Unsupervised Object Transfiguration with Attention

Zihan Ye[1] · Fan Lyu[2] · Linyan Li[3] · Yu Sun[1] · Qiming Fu[1] · Fuyuan Hu[1]

Received: 22 July 2018 / Accepted: 5 March 2019 / Published online: 8 April 2019

Abstract

Object transfiguration is a subtask of the image-to-image translation, which translates two independent image sets and has a wide range of applications. Recently, some studies based on Generative Adversarial Network (GAN) have achieved impressive results in the image-to-image translation. However, the object transfiguration task only translates regions containing target objects instead of whole images; most of the existing methods never consider this issue, which results in mistranslation on the backgrounds of images. To address this problem, we present a novel pipeline called Deep Attention Unit Generative Adversarial Networks (DAU-GAN). During the translating process, the DAU computes attention masks that point out where the target objects are. DAU makes GAN concentrate on translating target objects while ignoring meaningless backgrounds. Additionally, we construct an attention-consistent loss and a background-consistent loss to compel our model to translate intently target objects and preserve backgrounds further effectively. We have comparison experiments on three popular related datasets, demonstrating that the DAU-GAN achieves superior performance to the state-of-the-art. We also export attention masks in different stages to confirm its effect during the object transfiguration task. The proposed DAU-GAN can translate object effectively as well as preserve backgrounds information at the same time. In our model, DAU learns to focus on the most important information by producing attention masks. These masks compel DAU-GAN to effectively distinguish target objects and backgrounds during the translation process and to achieve impressive translation results in two subsets of ImageNet and CelebA. Moreover, the results show that we cannot only investigate the model from the image itself but also research from other modal information.

Keywords Multi-modalities · Object transfiguration · Image-to-image translation · Generative Adversarial Networks (GANs) · Attention mechanism · Deep learning

Introduction

Object transfiguration, which is a subtask of image translation, aims to transform a selected kind of object to another in a given image, for example, translating an orange to an apple as shown in Fig. 1. The task relates to image translation [1–4], object-related learning [5–7], and target detection [8–14], and most translation models are trained in supervised setting in which paired training examples are available [1–4]. However, because it is difficult and expensive to obtain paired examples, people use alternative ways to get training examples, for example, as researchers use Photoshop to synthetize training samples in the task of de-rain [15]. Also, some methods use only a few training samples [16] or even one single sample [17] to ease training. Recently, unsupervised image translation [18–22] based on Convolutional Neural Network (CNN) [23] became popular because they can easily translate images by matching image features extracted from CNN between two input images [19]. Specially, methods based on dual-structure Generative Adversarial Network (GAN) [20–22] achieve great performance. The "dual structure" means a cycle consisting of two translating directions (such as horse → zebra and zebra → horse). In such structures, the cycle-consistent loss is used to provide a measurement of image changes in this cycle.

However, most of the existing unsupervised approaches fail to distinguish target objects and backgrounds clearly. These methods tend to damage backgrounds (such as

Zihan Ye and Fan Lyu contributed equally to this work.

✉ Linyan Li
linyan_li@qq.com

✉ Fuyuan Hu
fuyuanhu@mail.usts.edu.cn

[1] Suzhou University of Science and Technology, Suzhou 215009, China

[2] Tianjin University, Tianjin 300072, China

[3] Suzhou Institute of Trade & Commerce, Suzhou 215009, China

deformation, discoloration). As shown in Fig. 1 a, the words on the billboard are blurred when being translated by CycleGAN. In Fig. 1 b, although green apples and bananas are not target objects, CycleGAN purples the green apples and makes the bananas red. To address this problem, in this paper, we propose an adversarial object transfiguration model by constructing the novel Deep Attention Unit (DAU) to preserve backgrounds of translated image. Our approach is inspired by the attention mechanism [24–26] which tries to select the most important information from a large amount of data. DAU computes attention masks that point out where the target objects are and then our GAN translates images under the guidance of these masks. In addition, we also construct a background-consistent loss and an attention-consistent loss to train the DAU-GAN. We evaluate our model on two subsets of ImageNet (apple–orange and horse–zebra) and CelebA. The results show that the proposed DAU-GAN keeps images' backgrounds and translates the target object effectively too. For example, in Fig. 1 a, the billboard translated by our model is clearer than the other translated by CycleGAN in Fig. 1 b, and our model successfully preserves other fruits' color.

Our main contributions are threefold: (1) introducing the attention mechanism into GAN to improve target cognitive ability; (2) proposing a novel framework called DAU-GAN, which can preserve backgrounds of translated images by the assistance of attention masks in object transfiguration task; (3) constructing a background-consistent loss and an attention-consistent loss to enhance the effectiveness of translation by multi-modal information.

Related Works

Generative Adversarial Network

GAN [27] was developed to solve many problems, like image generation [4, 28], image editing [29], image classification [30], and object recognition [31]. The key of GAN's success is by leading a generator and a discriminator to reciprocally compete. The generator generates data indistinguishable with real image, and the discriminator distinguishes between the generated image and real image. Isola et al. [2] studied the potential of GAN in image translation. After that, a cycle-consistent loss [21, 22] was proposed for unsupervised image translation. It assumes that, after an image is translated, it can be translated back to the original image, e.g., horse → zebra → horse.

Multi-Modal Deep Learning

Deep learning in single-modal tasks has been studied widely, and researchers gradually explore the influence of deep learning on multi-modal tasks. For example, Reed et al. [32] learned to map from images to captions and Zhang et al. [4] learned how to draw a painting with text prior. Multi-modal learning also exists between different kinds of image. Huang et al. [33] decomposed an image representation into a content code and a style code, and translated an image from one domain into many other domains. Inspired by this, we construct a model to simultaneously learn a translation modality and another special modality—the attention modality.

Attention Mechanism

We take the attention [33, 34] in image as a special modality, which can instruct the model to distinguish target objects and backgrounds during the learning process. The concept of attention originates in human physiology [11, 33, 35, 36] and was introduced to deep learning fields later [24, 26]. The essence of attention is a learned soft distribution to make the model only care about interesting parts and ignore meaningless parts. Some GAN-based models apply attention mechanism, e.g., DA-GAN [37], to predict attention frame and crop framed parts in image, then encode these parts. We also integrate the attention mechanism into GAN, but our approach extracts attention masks from feature maps, which is much different from DA-GAN. Additionally, we refer to some structure of residual networks [38–40] to decrease the impact of vanishing gradient.

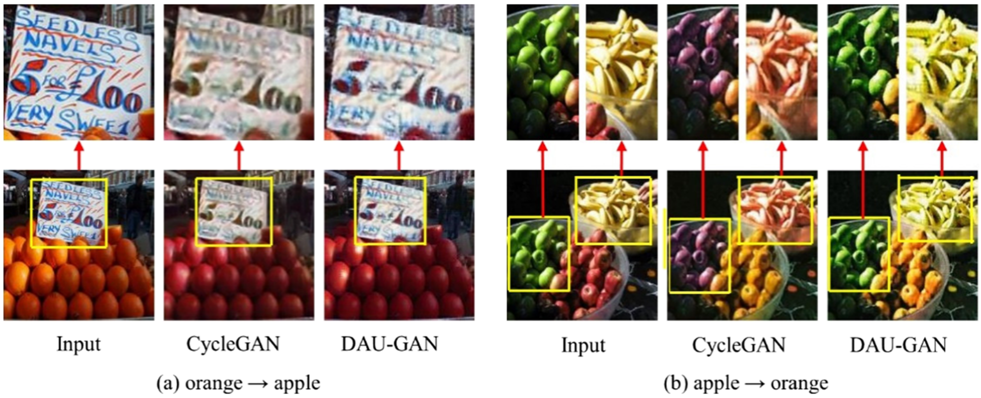


Fig. 1 Yellow rectangles represent the regions obviously damaged by CycleGAN during translating. For example, in **a**, the words on the billboard are exceedingly blurred; in **b**, green apples are purpled and bananas reddened

DAU-GAN

Proposed Architecture

The goal of object transfiguration is to transform a specific object in a given image to another different type. It is often considered as a domain-transforming problem: $G : X \rightarrow Y$, where G is the learned mapping function, X is the source domain, and Y is the target domain.

Our overall architecture is shown in Fig. 2. The upper part shows an overall framework for object transfiguration and the lower part shows the details of the generator. Our model is based on a dual structure, which consists of two generators, G and F, and two discriminators, D_X and D_Y. To ensure the generated images $G(\mathrm{x}) = y^{'}$ and $F(\mathrm{y}) = x^{'}$ are in the corresponding domains, we employ two discriminators D_X and D_Y to distinguish the real images from synthetic ones. D_X or D_Y takes an image as input and outputs a likelihood that the input is a natural image from domain X or Y.

Our model accepts $(\mathrm{x}, \mathrm{y}) \in (\mathrm{X}, \mathrm{Y})$ as inputs. For the translation $\mathrm{G} : \mathrm{X} \rightarrow \mathrm{Y}$, it outputs an image $\mathrm{G(x)}$ in domain Y that preserves the background $\mathrm{B(x)}$ of x and translates objects in the images. For the translation $\mathrm{F} : \mathrm{Y} \rightarrow \mathrm{X}$, the output is $\mathrm{F(y)}$ in domain X, preserving the background $\mathrm{B(y)}$ during the translation.

Deep Attention Unit

We propose a novel DAU on the top of dual learning inspired by the dramatic success of the attention mechanism [41, 42]. DAU helps the model filter out backgrounds and objects from an image to preserve backgrounds. In CNN, the feature maps from each convolutional layer can be divided into many regions. DAU can be attached to any convolutional layer, and compute attention masks for the corresponding feature maps. In other words, the DAU forces the model to focus the translation process on the specific objects for the attention mask which will highlight the translated objects.

In CNN, feature maps from each convolutional layer can be divided into many regions. DAU, which is attached to any convolutional layer, effectively computes the attention masks for the corresponding feature maps. In other words, DAU can force the model to focus the translation process on the specific

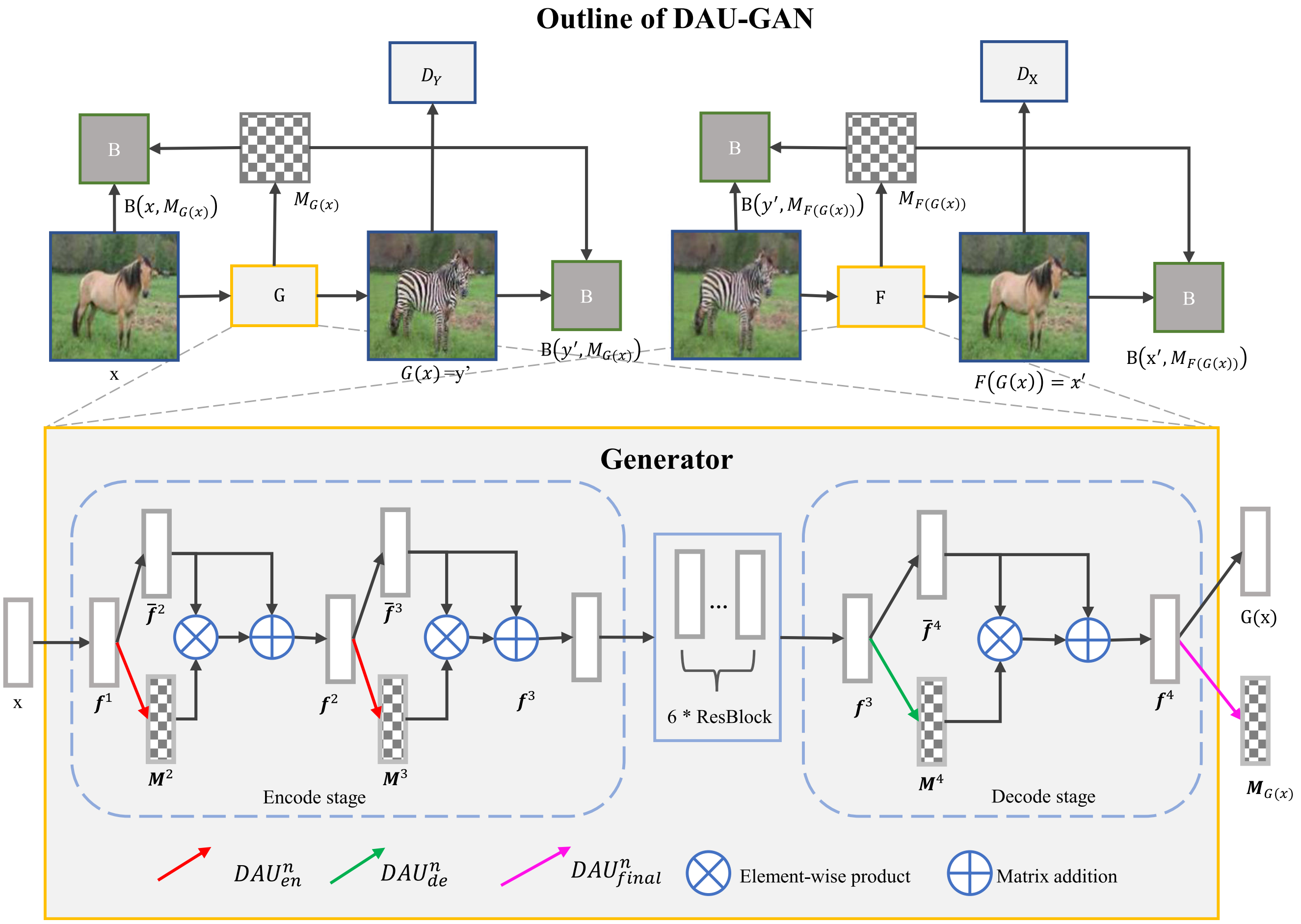


Fig. 2 Architecture of DAU-GAN: the upper part is the outline of our model, and the lower part is the generator in detail

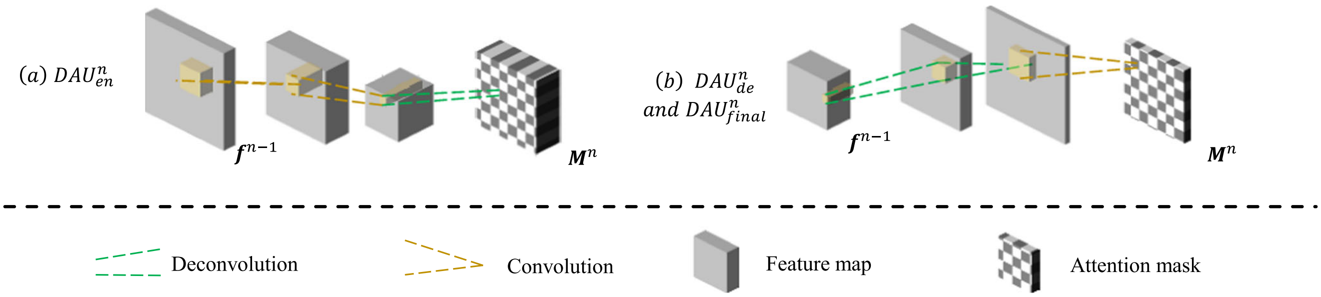


Fig. 3 Structures of different types of DAU. $\mathrm{DAU}_{\mathrm{en}}$ for the encode stage, $\mathrm{DAU}_{\mathrm{de}}$ for the decode stage, and $\mathrm{DAU}_{\mathrm{final}}$ for the output stage

objects. The main reason is that the attention mask can separate objects and backgrounds, and highlight translated objects.

In this paper, we construct three types of DAU, i.e., $\mathrm{DAU}_{\mathrm{en}}$, $\mathrm{DAU}_{\mathrm{de}}$,and $\mathrm{DAU}_{\mathrm{final}}$ for the encode stage, decode stage, and output stage respectively. For $\mathrm{DAU}_{\mathrm{en}}$ and $\mathrm{DAU}_{\mathrm{de}}$, we set the shapes equal to the corresponding feature map, while for $\mathrm{DAU}_{\mathrm{final}}$, we shrink the depth to 1 for the convenience of visualization.

We regard the feature map from the *n*th convolutional layer as $\mathbf{f}^n$. As shown in Fig. 3, DAU extracts the mask M_n from $\mathbf{f}^{n-1}$, i.e., $\mathbf{M}^n = \mathrm{DAU}(\mathbf{f}^{n-1})$. Specifically, following the dual adversarial learning architecture, we divide the DAU into two main categories for encoder (DAU_{en}) and decoder ($\mathrm{DAU}_{\mathrm{de}}$). For the encoder, $\mathbf{f}^{n-1}$ firstly passes through two convolutional layers with the ReLU [43] activation function (ReLU(x) = max(0, x)), and then is sent to one deconvolutional layer with the Sigmoid activation function ($\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}}$). The process is denoted as follows:

$$\mathrm{DAU}_{\mathrm{en}}^{\mathrm{n}}(\mathrm{x}) = \mathrm{Conv}^{-1}(\mathrm{Conv}(\mathrm{Conv}(\mathrm{x}))), \tag{1}$$

$$\mathbf{M}^{\mathrm{n}} = \mathrm{DAU}_{\mathrm{en}}^{\mathrm{n}}\left(\mathbf{f}^{\mathrm{n}-1}\right). \tag{2}$$

We define $\overline{\mathbf{f}}^{\mathrm{n}}$ as the untreated (*n*)th feature map. With the mask $\mathbf{M}^{\mathrm{n}}$, we obtain the enhancive feature map by the elements-wise product: $\mathrm{H}\left(\mathbf{M}^{\mathrm{n}}, \overline{\mathbf{f}}^{\mathrm{n}}\right)$. In addition, we take a residual architecture [38–40] that adds a shortcut to depress the impact of vanishing gradient. After these operations, the enhancive (*n*)th feature map is computed by:

$$\mathbf{f}^{\mathrm{n}} = \overline{\mathbf{f}}^{\mathrm{n}} + \mathrm{H}\left(\mathbf{M}^{\mathrm{n}}, \overline{\mathbf{f}}^{\mathrm{n}}\right). \tag{3}$$

In the decoder, the DAU is inverse. In other words, the feature map firstly passes through two deconvolutional layers and then go by one convolutional layer. The DAU of the decoder can be represented as follows:

$$\mathrm{DAU}_{\mathrm{en}}^{\mathrm{n}}(\mathrm{x}) = \mathrm{Conv}\left(\mathrm{Conv}^{-1}\left(\mathrm{Conv}^{-1}(\mathrm{x})\right)\right). \tag{4}$$

In the output layer, the translated image $\mathrm{y}' = \mathrm{G}(\mathrm{x})$ and the corresponding attention mask $\mathbf{M}_{\mathrm{G}(\mathrm{x})}$ would simultaneously be exported. As shown in Fig. 4, because the elements of mask represent the probabilities that the object is in the corresponding position, we get the background of input image x from the equation: $\mathrm{B}(\mathrm{x}, \mathbf{M}_{\mathrm{G}(\mathrm{x})}) = \mathrm{H}(\mathrm{x}, 1 - \mathbf{M}_{\mathrm{G}(\mathrm{x})})$. We integrate DAU in each layer of the generator to enhance related features, even though we use only the last attention mask to compute $\mathrm{L}_{\mathrm{att}}$ and L_{bg}.

Attention-Consistent Loss and Background-Consistent Loss

The attention mask $\mathbf{M}_{\mathrm{G}(\mathrm{x})}$ of G(x) should be the same with the attention mask $\mathbf{M}_{\mathrm{F}(\mathrm{G}(\mathrm{x}))}$ of the translated image

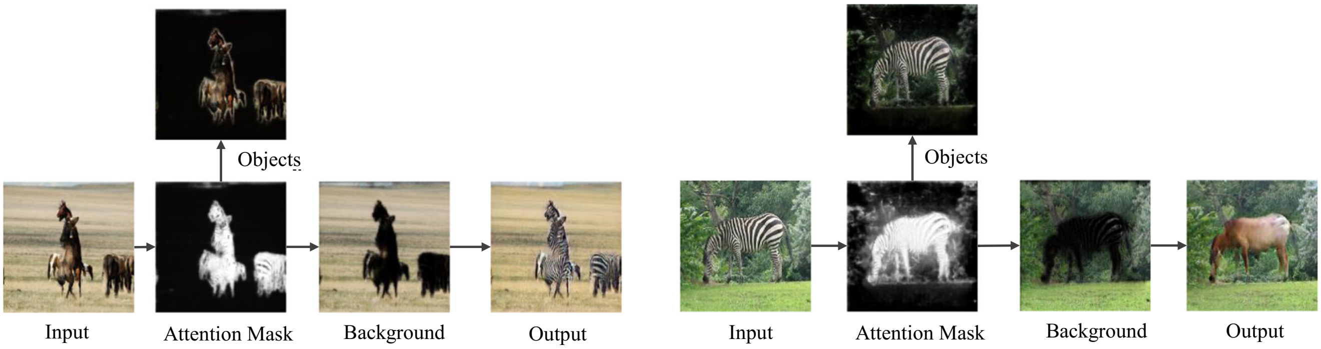
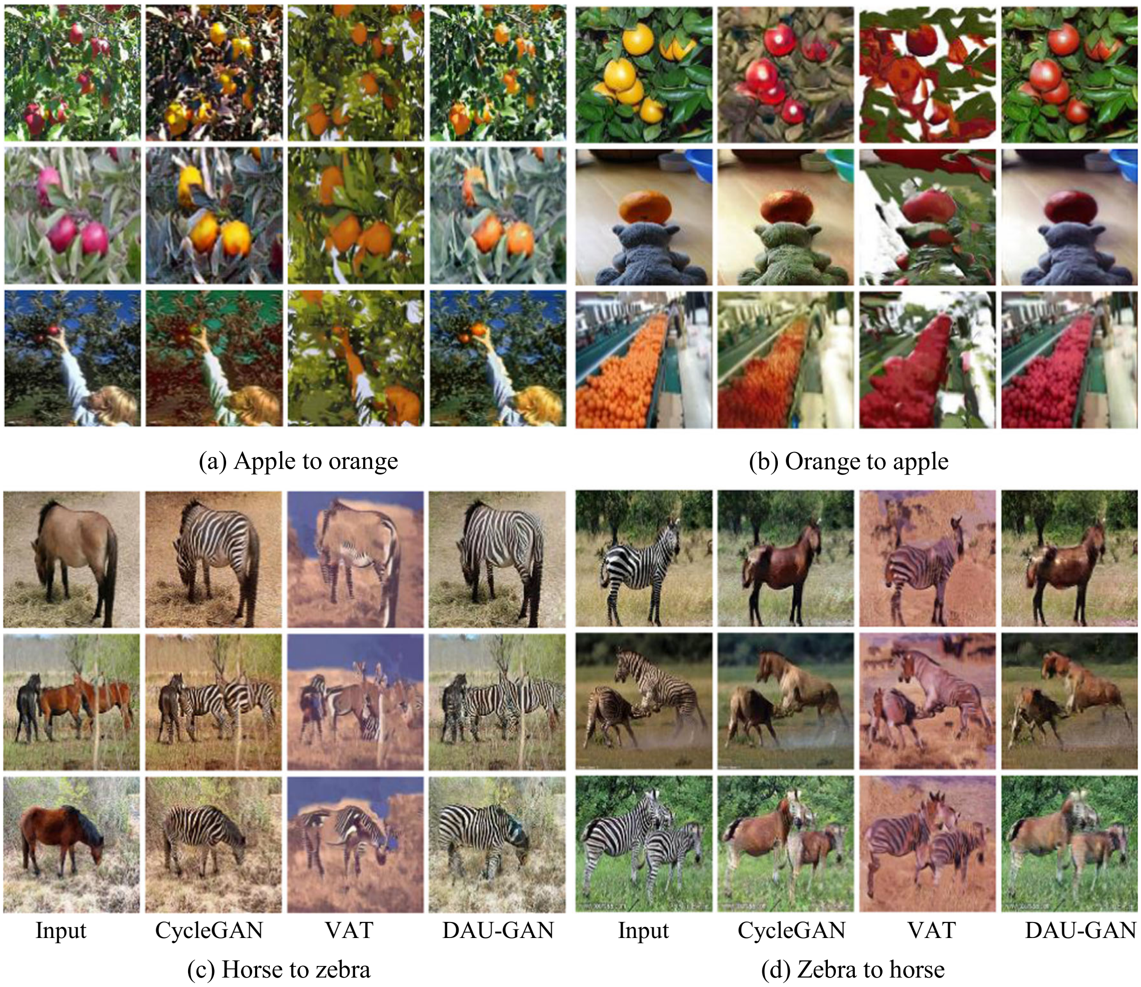


Fig. 4 Object(s) and background can be separated by DAU (horse → zebra)



(a) Apple to orange

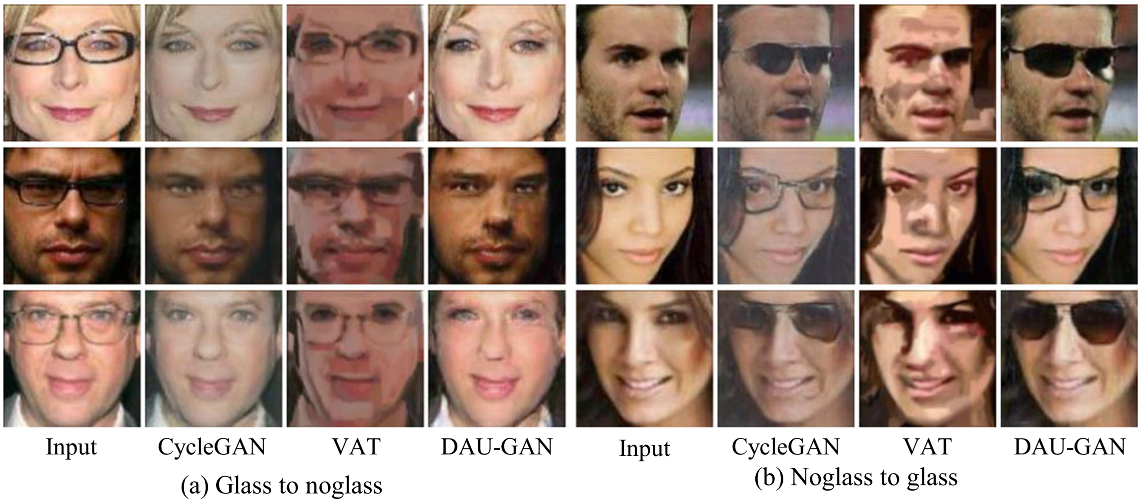
(b) Orange to apple

(c) Horse to zebra

(d) Zebra to horse

Fig. 5 Experimental results on two subsets of ImageNet

F(G(x)). For instance, in the translated process G, a horse is transformed into a zebra, while in the translated process F, the zebra is transformed back to a horse. The regions of the horse and zebra should be the same, i.e. $\mathbf{M}_{G(x)} = \mathbf{M}_{F(G(x))}$. Considering that the position of objects during translating is invariable, we construct an attention-consistent loss to improve the performance of predicting attention masks:



(a) Glass to noglass

(b) Noglass to glass

Fig. 6 Experimental results on CelebA

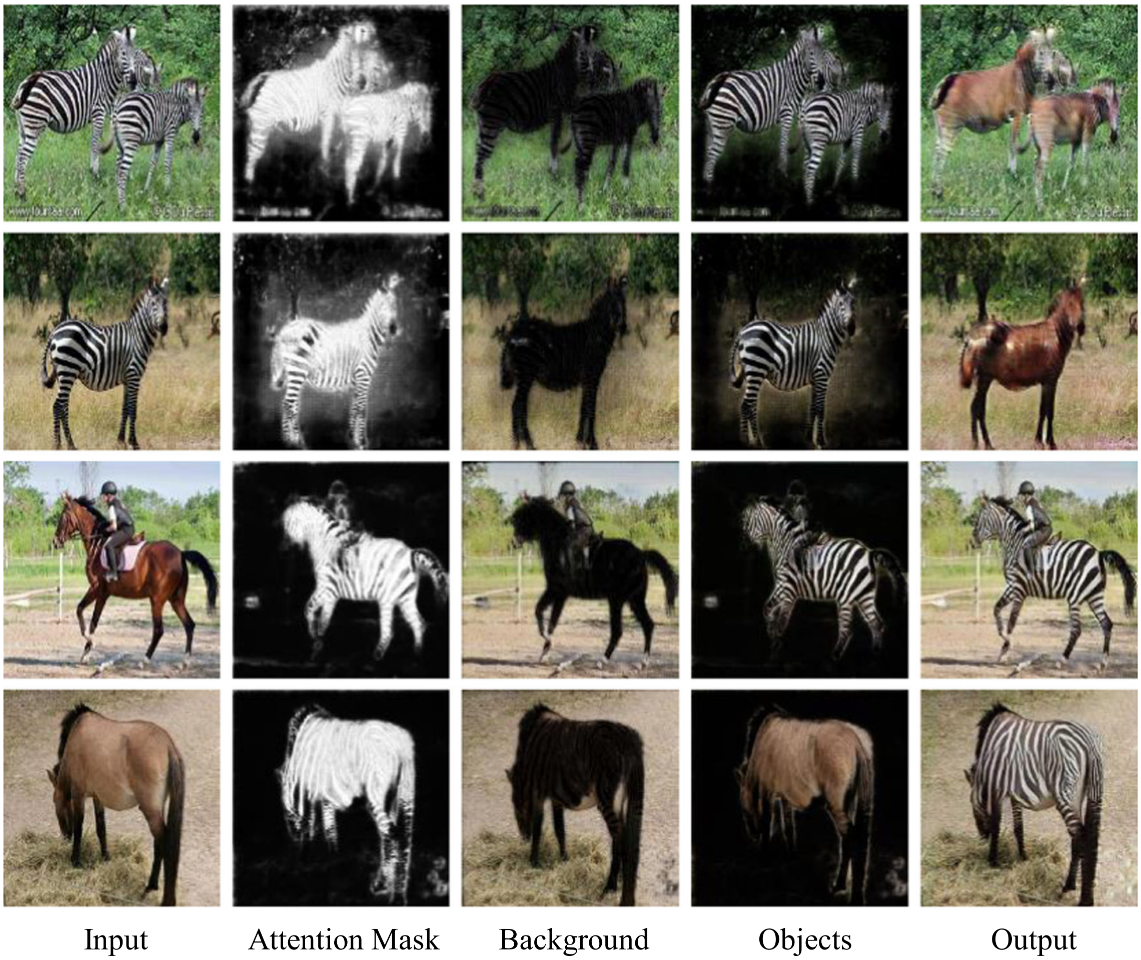


Fig. 7 Attention masks after different training steps (horse → zebra)

$$
\mathrm{L}_{\mathrm{att}}(\mathrm{x},\mathrm{G},\mathrm{F}) = \alpha * \left\| \mathbf{M}_{\mathrm{G(x)}} - \mathbf{M}_{\mathrm{F(G(x))}} \right\|_1 + \beta * \left(\mathbf{M}_{\mathrm{G(x)}} + \mathbf{M}_{\mathrm{F(G(x))}} \right), \tag{5}
$$

where the second term is a regularization, which prevents the overfitting problem. When DAU predicts an overweight attention mask, e.g., a mask that every element's value is 1, the value of this attention-consistent loss is too big. We add the term to encourage our model to focus on some special parts of an image instead of a whole image.

With an attention mask **M**, DAU-GAN can distinguish objects and backgrounds. Similar to regions of target objects, backgrounds should also be invariable. Thus, we construct a background-consistent loss to preserve backgrounds:

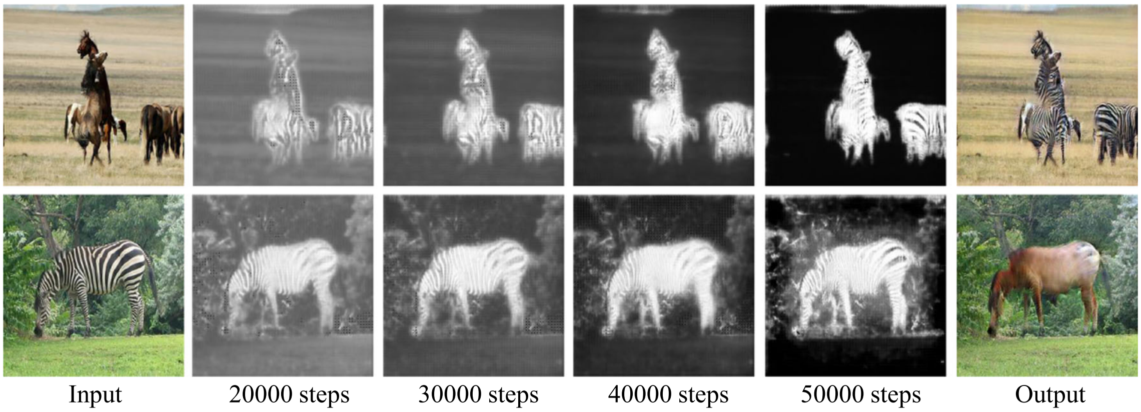
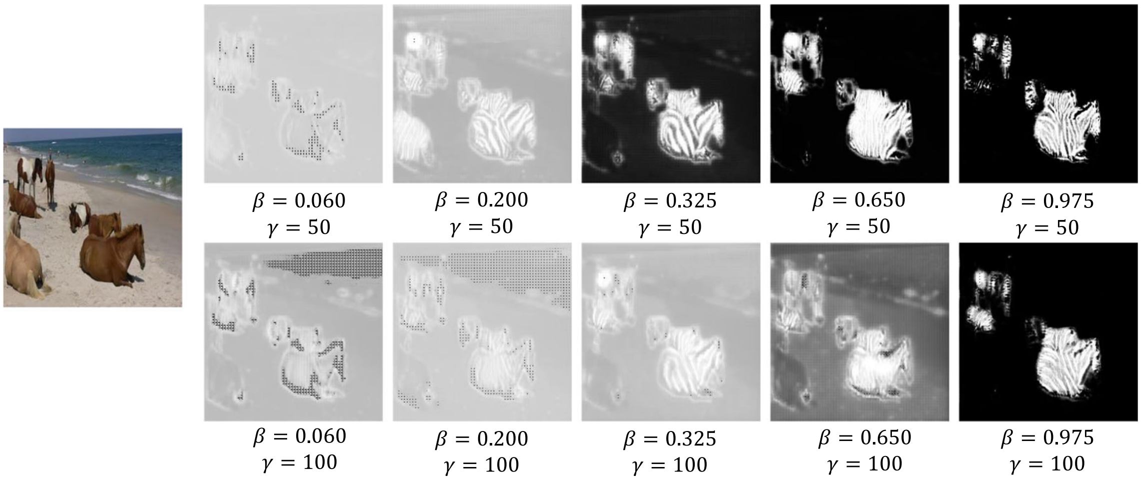


Fig. 8 Attention masks after different training steps (horse → zebra)


$\beta = 0.060$ $\gamma = 50$
$\beta = 0.200$ $\gamma = 50$
$\beta = 0.325$ $\gamma = 50$
$\beta = 0.650$ $\gamma = 50$
$\beta = 0.975$ $\gamma = 50$
$\beta = 0.060$ $\gamma = 100$
$\beta = 0.200$ $\gamma = 100$
$\beta = 0.325$ $\gamma = 100$
$\beta = 0.650$ $\gamma = 100$
$\beta = 0.975$ $\gamma = 100$

(a) Input

(b) Attention mask

Fig. 9 Attention masks with different 훽 and 훾

$$L_{bg}(x, G) = \gamma * \left\| B\left(x, \mathbf{M}_{G(x)} - B\left(G(x), \mathbf{M}_{G(x)}\right)\right\|_1. \quad (6)$$

Experimental results have verified that our model effectively preserves backgrounds during a successful translation process.

Full Objective

From [27], the standard adversarial losses are:

$$L_{GAN}(X, Y, G, D_Y) = \log(D_Y(y)) + \log\left(1 - D_Y\left(y'\right)\right), \quad (7)$$

$$L_{GAN}(Y, X, F, D_X) = \log(D_X(x)) + \log\left(1 - D_X\left(x'\right)\right). \quad (8)$$

Then, because we adopt the dual structure, we also construct a cycle-consistent loss to improve the performance of our model. The cycle-consistent loss assumes that, for each image x in domain X, our model is able to be able to translate $y' = G(x)$ to the original image. It is formulated as: $F(y') = F(G(x)) \approx x$. Similarly, for each image y in domain Y, the assumption is: $G(x') = G(F(x)) \approx y$. The cycle-consistent loss is as follows:

$$L_{cyc}(G, F) = \|F(G(x)) - x\|_1 + \|G(F(y)) - y\|_1. \quad (9)$$

Our full objective is given:

$$\begin{aligned} L(G, F, D_X, D_Y)| = {} & L_{GAN}(X, Y, G, D_Y) + L_{GAN}(Y, X, F, D_X) \\ & | + L_{att}(x, G, F) + L_{att}(y, F, G) \\ & | + L_{bg}(x, G) + L_{bg}(y, F) + L_{cyc}(G, F). \end{aligned} \quad (10)$$

Therefore, we solve the optimizing problem:

$$G^*, F^* = \arg\min_{G,F}\max_{D_X, D_Y} L(G, F, D_X, D_Y). \quad (11)$$

We alternately optimize our generators and our discriminators. For every round of optimization, first discriminators' parameters are updated at one time while generators' parameters are fixed. Then, generators' parameters are modified at one time while discriminators' parameters are fixed. At the time that discriminators' accuracy is the highest and discriminators cannot distinguish real images from fake images produced by generators, our model is optimal.

Table 1 The average change loss for pixels of each translated image's background

Model	Translation					
	Apple → orange	Orange → apple	Horse → zebra	Zebra → horse	Noglass → glass	Glass → noglass
VAT	75.79	154.03	107.79	46.53	69.21	138.10
CycleGAN	189.52	333.84	309.92	140.22	159.54	236.23
DAU-GAN	40.27	51.56	67.18	34.75	44.91	74.28

Experimental Results

Datasets and Setting

We first experiment our model on two subsets (apple–orange and horse–zebra) of ImageNet [44], which is a large-scale labeled image dataset for computer vision. The numbers of training images of apple, orange, horse, and zebra are 995, 1019, 1067, and 1334 respectively. We also experiment on CelebA that includes 10,177 face images. Each image has been labeled as different characteristics. We select 1800 images, half of them with glasses while the others are without. Then, we use these images in the translation: w/ glasses ↔ w/o glasses.

We run our code on one GeForce GTX 1080 GPU card. More specifically, we adapt the architecture [1] as the backbone of our generators, which includes six blocks for 128 × 128 training images, and nine blocks for 256 × 256 or higher resolution training images. Our discriminators are constructed by using 70 × 70 PatchGAN [2]. For ImageNet, we respectively set α, β, and γ on Eqs. (5) and (6) to 1, 0.325, and 50. For CelebA, β is set as 0.240. We use the Adam solver with a batch size of 1 and a learning rate of 0.0002 to optimize the model.

Performance Comparison

Figures 5 and 6 show the experimental results on ImageNet and CelebA, respectively. Obviously, the impact of backgrounds in those images translated by CycleGAN and VAT is very remarkable. For example, in the second column in Fig. 5 a and b, leaves are discolored from green to gray. As shown in Fig. 6, the translation of VAT completely fails. Although CycleGAN successfully removes the glasses in Fig. 6 b, the human skin is whitened. In contrast, our approach effectively preserves backgrounds besides successfully translating objects that we want to translate. For example, in Fig. 5 c, zebras produced by our approach have more natural streaks besides preserved backgrounds, and in Fig. 6 b, people are with glasses by our approach, although the people's eyes seem different after removal of glasses. It might be because our model missed pixel-level information in the convolutional process. We will try to use U-net structure [45] to preserve pixel-level information in the future.

Attention Masks Analysis

Moreover, we visualize attention masks in Fig. 7 to confirm DAU's effectiveness, where the regions in black and white represent backgrounds and objects, respectively. We can see that DAU can correctly predict those regions where objects occur. In addition, we export attention masks in different training steps in Fig. 8. The results show that high value gradually concentrates on where the object is. It verifies that DAU does help our model learn to distinguish backgrounds and objects expected to be translated.

To explore the effect of hyperparameter for attention masks, we export attention masks in different β and γ. The results are shown in Fig. 9. For each raw of Fig. 9 b, γ is fixed but β gets higher. And the contrast of attention masks is improved when β gets higher. It means that our model obtains more concentrated attention. Besides comparing the attention masks that have the same β but different γ, we find that the contrast of attention masks is improved when γ gets lower. We find that with a higher ratio of β to γ, there is more concentrated attention.

Quantitative Evaluation

At present, researchers have proposed some metrics to measure quantitatively the models' performances [46, 47]. To accurately demonstrate our approach, we use the average pixel difference of translated images to measure quantitatively our model's effect on test sets. For every translated image, we sum the average pixel difference of backgrounds in every channel. The result is shown in Table 1. Although, sometimes, the difference of VAT is less than CycleGAN's, the translation results are dissatisfied. Not only does our model accomplish translation task in Fig. 7, but it also has the minimum difference in backgrounds for all kinds of translation as shown in Table 1. It is a strong evidence that indicates that our model preserves the background during the object transfiguration task.

Conclusions

In this paper, we propose Deep Attention Unit (DAU) and construct the DAU-GAN for task of object transfiguration. DAU helps our model focus on the most significant information by predicting attention masks, which instructs DAU-GAN to effectively distinguish specific objects and backgrounds during the translation process and achieve impressive translation results in two datasets. Moreover, the results show that our model can learn from the image, and can also learn from other modal information. In the future, we plan to design more robust DAU structures to make model train easier.

Funding This work was supported by the National Natural Science Foundation of China (No. 61876121, 61472267, 61728205, 61502329, 61672371), Primary Research & Development Plan of Jiangsu Province (No. BE2017663), Aeronautical Science Foundation (20151996016), and Jiangsu Key Disciplines of Thirteen Five-Year Plan (No. 20168765) and Suzhou Institute of Trade & Commerce Research Project(KY-ZRA1805).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. ECCV. 2016:694–711.
2. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. CVPR. 2017:1125–34.
3. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image superresolution using a generative adversarial network. CVPR. 2017:4681–90.
4. Zhang H, Xu T, Li H, Zhang S, Huang X, Wang X, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. CVPR. 2017:5907–15.
5. Feng Y, Ren J, Jiang J. Object-based 2d-to-3d video conversion for effective stereoscopic content generation in 3d-tv applications. IEEE Trans Broadcast. 2011;57(2):500–9.
6. Ren J, Jiang J, Wang D, Ipson S. Fusion of intensity and inter-component chromatic difference for effective and robust colour edge detection. IET Image Process. 2010;4(4):294–301.
7. Zabalza J, et al. Novel segemented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. Neurocomputing. 2016;185:1–10.
8. Han J, Zhang D, Hu X, Guo L, Ren J, Wu F. Background prior-based salient object detection via deep reconstruction residual. TCSVT. 2015;25(8):1309–21.
9. Yan Y, Ren J, Zhao H, Sun G, Wang Z, Zheng J, et al. Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn Comput. 2018;10(1):94–104.
10. Han J, Zhang D, Cheng G, Guo L, Ren J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans Geosci Remote Sens. 2015;53(6):3325–37.
11. Gao F, Zhang Y, Wang J, Sun J, Yang E, Hussain A. Visual attention model based vehicle target detection in synthetic aperture radar images: a novel approach. Cogn Comput. 2015;7(4):434–44.
12. Gao F, You J, Wang J, Sun J, Yang E, Zhou H. A novel target detection method for SAR images based on shadow proposal and saliency analysis. Neurocomputing. 2017;267:220–31.
13. Gao F, Ma F, Wang J, et al. Visual saliency modeling for river detection in high-resolution SAR imagery. IEEE Access. 2018;6: 1000–14.
14. Gao F, Ma F, Zhang Y, Wang J, Sun J, Yang E, et al. Biologically inspired progressive enhancement target detection from heavy cluttered SAR images[J]. Cogn Comput. 2016;8(5):955–66.
15. Fu X, Huang J, Zeng D, Huang Y, Ding X, Paisley J. Removing rain from single images via a deep detail network. CVPR. 2017:3855–63.
16. Shufei Zhang et al. Learning from few samples with memory network, cognitive computation, 2018; 10(1) 15–22.
17. Luo C, et al. Zero-shot learning via attribute regression and class prototype rectification. IEEE Transactions on Image Processing. 2018;27(2):637–48.
18. Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks. Advances in Neural Information Processing Systems. 2017:700–8.
19. Liao J, Yao Y, Yuan L, Hua G, Kang SB. Visual attribute transfer through deep image analogy. ACM Trans Graph. 2017;36(4):120.
20. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J. Stargan: unified generative adversarial networks for multi-domain image-to-image translation. arXiv preprint. 2017;arXiv:1711.09020.
21. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. CVPR. 2017: 2223–32.
22. Yi Z, Zhang H, Tan P, Gong M. Dualgan: unsupervised dual learning for image-to-image translation. CVPR. 2017:2849–57.
23. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep-convolutional neural networks. NIPS. 2012:1097–105.
24. Zhao B, Feng J, Wu X, Yan S. A survey on deep learning-based fine-grained object classification and semantic segmentation. Int J Autom Comput. 2017;14(2):119–35.
25. Yan Y, Ren J, Sun G, Zhao H, Han J, Li X, et al. Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. Pattern Recogn. 2018;79:65–78.
26. Aboudib A, Gripon V, Coppin G. A biologically inspired framework for visual information processing and an application on modeling bottom-up visual attention. Cogn Comput. 2016;8(6): 1007–26.
27. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: NIPS. 2014:2672–80.
28. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint. 2015;arXiv:1511.06434.
29. Zhu JY, Kr¨ahenb¨uhl P, Shechtman E, Efros AA. Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision. 2016:597–613.
30. Gao F, Huang T, Wang J, Sun J, Hussain A, Yang E. Dual-branch deep convolution neural network for polarimetric SAR image classification. Appl Sci. 2017;7(5):447.
31. Gao F, Yang Y, Wang J, et al. A deep convolutional generative adversarial networks (DCGANs)-based semi-supervised method for object recognition in synthetic aperture radar (SAR) images. Remote Sens, 2018, 10(6).
32. Reed, Scott and Akata, Zeynep and Yan, Xinchen and Logeswaran, Lajanugen and Schiele, Bernt and Lee, Honglak.: Generative adversarial text to image synthesis. In: ICML. 2016: 1060–1069.
33. Huang X, Liu MY, Belongie S, et al. Multimodal unsupervised image-to-image translation. arXiv preprint. 2018;arXiv: 1804.04732.
34. Zhu JY, Zhang R, Pathak D, Darrell T, Efros AA, Wang O, et al. Toward multimodal image-to-image translation. NIPS. 2017:465–76.
35. Briggs F, Mangun GR, Usrey WM. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature. 2013;499(7459):476–80.
36. Wang Z, Ren J, Zhang D, et al. A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing. 2018;289:68–83.
37. Ma S, Fu J, Chen CW, Mei T. DA-GAN: instance-level image translation by deep attention generative adversarial networks (with supplementary materials). CVPR. 2018:5657–66.
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CVPR. 2016:770–8.
39. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. Residual attention network for image classification. CVPR. 2017:3156–64.
40. Liu X, Deng Z. Segmentation of drivable road using deep fully convolutional residual network with pyramid pooling. Cogn Comput. 2018;10(2):272–81.
41. Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. CVPR. 2017:4438–46.
42. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: neural image caption generation with visual attention. ICML. 2015:2048–57.
43. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. ICAIS. 2011:315–23.

44. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. CVPR. 2009:248–55.

45. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234–241.

46. Yang P, Huang K, Liu CL. Geometry preserving multi-task metric learning. Mach Learn. 2013;92(1):133–75.

47. Yang X, Huang K, Zhang R, et al. Learning latent features with infinite nonnegative binary matrix trifactorization. TETCI. 2018;99:1–14.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.