low probability event. The insufficiency of training images leads to a great difficulty for developing reliable CNN architectures suitable for prohibited item detection.

This motivates the use of synthetic image for data augmentation to meet the requirements of model training on database. Threat Image Projection (TIP)[2] method takes into account the consistency of transparency and contrast between the background image and the target image. However, the colors of different materials in the security X-ray image are significantly different, so the two parameters need to be reset for each image, which increases the difficulty and complexity

* This work has been supported by the National Natural Science Foundation of China (No.61806208).

** E-mail: huxb1975@163.com

of image blend. The X-ray Image-Synthesis-Generative Adversarial Networks (XS-GAN)[5] method is used, and the results are relatively real. However, only a prohibited item is considered.

Here, we attempt to address two main issues of the most advanced methods above: (1) the lack of nature and authenticity in the TIP results and (2) the difficulty of synthesizing multiple prohibited items with GANs. In this paper, we show a new approach that synthesis more natural X-ray images with multiple prohibited items from semantic label maps and image blend steps are no longer required. This work may enrich the study of synthesizing X-ray security images with multiple prohibited items.

Our main contributions can be summarized as follows:

a) We build an image synthesis model to achieve the synthesis of X-ray security images with multiple prohibited items.
b) A novel residual block is introduced to improve the multi-scale feature extraction capability for generator.
c) We extend our approach by establishing the semantic label library to synthesize as many X-ray security images as needed.
d) The SIP method is proposed to synthesize semantic label images.

In the evaluation phase, on one hand, Fréchet Inception Distance (FID)[6] score is used to evaluate the authenticity and diversity of the synthetic images. On the other hand, two classification networks, Fully Convolutional Networks (FCNs)[7] and Single Shot MultiBox Detector (SSD)[8], are used to evaluate the feasibility of using datasets generated by our model.

The generative adversarial network (GAN)[9] generator learns the distribution of the real images, and the discriminator is responsible for distinguishing the real images from the generated image. The image-to-image translation is a method based on GANs to synthesize images. The input image is often used as sketch image, semantic label maps[10] or instance image[11], etc, and it is translated into real scene image. The pix2pix[12] method is first proposed for image-to-image-translation, which needs the paired dataset. After that, CycleGAN[13], DualGAN[14] and DiscoGAN[15] transform directly on the image domain and solve the problem of datasets in pairs. InstaGAN[11] introduces attributes of instance to implement translation tasks involving shape changes. In addition, the synthesis of complex high-resolution images has always been a challenging problem. Recently, the proposed pix2pixHD[16] and SPADE[17] synthesize high-resolution images which transform semantic label maps. However, the result of pix2pixHD is unstable for multi-object images in which prohibited items have multiple scales. Some prohibited items are shown in Fig.1. Zhao K et al[18] suggested that ResNet could not extract multi-scale features precisely. They proposed Res2Net to solve this problem, and its applications in instance segmentation and object detection network show the superiority of extracting multi-scale features. Inspired by this, we design a new generator architecture to improve pix2pixHD, in order to generate stably multi-object images.

We can see complete synthesis method from Fig.2. First, the semantic label images are obtained by annotation tool. Next, the database including semantic label images and X-ray images is used to train synthesis model. Then, the semantic label library is established. Finally, we can choose image from semantic label library as input of image synthesis model to generate X-ray security images with prohibited items.

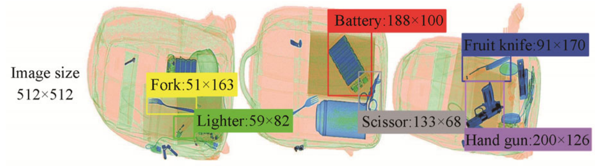


Fig.1 Multi-scale prohibited items

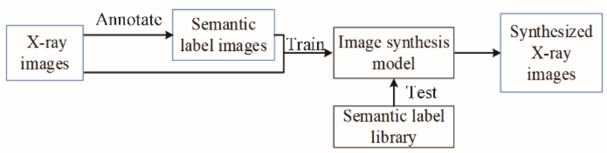
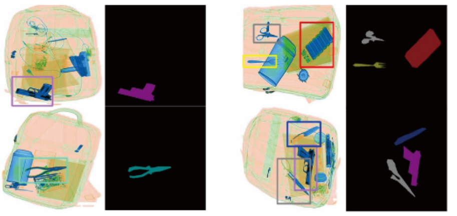


Fig.2 Flowchart of synthesis

We collect X-ray security image using security X-ray machines. They include seven categories, including handgun, pliers, scissors, fruit knives, forks, batteries, lighters. Each image contains different amounts of prohibited items. To be specific, we divide the database into two sets, one set contains single-object image containing only one prohibited item and the other set is about multi-object image containing two or three prohibited items. The number of single-object images is 1 800, the number of multi-object images is 1 688 and their size is 512×512 (for examples, as shown in Fig.3).



(a) Single-object image (b) Multi-object image

Fig.3 Database image

We obtain the semantic label maps shown in Fig.3 using an annotation tool. For representing semantic la bel maps, we introduce a dictionary, which form is {prohibited item: pixel value, color}. Such as {handgun: 1, purple}, {batteries: 2, red}, {fruit knives: 3, blue}. The background pixel value is 0, which is black.

Our image synthesis method is achieved by learning the mapping from semantic label image to real image, so the diversity of semantic label images directly affects the diversity of real image. In general, the prohibited items are multi-posture and varied in categories. Generally we can only secure a limited number of semantic label images which cannot represent all postures. Therefore, we used the approach shown in Fig.4 to build a semantic label library, which contains a large number of semantic label images with more postures and categories.

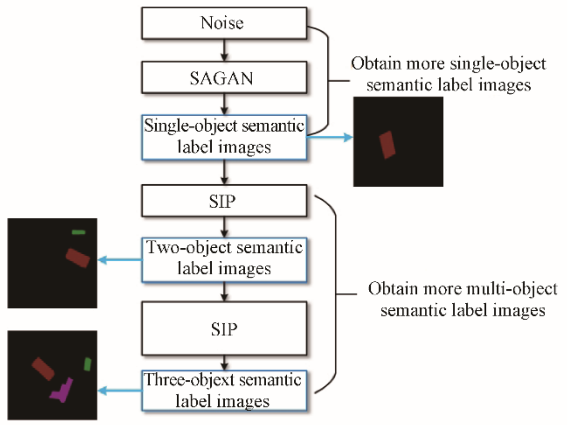

Noise
SAGAN
Single-object semantic label images
Obtain more single-object semantic label images
SIP
Two-object semantic label images
SIP
Three-object semantic label images
Obtain more multi-object semantic label images

Fig.4 Illustration of establishing semantic label library

First, the Self-Attention Generative Adversarial Network (SAGAN)[19] model is used to generate a large number of single-object semantic label images with various postures from random noise. Because the uncontrollability of generating multi-object images from noise makes it difficult to synthesize multi-project semantic image of a specified type using GAN. Based on the same background of semantic label images, we propose an approach called Semantic Image Projection (SIP), which is described by Eqs.(1) and (2), to obtain the multi-object semantic label images.

$$I_{\mathrm{d}}(i,j)=I_{\mathrm{s}_1}(i,j)+I_{\mathrm{s}_2}(i,j)\,, \tag{1}$$

$$I_{\mathrm{SIP}}(i,j)=\begin{cases} I_{\mathrm{d}}(i,j)+I_{\mathrm{s}_2}(i,j) \\ I_{\mathrm{s}_1}(i,j)+I_{\mathrm{s}_2}(i,j)+I_{\mathrm{s}_3}(i,j) \end{cases}, \tag{2}$$

where $I_s(i, j)$ represents the value of pixel in ith row and jth column for single-object label image, $I_d(i, j)$ $I_d(i,j)$ corresponds to double-object label image, and $I_{SIP}(i,j)$ corresponds to three-object label image.

Our image synthesis model as illustrated in Fig.5 consists of two generators {G_1, G_2} and two discriminators {D_1, D_2}. The two generators are responsible for learning global information and detailed information respectively, and the two discriminators have a uniform architecture but operate at different image resolutions.

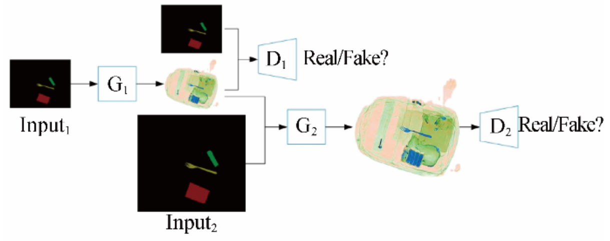


Fig.5 Illustration of image synthesis architecture

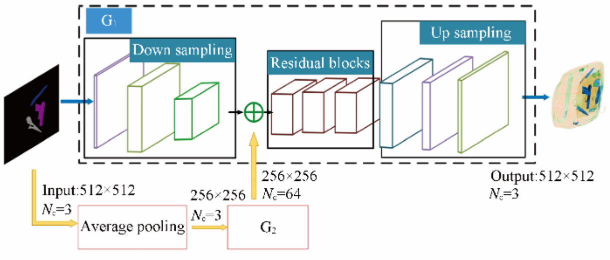


Fig.6 Illustration of generator architecture

We can see complete generator architecture from Fig.6. A semantic label image of resolution 512×512 is used directly as the input of G_1 and also is passed through an average pooling layer to G_2. The G_2 called global generator is described in detail in section 3.2.2. The G_1 called detailed generator as in Fig.6 consists of 4 components: the down sampling, the append operation, residual network, and the up sampling. The down sampling denotes a reflection padding layer, a 7×7 Convolution-InstanceNorm-ReLU layer, a 3×3 Convolution-InstanceNorm-ReLU layer with 64 filters and stride 1. The append operation denotes the addition of the output of the down sampling to the output of G_2. The residual network denotes three residual blocks that contains two 3×3 convolutional layers with 64 filters on both layers. The up sampling denotes 3×3 Transposed Convolution-InstanceNorm-ReLU layer with 32 filters and stride 2, the reflection padding layer and the 7×7 Convolution-InstanceNorm-ReLU layer.

Global generator architecture: Considering that the input and output have the same structure although they are different in appearance[12], we design the global generator G_2. In this generator G_2 shown in Fig.7, the input is down-sampled image through a series of convolutional layers. And then through the residual network which consists of nine residual blocks to deepen the network structure. Last it is up-sampled using transpose convolution.

Novel residual blocks: One of the characteristics of the multi-object image synthesis problem is that the multi-object image has different sizes. For example, the size of a cigarette lighter is significantly smaller than that of a gun. Therefore, the generator is required to have stronger multi-scale feature extraction capability. Therefore, instead of ResNet, the Res2Net as shown in Fig.8(a) is used in our generator as residual blocks.

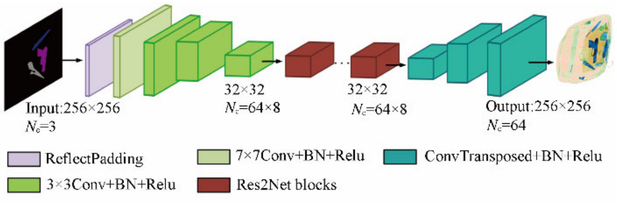
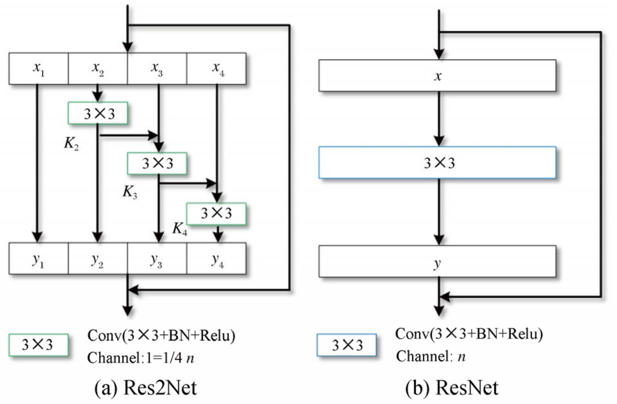


Fig.7 Global generator G₂ architecture



(a) Res2Net (b) ResNet

Fig.8 Comparison between the Res2Net module and ResNet module

Different from the ResNet, the Res2Net[18] replaces the 3×3 filters of n channels by a set of 3×3 filter groups, denoted by $f_i()$, each with m channels ($n=s\times m$). First, it divides the input feature maps into s groups. The first group of feature maps, denoted by x_1, remains unchanged. Then, the second group, denoted by x_2, is extracted by a filter $f_2()$, and this output feature y_2 is added to the third group x_2, which is also taken as the input of a filter $f_3()$. Repeat the process until all input characteristics are processed. Thus, the output y_i can be defined as

$$y_i=\begin{cases}x_i, i=1\\ f_i(x_i), i=2\\ f_i(x_i+y_{i-1}), 1<i\le s\end{cases}. \tag{3}$$

In this way, we can obtain outputs of different receptive fields. In our residual blocks, we set the scale s=4, the receptive field of y_2 is 3×3, the receptive field of feature y_3 is 5×5, the receptive field of feature y_4 is 7×7. The size of receptive field can be described by

$$y_2=x_2*(3\times3)=K_2, \tag{4}$$

$$y_3=(K_2+x_3)*(3\times3)=K_3, \tag{5}$$

$$y_4=(K_3+x_4)*(3\times3)=K_4. \tag{6}$$

Finally, feature maps from all groups are concatenated and sent to 1×1 convolution layer to fuse information altogether. The larger the s, the stronger the multi-scale capability. This split-first-fuse strategy allows convolution to process multi-scale features more efficiently.

Evaluating the quality of synthetic images is a big challenge. To quantify the quality of our results, we used two strategies. We assess the authenticity and diversity using FID[6] scores that measure the distance between the real distribution and the pseudo-distribution, so the lower the score, the better.

Furthermore, two classification networks are used to evaluate the data enhancement effect of synthesized multi-object X-ray security image. We hold that if the synthetic image makes a difference in the training model, the classification network pre-trained with synthetic images can also correctly predict the labels of the real images. The first classification network is FCN-8s, a popular full convolution network segmentation, which combines the feature hierarchy of layers with refined spatial precision output[7]. The second classification network is SSD, a multi-object detection algorithm that directly predicts object categories and bounding box.

TIP[2] and XS-GAN[5] are common method used to synthesis X-ray security image. As shown in Fig.9, TIP is utilized to synthesis image. The fruit knife and scissor cannot blend naturally into the background. The results of XS-GAN are more natural and genuine. Unfortunately, only a prohibited item is considered by both methods.

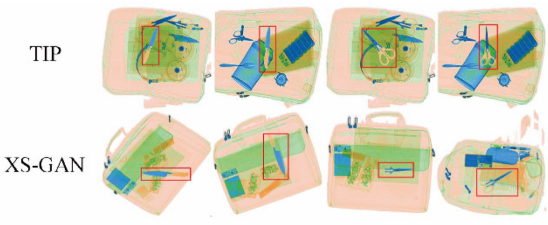


Fig.9 Image samples

Four different GANs are used for comparative experiments. The results are qualitatively evaluated visually and quantitatively evaluated from FID scores.

Visually, as shown in Fig.10, Cycle-GAN is utilized to synthesis the image. The visual quality of the composite image is poor, and the background is seriously distorted. Although the pix2pix model improves the image quality, the resolution is only 256×256. Pix2pixHD network can synthesis images only containing the general outline information and color information, and the detail information is lost for multi-object images where objects have multi-scale. For example, from the multi-scale prohibited items shown in Fig.11, one can see the loss of the internal contour information of the battery, the deformation of the internal structure of the lighter, and the deformation of the fruit knife handle. Compared with the above model, the X-ray security image with multi-scale prohibited items synthesized by the proposed model is more realistic, and the resolution also meets the visual requirements.

Tab.1 shows the FID scores of the four models. It can be seen that the score of our model is the lowest, reaching 30.55, which is better than 47.99, the score of the pix2pixHD model.

Our method is extended by semantic label library, so it can be used to synthesize the image of specified category. As shown in Fig.12, input$_1$ represents two-object semantic image and input$_2$ represents three-object image. We can vary position, kind in the semantic image to generate abundant X-ray security image with prohibited items.

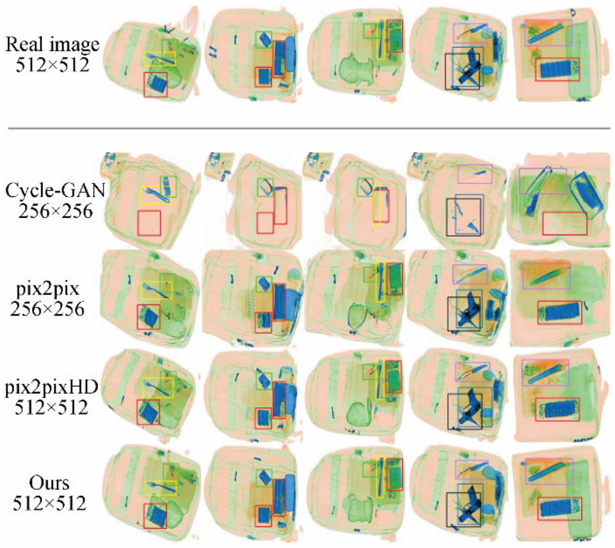


Fig.10 Images samples synthesized by different GANs

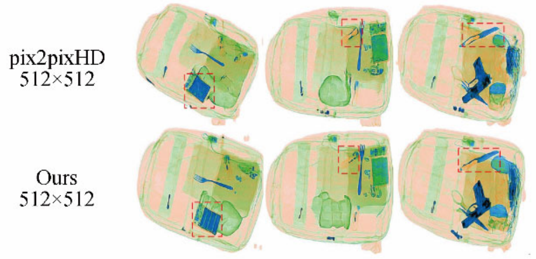


Fig.11 Image samples synthesized by different GANs

Tab.1 FID scores

Model	Score
Cycle-GAN	172.00
pix2pix	78.34
pix2pixHD	47.99
Ours	**30.55**

FCN-8s are pre-trained to use the same number of synthesized images and the same number of real images, respectively, to verify the labels of real images. The results are shown in Tab.2. The Mean IoU of the images synthesized by our model is the best, and it is very close to the result of the oracle image, proving that the synthesized image can also make a difference for training semantic segmentation model.

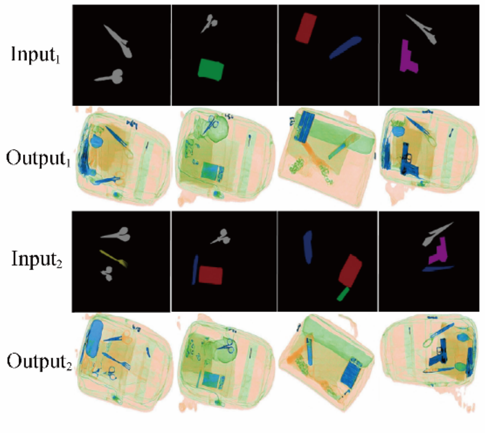


Fig.12 Image samples synthesized by our model

Tab.2 FCN scores

	pix2pixHD	Ours	Oracle
Pixel acc	98.54	98.85	99.00
Mean IoU	0.648 4	0.749 4	0.702 0

SSD is trained by three datasets respectively, as shown in Tab.3. D_{real} is the real image, D_{syn} is the image synthesized by our model, and half of the images in $D_{\text{real+syn}}$ come from D_{real}, and the other half come from D_{syn}. These three datasets have the same number of images and the same test dataset. Although the verification results of the training model using real images are the best, the mAP of using synthetic images reached 0.825, which is also highly accurate, proving the reliability of synthetic images for training classification networks.

Tab.3 Results of SSD with different items

Data	battle	gun	plier	mAP
D_{real}	0.908	0.907	0.899	0.905
D_{syn}	0.843	0.836	0.797	0.825
$D_{\text{real+syn}}$	0.904	0.873	0.804	0.860

In this paper, we propose an approach to address image synthesis task of X-ray security image with multi-scale prohibited items. Furthermore, we build a semantic label library that extends this approach to achieve convenient synthesis of a large number of X-ray security images. The FID score proves the superiority of our model and the authenticity of the synthesized image. In addition, the FCN score and the result of SSD prove the feasibility of using synthetic images to train the classification network.

References

[1] D. Mery, E. Svec and M. Arias, Object Recognition in Baggage Inspection Using Adaptive Sparse Representations of X-ray Images, Cham, Switzerland: Springer, 709 (2015).

[2] Bhowmik N, Wang Q and Gaus Y F A, The Good, the Bad and the Ugly: Evaluating Convolutional Neural Networks for Prohibited Item Detection Using Real and Synthetically Composited X-ray Imagery, arXiv preprint arXiv:1909.11508, (2019).

[3] Mery D, Riffo V and Zscherpel U, Journal of Nondestructive Evaluation **34**, 42 (2015).

[4] Miao C, Xie L, Wan F, Su C, Liu H, Jiao J and Ye Q, SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images, Conference on Computer Vision and Pattern Recognition, 2119 (2019).

[5] Zhao T, Zhang H, Zhang Y and J Yang, X-Ray Image with Prohibited Items Synthesis Based on Generative Adversarial Network, Chinese Conference on Biometric Recognition, 379 (2019).

[6] Heusel M, Ramsauer H, Unterthiner T and Nessler B, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Advances in Neural Information Processing Systems, 6626 (2017).

[7] Hinton G E and Salakhutdinov R, Science **313**, 504 (2006).

[8] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y and Berg A C, SSD: Single Shot Multibox Detector, European Conference on Computer Vision, 21 (2016).

[9] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y, Generative Adversarial Nets, Advances in Neural Information Processing Systems **27**, 2672 (2014).

[10] Bau D, Strobelt H, Peebles W and Wulff J, ACM Transactions on Graphics (TOG) **38**, 1 (2019).

[11] Mo S, Cho M and Shin J, InstaGAN: Instance-aware Image-to-Image Translation, arXiv preprint arXiv:1812.10889, (2018).

[12] Isola P, Zhu J Y and Zhou T, Image-to-Image Translation with Conditional Adversarial Networks, Conference on Computer Vision and Pattern Recognition, 1125 (2017).

[13] Zhu J Y, Park T and Isola P, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, IEEE International Conference on Computer Vision (ICCV), 2223 (2017).

[14] Yi Z, Zhang H, Tan P and Gong M, DualGAN: Unsupervised Dual Learning for Image-To-Image Translation, IEEE International Conference on Computer Vision (ICCV), 2849 (2017).

[15] Kim T, Cha M and Kim H, Learning to Discover Cross-Domain Relations with Generative Adversarial Networks, The 34th International Conference on Machine Learning **70**, 1857 (2017).

[16] Wang T C, Liu M Y and Zhu J Y, pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2018).

[17] Park T, Liu M Y, Wang T C and Zhu J Y, Semantic Image Synthesis with Spatially-Adaptive Normalization, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2337 (2019).

[18] Gao S H, Cheng M M and Zhao K, Res2Net: A New Multi-scale Backbone Architecture, arXiv preprint arXiv:1904.01169, (2019).

[19] Mejjati Y A, Richardt C, Tompkin J, Cosker D and Kim K I, Unsupervised Attention-Guided Image-to-Image Translation, The 32nd International Conference on Neural Information Processing, 3693 (2018).