A Novel Visual Representation on Text Using Diverse Conditional GAN for Visual Recognition

Tao Hu, *Member, IEEE*, Chengjiang Long, *Member, IEEE*, and Chunxia Xiao, *Member, IEEE*

***Abstract*— Automatic image visual recognition can make full use of largely available images with text descriptions on social media platforms to build large-scale image labeled datasets. In this paper, we propose a novel visual text representation, named DG-VRT (Diverse GAN-Visual Representation on Text), which extracts visual features from synthetic images generated by a diverse conditional Generative Adversarial Network (DCGAN) on the text, for visual recognition. The DCGAN incorporates the current state-of-the-art text-to-image GANs and generates multiple synthetic images with various prior noises conditioned on a text. Then we extract deep visual features from the generated synthetic images to explore the underlying visual concepts and provide a visual transformation on text in feature space. Finally, we combine image-level visual features, text-level features and visual features based on synthetic images together to recognize the images, and we also extend the proposed work to semantic segmentation. We conduct extensive experiments on two benchmark datasets and the experimental results demonstrate the efficacy of our proposed representation on text for visual recognition.**

***Index Terms*— Visual representation, diverse conditional GAN, visual recognition.**

I. Introduction

NOWADAYS, images are being taken and shared to be commented on an unprecedented rate among social networks like Facebook, Twitter, and Flickr. Images in these social media platforms do not exist in isolation and most images on the web carry rich text information including informative and semantic signals like who takes the photo, and where and with whom. Therefore, it is desirable to explore social media context, especially text context information jointly with pixel information, to aid visual recognition tasks on images.

Prior work takes advantage of text context information to improve visual recognition by various treatments like selecting top frequently used words and user-generated tags [1]–[4], extracting text-level feature representation with Text CNNs [5], and exploring visual feature representation from a set of neighbor images [6] defined based on Jaccard similarity between image metadata. The intuition behind is that images with similar text context information tend to depict similar scenes.

To improve the accuracy of recognition on small sample image dataset, we intend to use the text information of each image sample, which effectively expand the original image features. Inspired by the development of text-to-image Generative Adversarial Networks (GANs) [7]–[12], which can generate high-resolution and photo-realistic synthetic images conditioned on the text, we propose a novel visual representation on text, named as "DG-VRT", by representing text information using visual concepts extracted from a series of visually plausible synthetic images generated by a diverse conditional GAN (DCGAN), as illustrated in Fig. 2. This is consistent with a popular saying "As there are 1000 Hamlets, there are 1000 readers." Usually, given a text that describes a specific scene, different readers can imagine different relevant visual scenes in their brains. One synthetic image is not sufficient to simulate what multiple readers can visually imagine from the single text information.

Instead of using K individual text-to-image GANs like StackGAN++ [9] and AttnGAN [7] to generate K synthetic images directly, our DCGAN generates K synthetic images using K generators with different prior noise vectors and one shared discriminator to ensure the diversity existing among them. The intuition behind is that each reader imagines Hamlet based on his/her prior knowledge, and different prior knowledge leads to a different image of Hamlet in his/her mind. As illustrated in Fig. 1, the synthetic images generated by DCGAN not only cover most of the content in text information, but also provide much information about the background underlying in text information. This is also consistent with our human understanding of text information. Therefore, we can fully explore this kind of visual representation on text to improve the accuracy of image recognition or semantic segmentation, and automatically collect a large-scale labeled image dataset from the images with text description largely available on social media platforms.

We apply an image-level CNN to extract visual features for each synthetic image. It worth emphasizing that we care much more about the common visual representation among these K synthetic images rather than each individual. Therefore, to obtain a compact visual feature representation, we first apply an affine transformation with a ReLU layer to adjust feature maps and reduce the channels before we apply an element-wise pooling to extract the common feature

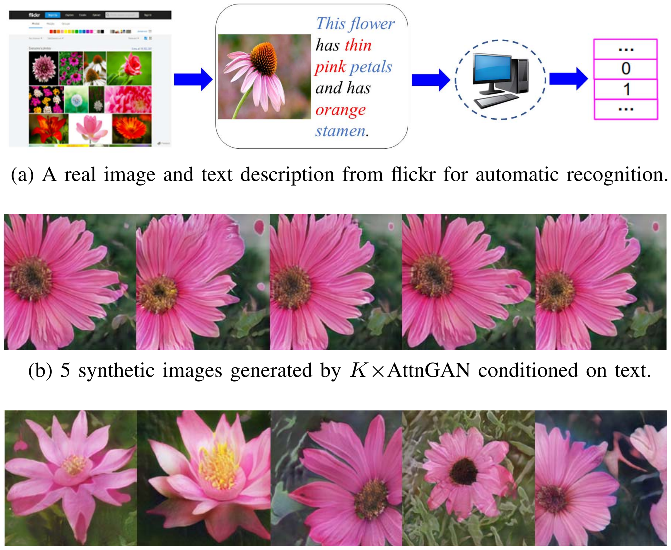
Manuscript received March 10, 2020; revised October 31, 2020 and February 10, 2021; accepted February 12, 2021. Date of publication March 5, 2021; date of current version March 11, 2021. This work was supported in part by the Key Technological Innovation Projects of Hubei Province under Grant 2018AAA062; in part by the National Natural Science Foundation of China under Grant 61562025, Grant 61972298, and Grant 61962019; and in part by the Chinese Association for Artificial Intelligence, CAAI-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lucio Marcenaro. *(Corresponding author: Chunxia Xiao.)*

Tao Hu is with the School of Information Engineering, Hubei Minzu University, Enshi 445000, China, and also with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: hutao_es@foxmail.com).

Chengjiang Long is with JD Tech at Silicon Valley R&D Center, Mountain View, CA 94043 USA (e-mail: chengjiang.long@jd.com).

Chunxia Xiao is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: cxxiao@whu.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3061927



(a) A real image and text description from flickr for automatic recognition.

(b) 5 synthetic images generated by $K\times$AttnGAN conditioned on text.

(c) 5 synthetic images generated by our DCGAN conditioned on text.

Fig. 1. Two visual representation methods on text for automatic image visual recognition of the collected a large-scale labeled image dataset from social media platform like Flickr. One is using AttnGAN [7] (b) for K times conditioned a text with K different noises, and the other one is our proposed DCGAN method (c) conditioned on text. Both these methods can generate K high-resolution and photo-realist synthetic images. Our goal is to visually represent text information and extract visual feature representations from synthetic images to boost the accuracy of automatic recognition on real images.

representation, which can be considered as a visual representation of text in the visual feature space. Take image recognition as an example, considering that feature fusion has been proved to be effective in improving the performance of image recognition, we combine the real image feature extracted from the image-level CNN, text feature extracted from a text-level CNN, and the common feature representation from K synthetic images via multi-source feature fusion, and then we feed the final feature representation into a fully connected layer as a classifier for visual recognition.

To sum up, the main contributions of our paper are three-fold as follows:

(1) We propose a novel method called DG-VRT to represent text information with K visually plausible synthetic images generated via our proposed DCGAN, which is composed of K generators associated with different noise priors and a single shared discriminator.

(2) We extract a common and compact visual feature representation from synthetic images conditioned on a text, and combine it with an image-level feature from a real image, and a text-level feature together via a multi-source feature fusion to boost the performances of image recognition.

(3) The extensive experiments on two benchmark datasets have demonstrated the efficacy of our recognition framework. We also extend DG-VRT to solve the semantic segmentation task and the experimental results strongly demonstrate the efficacy of our approach.

II. Related Works

Considering that our work is mainly to verify the effect of visual features of text on improving image recognition performance, we mainly review the related works in three fields, *i.e.*, *text information for image recognition*, *attention mechanism*, and *diverse conditional text-to-image GANs*.

A. Text Information for Visual Recognition

As text information provides informative and semantic signals for the image, it can improve the accuracy of image recognition combining image feature with text information. Huang *et al.* [13] proposed a deep multimodal attention network to embed text description and visual content, which is effective in multi-label classification. Johnson *et al.* [14] proposed a deep convolutional neural network to combine both the visual information of images and their neighboring images defined based on the shared tags from a text. Rawat *et al.* [15] proposed ConTagNet to predict multiple tags for an image based on the text content. Long *et al.* [5] extracted deep text-level features using multiple text CNNs [16], [17] for text representation in image labeling. Hu *et al.* [18] proposed a joint vision and language model for image segmentation from referring expressions, which utilizes existing large scale vision-only and text-only dataset. To solve the referring image segmentation, Chen *et al.* [19] proposed See-through-Text grouping to reveal segmentation cues of the pixel by ConvRNN. These works [18], [19] just use textual information to directly deal with image segmentation, and do not consider the hidden visual information of the text. Different from the mentioned works, we adopt diverse conditional text-to-image GANs to generate multiple photo-realistic synthetic images to extract visual feature representation for the text description. We then make full use of the extracted visual feature representation as a novel text interpretation to improve the performance of image visual recognition [20]–[25] and semantic segmentation [26].

B. Attention Mechanism

is an important part of sequence translation models, which can be described as mapping a query and a set of key-value pairs to an output [27]. It has been successfully used in modeling multi-level dependencies in machine translation [28], video understanding [29], [30], image labeling [31] and other computer vision applications [32]–[35]. The attention mechanism can enable GANs to generate fine-grained high-resolution and photo-realistic synthetic images with multi-level conditioning. Xu *et al.* [7] firstly explored the attention mechanism in GANs, which is able to compute the similarity between the synthetic images and the input text description. Qiao *et al.* [10] also used the attention mechanism to guarantee semantic consistency based on word-level attention and sentence-level attention model. Chen *et al.* [36] proposed FTGAN to train the text encoder and image decoder simultaneously based on attention mechanism for fine-grained text-to-face synthesis. Like AttnGAN [7], our proposed diverse conditional GANs explore an attention mechanism to compute the similarity between each synthetic image and the input text description. We also use the attention mechanism to fuse visual representation feature, text feature, and real image feature into a combined feature representation for image recognition.

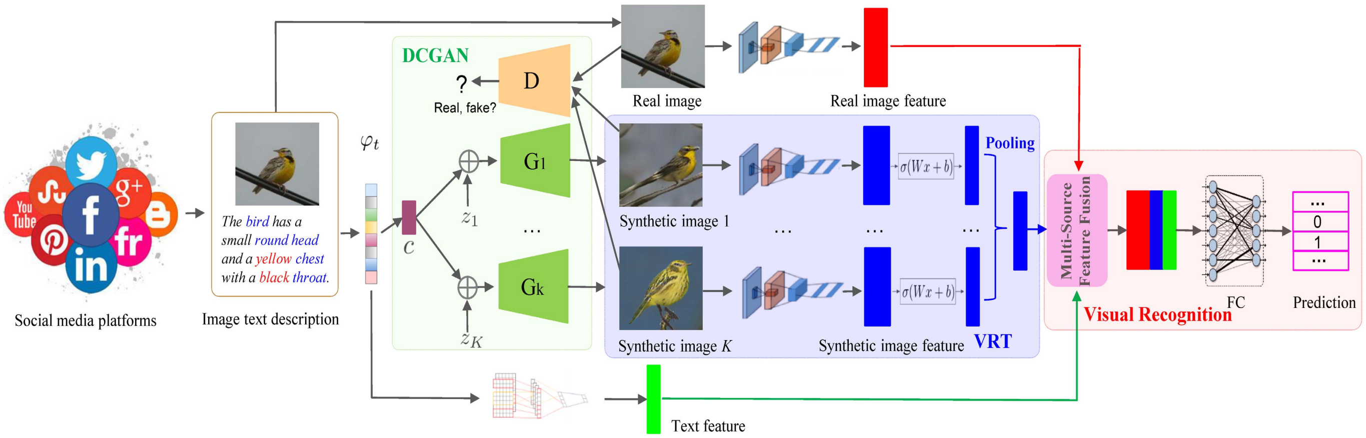


Fig. 2. Overview of our proposed framework DG-VRT for visual recognition on real images with text description. DCGAN is designed to generate K high-resolution and photo-realistic synthetic images conditioned on a text. VRT refers to visual feature representation in blue from the K synthetic images, which are extracted to represent the text in the visual feature space. Finally, a multi-source feature fusion is applied to formulate the final feature representation for predicting the category of the given real image.

C. Diverse Conditional Text-to-Image GANs

are proposed to effectively generate high-resolution images from text descriptions, which have shown promising results in text-to-image synthesis [7]–[11], [37]–[40]. Reed *et al.* [38] developed a GAN to translate visual concepts from characters to pixels and generate 128 × 128 images. Zhang *et al.* [8] proposed StackGAN to synthesize images from text description in two separate stages, and later they extended StackGAN to StackGAN++ [9] using multiple generators and discriminators with different resolution scales. Xu *et al.* [7] proposed AttnGAN to pay attention to relevant words with different sub-regions of the image. Qiao *et al.* [10] introduced MirrorGAN to guarantee semantic consistency between text and synthetic image. HDGAN [40] used hierarchical-nested adversarial objectives to regularize mid-level representations and designed an extensile single stream generator architecture to push generated images up to high resolutions. Chen *et al.* [36] proposed FTGAN for fine-grained text-to-face synthesis. TAGAN [11] used word-level local discriminators to optimize the performance of a single generator. In comparison, our proposed DCGANs use one discriminator to optimize multi-generators in one scale stage.

In this paper, we focus on extracting visual representation for a text description by generating high-resolution and photo-realistic synthetic images. Hence we implement two versions of DCGAN, denoted as "DCGAN-A" incorporating AttnGAN and "DCGAN-S" incorporating StackGAN++. Corresponding to "DCGAN-S" and "DCGAN-A", there are two versions of the proposed DG-VRT framework, denoted as "DG$_S$-VRT" and "DG$_A$-VRT". Our DCGAN-S and DCGAN-A extend one generator to K generators at each stage to generate K synthetic images with different prior noise vectors. Different from Hoang *et al.*'s MGAN [41], our DCGAN-S and DCGAN-A use multiple generators with various prior noise vectors, a single discriminator, and an attention similarity model at each scale stage to generate multiple synthetic images. Our methods can effectively represent visual concepts information embedded in the text description.

III. Proposed Method

As illustrated in Fig. 2, our proposed framework DG-VRT consists of three key components, *i.e.*, generating K synthetic images using a diverse conditional GAN (DCGAN), extracting visual feature representation from the synthetic images, and combining the synthetic image feature with real image feature and text feature using a multi-source feature fusion to conduct a classification task for image recognition. We discuss with details in the following subsections.

A. Diverse Conditional GAN

Intuitively, given a text, different people may imagine a different visual scene, and one text-based synthetic image is not sufficient to cover the underlying information behind the text itself. Therefore, to better explore the visual feature representation of the text information, we propose a DCGAN which includes K generators $\left[G_i^1, G_i^2, \ldots, G_i^K\right]$ and one shared discriminator D_i to generate K synthetic images at the i-th resolution scale stage for each text information. Note that such K generators are corresponding to K different prior noise vectors $z_1, \ldots, z_K$ (we sample values for each z_k from a normal distribution) as input. For the alternative training purpose, the loss function of discriminator D_i at the i-th resolution scale stage is designed as Equ (1), and the loss function of generators $\left[G_i^1, G_i^2, \ldots, G_i^K\right]$ is designed as Equ (2).

$$\begin{aligned}\mathcal{L}_{D_i} = {} & K\mathbb{E}_{X_i\sim p_{data_i}}\left[-\log(D_i(X_i))\right]\\ & + \sum_{k=1}^{K}\mathbb{E}_{s_i^k\sim p_{G_i^k}}\left[-\log(1-D_i(s_i^k))\right]\\ & + K\mathbb{E}_{X_i\sim p_{data_i}}\left[-\log(D_i(X_i,c))\right]\\ & + \sum_{k=1}^{K}\mathbb{E}_{s_i^k\sim p_{G_i^k}}\left[-\log(1-D_i(s_i^k,c))\right].\end{aligned} \tag{1}$$

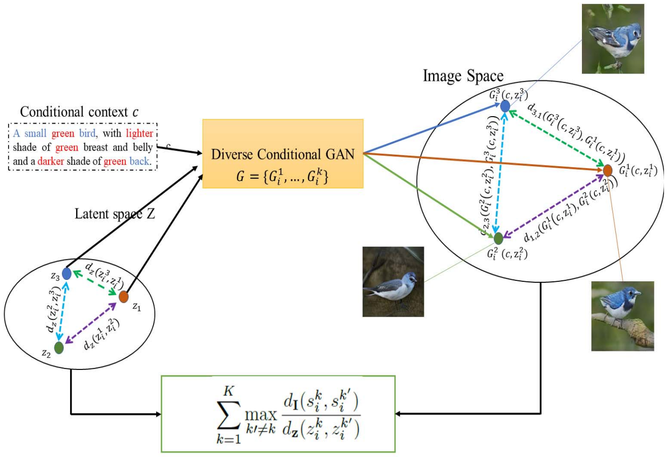

Conditional context c
A small green bird, with lighter shade of green breast and belly and a darker shade of green back.
c
Diverse Conditional GAN
$G = \{G_i^1, \ldots, G_i^k\}$
Latent space Z
z_1
z_2
z_3
$d_z(z_i^3, z_i^1)$
$d_z(z_i^2, z_i^3)$
$d_z(z_i^1, z_i^2)$
Image Space
$G_i^3(c, z_i^3)$
$G_i^1(c, z_i^1)$
$G_i^2(c, z_i^2)$
$d_{3,1}(G_i^3(c, z_i^3), G_i^1(c, z_i^1))$
$d_{2,3}(G_i^2(c, z_i^2), G_i^3(c, z_i^3))$
$d_{1,2}(G_i^1(c, z_i^1), G_i^2(c, z_i^2))$
$\sum_{k=1}^{K} \max_{k\prime \neq k} \frac{d_{\mathbf{I}}(s_i^k, s_i^{k'})}{d_{\mathbf{z}}(z_i^k, z_i^{k'})}$

Fig. 3. Diverse loss of the proposed DCGAN.

$$\begin{aligned}\mathcal{L}_{G_i} &= \sum_{k=1}^{K} \mathcal{L}_{G_i^k} + \mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K) \\ &\quad + \mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K). \qquad (2)\\ \mathcal{L}_{G_i^k} &= \mathbb{E}_{s_i^k \sim p_{G_i^k}}\left[-\log(D_i(s_i^k))\right] \\ &\quad + \mathbb{E}_{s_i^k \sim p_{G_i^k}}\left[-\log(D_i(s_i^k, c))\right]. \qquad (3)\end{aligned}$$

where $\mathbf{X}$ is from the true image distribution p_{data}, s_i^k is from the mode distribution p_{G_k}, c is text conditional parameter, $\mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K)$ is a image-text semantic matching loss which calculates the similarity between image and word-level text.

The term $\mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K)$ calculates the dot-product attention between the word or text and the sub-region of the image at the i-th scale stage. It uses the Cosine similarity to represent the relation between text and images. Different from work [7], $\mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K)$ adds up the attention between the text and the corresponding K synthetic and it is defined as:

$$\mathcal{L}_{\text{similarity}}(s_i^1, \ldots, s_i^K) = \sum\nolimits_{k=1}^{K}(\mathcal{L}_{1,k}^{w} + \mathcal{L}_{2,k}^{w} + \mathcal{L}_{1,k}^{s} + \mathcal{L}_{2,k}^{s}). \quad (4)$$

where $\mathcal{L}_{1,k}^{w} = -\sum_{i=1}^{M} log P(c_i|s_i^k)$ is the negative log posterior probability that the images are matched with their corresponding word-level description for the k-th generator, and $log P(c_i|s_i^k)$ is the posterior probability the sentence c_i matching with its corresponding image s_i^k. $\mathcal{L}_{2,k}^{w} = -\sum_{i=1}^{M} log P(s_i^k|c_i)$ is the negative log posterior probability that the word-level description matches with its corresponding image. Similarly, the $\mathcal{L}_{1,k}^{s}$ and $\mathcal{L}_{2,k}^{s}$ are the negative log posterior probabilities between image and text-level description for the k-th generator.

The term $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$ measures the diversity of the K synthetic images at the i-th scale stage. Inspired by [39], we can introduce a ratio, which calculates the difference between the noise vector and the corresponding feature difference of synthetic images to optimize the generator. Considering that our DCGAN has K generators and one shared discriminator, the diverse loss of the proposed DCGAN is described as Fig 3. We first calculate the feature difference $d_{\mathbf{I}}(s_i^k, s_i^{k'})$ between each pair of synthetic images in image space and the difference $d_{\mathbf{z}}(z_i^k, z_i^{k'})$ between each pair of corresponding noise vectors in latent space. Then, we set the cumulative value of the ratios between $d_{\mathbf{I}}(s_i^k, s_i^{k'})$ and $d_{\mathbf{z}}(z_i^k, z_i^{k'})$ as the diverse loss $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$ of the K generators. The $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$ is defined as:

$$\mathcal{L}_{diverse}((s_i^1, \ldots, s_i^K) = \sum_{k=1}^{K} \max_{k\prime \neq k} \frac{d_{\mathbf{I}}(s_i^k, s_i^{k'})}{d_{\mathbf{z}}(z_i^k, z_i^{k'})}. \qquad (5)$$

where $d_{\mathbf{I}}$ is the distance between synthetic image features, and $d_{\mathbf{Z}}$ means the distance between corresponding noise vector. Then the discriminator D_i and generators $\left[G_i^1, G_i^2, \ldots, G_i^K\right]$ can be optimized in a joint form by alternatively maximizing $\mathcal{L}_{D_i}$ and minimizing $\mathcal{L}_{G_i}$ until convergence. It is worth noting that $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$ uses K pairs of noise and image feature differences. The work [39] only uses one pair of noise and image feature difference. Therefore, $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$ can better utilize the relationship between the noises in the latent space and the image features in the image space to improve the diversity of the synthetic images.

For the theoretical analysis of our DCGAN, its generators are trained to combine K different prior noise vectors, as well as the text embedding vector to interpolate K different synthetic images, while its discriminator has been trained to predict whether the synthetic images and the text match or not. Therefore, the images from interpolated text embeddings can fill in the gaps in the data manifold that were presented during training. With the diverse term in Equation 5 included, we are able to ensure the diversity of the synthetic images. Such K diverse synthetic images are generated based on the correlation between text embeddings and the corresponding real images. That's why the visual feature representation extracted from the synthetic images can help boost the performance of image recognition on the real images.

It is worth mentioning that all the $\left[G_i^1, G_i^2, \ldots, G_i^K\right]$ share the same architecture. Therefore, in principle, any text-to-image GAN can be incorporated into our DCGAN framework. We prefer to incorporate those text-to-image GANs like StackGAN++ [9] and AttnGAN [7].

Discussion: Our DCGAN incorporates a text-to-image GAN to generate the diverse synthetic images. Note that the K generators in DCGAN can share same weights. In this way, we can control the training cost compared with training K single text-to-image GAN separately. We observe that training a DCGAN with shared weights is less expensive than training K single text-to-image GANs separately, and the K synthetic images generated by DCGAN are more diverse. As illustrated in Fig. 4, our DCGAN generates more diverse synthetic images, either with StackGAN++ or with AttnGAN.

To better understand the efficiency of attention mechanism in DCGAN-A, we visualize $K = 5$ synthetic images conditional on a simple text in Fig. 5. We observe that some sub-regions of synthetic images can be inferred from word-context features and small scale synthetic image feature by DMASM [7]. DCGAN-A allocates attention to all words and projects the attention to those sub-regions, as shown in Fig. 5. On the Caltech-UCSD Birds-200-2011 Dataset,

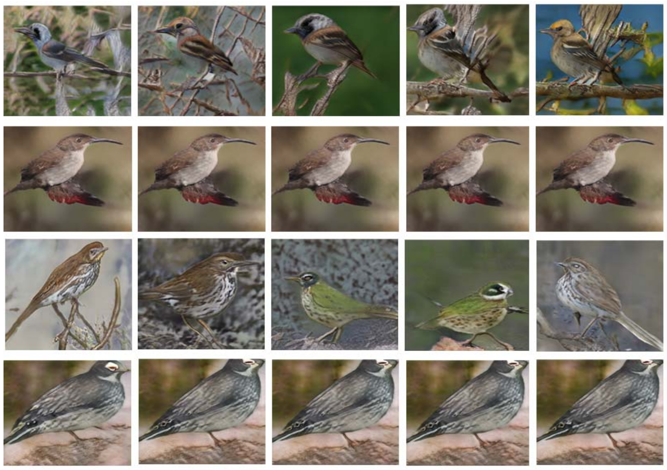


Fig. 4. $k = 5$ synthetic images generated conditioned on the text "A bird with a tiny pointed bill, large eyes with white eyering, small head and white breast." From top to bottom are generated by DCGAN with StackGAN++, $K\times$StackGAN++, DCGAN with AttnGAN, and $K\times$AttnGAN, respectively.

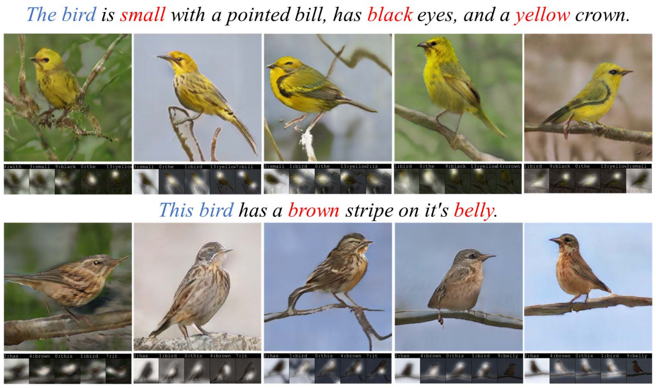

The bird is small with a pointed bill, has black eyes, and a yellow crown.
This bird has a brown stripe on it's belly.

Fig. 5. Visualization of $K = 5$ 256×256 synthetic images generated by our DCGAN-A on the text. Below each synthetic image are the visualization of top-5 most attended words by DMASM [7] in DCGAN-A.

the words like "birds", "the" and "this" are usually attended by the similarity model for locating the bird. The words like "black", "yellow", "brown", "small" and "belly", which describe the attributes (like color or shape) of birds, are also attended to draw the detail of birds. Obviously, each synthetic image effectively reflects the attended words and looks different from others. Such a visual diversity among synthetic images demonstrates that our DCGAN-A is able to understand the detailed semantic information from a text, and generate diverse and informative photo-realistic synthetic images for visually interpreting the text.

B. Visual Representation on Text

Given a text t, we can generate K synthetic images $\mathbf{s} = \{s^1, \ldots, s^K\}$ with our DCGAN. We use a pretrained ResNet [42] model as feature extractor, which is denoted as "$\phi(\cdot)$". We feed the K generated synthetic images into $\phi(\cdot)$ to extract visual feature at the second last layer (denoted as "$\phi(s^k)$") with size of $7 \times 7 \times 2048$ for each synthetic image s^k.

The early fusion often cannot make full use of the complementarity between K synthetic images, and the feature extracted from K synthetic images usually contains a lot of redundant information. Due to the errors from late fusion are often uncorrelated and do not affect each other, we use late fusion to process all K visual features in the visual representation on text. In order to fuse visual features for these K synthetic image, we compute an h-dimensional hidden state for each image by applying an affine transformation and an element-wise ReLU nonlinearity $\sigma(\varepsilon) = \max(0, \varepsilon)$ to its feature. To treat hidden states for each synthetic image differently, we apply distinct transformations to $\phi(s^k)$ with parameters $W_k \in \mathbb{R}^{d\times h}$ and $b_k \in \mathbb{R}^h$, and then we arrives at hidden states $\mathbf{v}_{s^k} \in \mathbb{R}^h$ for $s^k \in \mathbf{s}$. To generate a single hidden state $\mathbf{v}_s \in \mathbb{R}^h$ for all the synthetic images $\mathbf{s}$, we apply an element-wise max-pooling on each $\mathbf{v}_{s^k}$ so that $\mathbf{v}_s = \max_k \mathbf{v}_{s^k}$, *i.e.*,

$$\mathbf{v}_s = \max_k \mathbf{v}_{s^k} = \max_k(\sigma(\mathbf{W}_k \phi(s^k) + \mathbf{b}_k)). \tag{6}$$

where $\mathbf{W}_k$ and $\mathbf{b}_k$ is the parameters of distinct transformations.

Discussion: we choose synthetic images rather than visual features because we want to visually map the text using high quality synthetic images so that the human can view directly, while visual features may miss some details.

C. Multi-Feature Fusion for Visual Recognition

Regarding text information, standard deep learning model for text classification and sentiment analysis usually uses a word embedding layer and one-dimensional convolutional neural network [43]. The model can be expanded by using multiple parallel convolutional neural networks that read the source document using different kernel sizes. This, in effect, creates a multichannel convolutional neural network for text that reads text using different n-gram sizes (groups of words). We follow Kim's multi-channel model to implement a merged model with 3 text CNNs with kernels of different sizes (4-gram, 6-gram, and 8-gram [44]), denoted as 3-Text-CNNs, to extract 512-dimensional text-level feature $\mathbf{v}_t$ from the second last layer.

Considering that the image feature and text feature are multi-modal features, they have different feature dimensions. It is not suitable to directly connect image feature and text feature. However, we can calculate the similarity between image features and text features to connect them. In this paper, we combine the visual feature $\mathbf{v}_s$ from all the K synthetic images $\mathbf{s}$ and the text-level feature $\mathbf{v}_t$ of input text as $Attn(\mathbf{v}_s, \mathbf{v}_t)$ by an attention model. The attention model is used to concatenate the visual feature $\mathbf{v}_r$ of real image with $Attn(\mathbf{v}_s, \mathbf{v}_t)$ to be the final feature representation $Attn(\mathbf{v}_r, Attn(\mathbf{v}_s, \mathbf{v}_t))$. Finally, we feed $\mathbf{x}' = Attn(\mathbf{v}_r, Attn(\mathbf{v}_s, \mathbf{v}_t))$ into two connected layers to conduct a classification task for image recognition. The whole process of multi-features fusion for image recognition is described in Algorithm 1.

There are three common attention models, *i.e.*, additive attention, dot-product attention and multiplicative attention [27]. In this paper, we adopt additive attention to formulate the multi-source feature fusion. For example, we use the Eqn (7) to calculate the similarity between a vector $\mathbf{p}$ and the

Algorithm 1 The Training Procedure Using Synthetic Images Generated by DCGAN for Visual Recognition

```
Input: Real image $\mathbf{x}$, synthetic images $\mathbf{s} = \{s^1, \ldots, s^K\}$,
    input text $t$, the max training epoch $N_e$, the learning rate
    of optimizer $\alpha$, the decay rate of learning rate $\beta$.
Output: The network $f(\Theta)$ for real image $\mathbf{x}$ category.
Extract the visual feature $\mathbf{v}_r$ of real image $\mathbf{x}$ by ResNet.
Extract the textual feature $\mathbf{v}_t$ of input text $t$ by 3-Text-CNNs.
for $k \in [1, K]$ do
    Extract visual feature $\phi(s^k)$ of $s^k$ by ResNet.
end for
Initialize the network $f$ parameters $\Theta$.
for $i \in [1, N_e]$ do
    Calculate $\mathbf{v}_s$ based on Eqn 6.
    $Attn(\mathbf{v}_s, \mathbf{v}_t) = \mathbf{Sim}_{add}(\mathbf{v}_s, \mathbf{v}_t) \oplus \mathbf{v}_s$.
    $\mathbf{x}' = \mathbf{Sim}_{add}(\mathbf{v}_r, Attn(\mathbf{v}_s, \mathbf{v}_t)) \oplus \mathbf{v}_r$.
    $\hat{y} = \mathbf{argmax}(FC(\mathbf{x}'))$.
    $J(\Theta) = -\sum_j \sum_n y_j^n \log(\hat{y}_n^j)$.
    $\Theta_{i+1} = \mathbf{Adam}(f, \Theta_i, \alpha, \beta)$
end for
return $\Theta_i$
```

TABLE I

Image Recognition Comparisons With Different Attention Methods, *i.e.*, Additive Attention (ADDAttn), Dot-Product Attention (DPAttn), and Multiplicative Attention (MPAttn), Used in DG$_S$-RT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ on the Caltech-UCSD Birds-200-2011 Dataset. (Unit: %)

Drop rate	Fusion	ADDAttn	DPAttn	MPAttn
0.5	90.76	93.44	90.92	92.06
0.2		93.19	90.92	92.30
0.1		**93.92**	91.90	90.41
None		90.06	92.79	91.41
Accuracy	90.76	**92.65**	91.63	91.55

query vector $\mathbf{q}$. The V, W_k, W_q are the initialization weights, $W_k \in \mathbb{R}^{d_m \times d_k}$, $W_q \in \mathbb{R}^{d_m \times d_k}$, $V \in \mathbb{R}^{d_m \times d_v}$. d_m is the output of the “And" operation of two key vectors d_k and d_v of $\mathbf{q}$.

$$\mathbf{Sim}_{add}(p, q) = \tanh(W_k \cdot \mathbf{p} + W_q \cdot \mathbf{q}) \cdot V. \tag{7}$$

Discussion: To verify the effectiveness of additive attention model and explain why we will use additive attention model in proposed framework, we compare it with another two attention models on the Caltech-UCSD Birds-200-2011 dataset [45]. We first use the DCGAN-S to generate $K = 2$ synthetic images. Then we fuse $\mathbf{v}_t$, $\mathbf{v}_s$ and $\mathbf{v}_r$ directly to predict the category of real image. We also set the parameter dropout rate with different values to adjust label prediction accuracy of real image with three attention models.

The statistic information of recognition accuracy is summarized in Table I, from which we can observe: (1) multi-source feature fusion with attention can effectively improve performance of image recognition. (2) When the dropout rate is 0.1, we get the best recognition accuracy with the additive attention, and the average recognition accuracy is 92.65%, which is the highest score between the three attention methods. In the following experiments, we use additive attention model to fuse multi-source features and the dropout rate is set as 0.1.

D. Implementation Details

The training parameters of our DG$_A$-VRT involves the parameters of text encoder and image encoder of an attentional similarity model, the parameters of K generators and one discriminator in DCGAN-A, the parameters in ResNet-50s and 3-text CNNs, the affine transformation parameter $\mathbf{W}_k$ and $\mathbf{b}_k$, the parameters of attention similarity between real image feature and synthetic images-text wise, and the parameters in the last fully connected layer. The DG$_A$-VRT training procedure is divided into three phases.

At Phase I, we use pairs of a real image and its corresponding text to train text decoder and image decoder based on DMASM [7]. At Phase II, based on the trained text decoder and image decoder we use the pairs of a real image and its corresponding text again to train DCGAN-A in an alternative optimization procedure until convergence. At Phase III, we apply the trained DCGAN-A to generate K high-resolution, word-related and photo-realistic synthetic images. Then we feed n real images, nK synthetic images and n text to learn the rest parameters in the entire framework with attention mechanism and softmax cross-entropy loss function. Similarly, the DG$_S$-VRT takes almost the same training except the procedure of DMASM part.

Note that we follow the tricks in StackGAN-v2 [9] to train DCGAN with StackGAN++ at each stage at Phase II with a batch size of 12 for 350000 iterations. We also follow the tricks in AttnGAN to train DCGAN with AttnGAN [7] at Phase I with a batch size of 90 for 76,800 iterations, and at each stage at Phase II with a batch size of 40 for 105,000 iterations. At Phase III, with a minibatch size of 64, we initialize all parameters with pre-trained models (ResNet-50 and 3 text CNNs), attention similarity between real image feature and synthetic images-text wise. We set the learning rate $\alpha = 0.0001$ and the decay rate $\beta = 0.00001$ in the training process of image recognition. We use stochastic gradient descent with a dynamic learning rate which is reduced by 0.1 when the number of val_loss decreasing is greater than 4. We use Adam as the optimization.

IV. Experiments

Our experiments are conducted on two datasets, *i.e.*, the Oxford 102 Category Flower Dataset [46], and the Caltech-UCSD Birds-200-2011 Dataset [45]. We use accuracy as the metric to measure performance of image recognition.

A. Experiments on Oxford 102 Category Flower Dataset

The Oxford 102 Category Flower Dataset [46] consists of 8,189 images with 102 categories of flowers which commonly occurs in the United Kingdom, and each category has 40 to 258 images. The images have large scale, pose and light variations. The text context information is provided by [47] with 10 descriptions for each image. We train our DCGAN on the texts and their corresponding real images. With the learned DCGAN, we are able to generate K visual plausible synthetic images conditioned on a text for experiments.

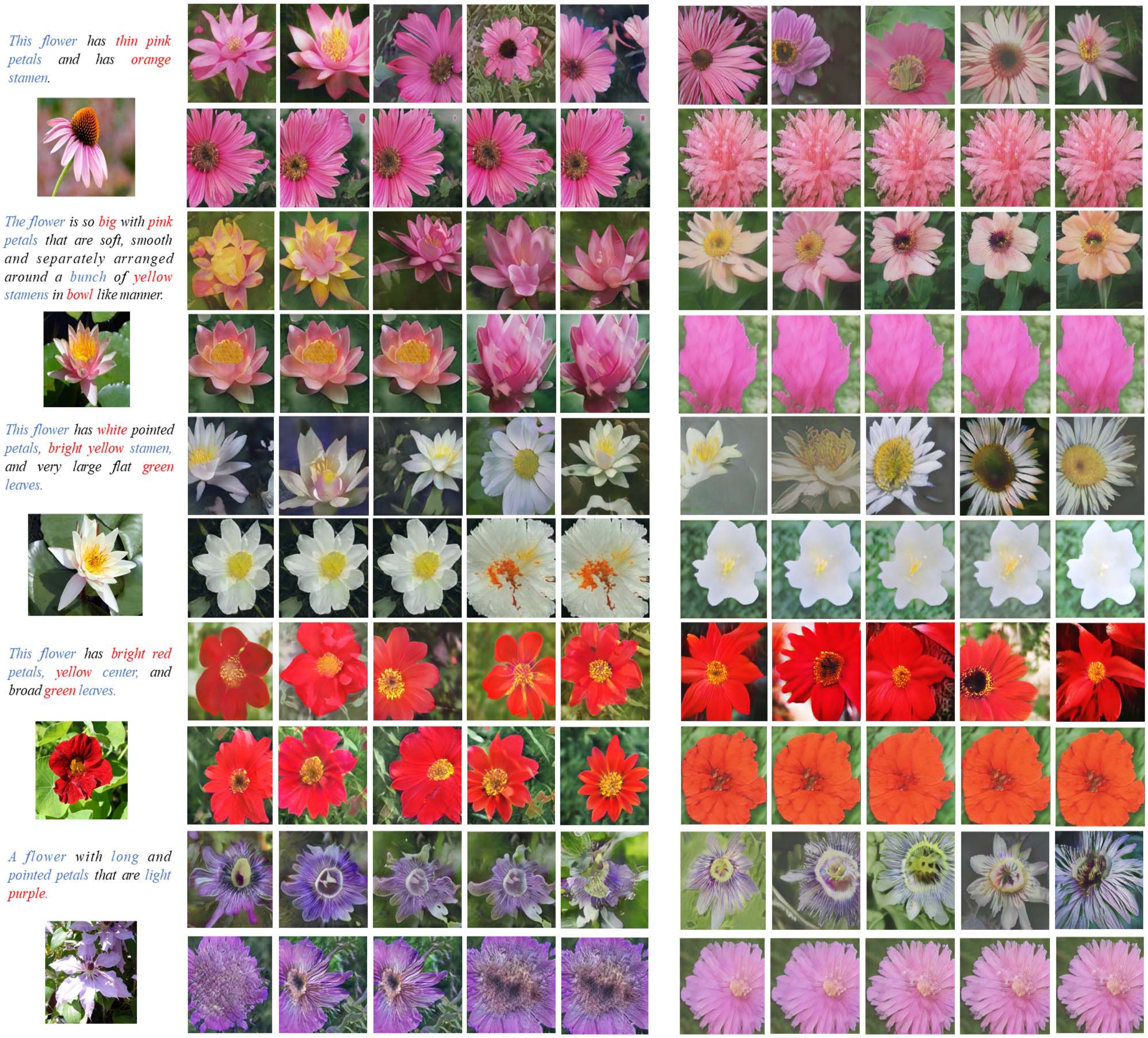

This flower has thin pink petals and has orange stamen.
The flower is so big with pink petals that are soft, smooth and separately arranged around a bunch of yellow stamens in bowl like manner.
This flower has white pointed petals, bright yellow stamen, and very large flat green leaves.
This flower has bright red petals, yellow center, and broad green leaves.
A flower with long and pointed petals that are light purple.

Fig. 6. Visualization of $K = 5$ high-resolution and photo-realistic synthetic images conditioned on a text, and compared with the corresponding real images (left region) on the Oxford 102 Category Flower Dataset. The synthetic images generated by DCGAN-A are located in the top of the central region, and the synthetic images generated by AttnGAN are located in the below of the central region. The synthetic images generated by DCGAN-S are located in the top of the right region, and the synthetic images generated by StackGAN++ are located in the below of the right region.

1) Effectiveness of DCGAN: To verify the effectiveness of our DCGAN, we conduct experiments to check the diversity of the synthetic images, and visual concept consistency between real images and the generated synthetic images. It's worth explaining that if DCGAN-S removes $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$, then DCGAN-S will degenerate into $K\times$StackGAN++. DCGAN-A removes $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K)$, then DCGAN-A will degenerate into $K\times$AttnGAN. If AttnGAN removes $\mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K)$, it will degenerate into StackGAN++. So we can describe the DCGAN-S as $\mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K) + K \times StackGAN{+}{+}$, and DCGAN-A as $\mathcal{L}_{similarity}(s_i^1, \ldots, s_i^K) + \mathcal{L}_{diverse}(s_i^1, \ldots, s_i^K) + K \times StackGAN + +$. This subsection can be considered as ablation study of DCGAN.

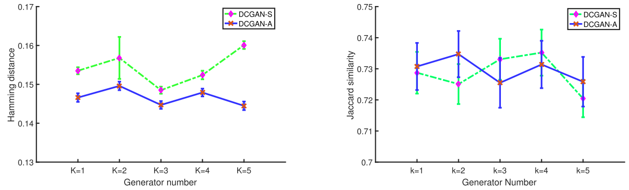
We use the metrics FID [48] and LPIPS [49] to measure the diversity on DCGAN-A and DCGAN-S. Note that lower FID values and higher LPIPS values indicate more diversity existing among synthetic images. The statistic is summarized in Table II, from which we can see the synthetic images generated by DCGAN-S are more diverse than by StackGAN++, and the synthetic images generated by DCGAN-A are most diverse.

We visualize some synthetic images by DCGAN-S and DCGAN-A in Fig. 6. As we can see, our generated K synthetic images not only cover main content elements in the text, but also provide underlying background and other rich visual information like size, shape and pose variations which are not mentioned in the text, due to various prior noise

TABLE II

DIVERSE PERFORMANCE COMPARISON ON THE OXFORD 102 CATEGORY FLOWER DATASET. THE SMALLER FID VALUES, THE BETTER QUALITY. THE LARGER LPIPS VALUES, THE MORE DIVERSE

Methods	FID	LPIPS
$K\times$StackGAN++	64.13±0.8832	0.2347±0.0163
$K\times$AttnGAN	42.41±0.1902	0.3304±0.0084
DCGAN-S	45.03±1.0726	0.3550±0.0013
DCGAN-A	33.11±0.1134	0.3431±0.0018



(a) Hamming distance.

(b) Jaccard similarity.

Fig. 7. The correlation measured using Hamming distance and Jaccard similarity between the synthetic images generated by our DCGAN-S and DCGAN-A conditioned on a text and the corresponding real images.

vector z_k. These observations are consistent with human's behavior to represent a text based on his/her prior knowledge. In other words, even given the same text, different people with different backgrounds will imagine different visual pictures in their brains. Such diverse representations are complementary to each other and can be merged to formulate a more representative format.

We also measure the correlation between synthetic images conditioned on a text with the corresponding real images with two distance metrics, *i.e.*, Hamming distance and Jaccard similarity, between their visual feature vectors extracted by the Pooling layer of VGG16 model [50]. Note that lower Hamming distance and higher Jaccard similarity indicate more correlative to each other in the given feature space. For Hamming distance in the range $[0, 1]$, smaller value means higher similarity between images. For Jaccard similarity in the range $[0, 1]$, the value closer to 1.0 means two compared images are more correlative to each other in the given feature space.

We plot both Hamming distance value and Jaccard similarity value between our generated synthetic images and the corresponding real images in Fig. 7. The hamming distance of DCGAN-S is 0.1543 ± 0.0039 and the Jaccard similarity of DCGAN-S is 0.7285 ± 0.0053; the hamming distance of DCGAN-S is 0.1467 ± 0.0019 and the Jaccard similarity of DCGAN-A is 0.7296 ± 0.0035. As we can observe, none of Hamming distances is larger than 0.16 and all Jaccard similarity values are over 0.70, which indicates high correlation between each synthetic image and the corresponding real image in the visual feature space.

2) Performance Comparison: We compare our proposed DG-VRT with Johnson *et al.*'s Convolutional Neural Network with Nearest Neighborhood [6], denoted as "JCNN-NN," which explores the related neighboring images to represent image metadata especially including tags from text. To our best acknowledge, JCNN-NN is the most closely related work to our VRT because it can be represented as a visual explanation of image metadata with the related neighboring images. To make the comparison fair, we utilize the same ResNet [42] as the visual feature extractor in JCNN-NN. In addition, we add two simple baseline algorithms, *i.e.*, ResNet [42] and 3-Text-CNNs [44], which indicate using $\mathbf{v}_r$ only and using $\mathbf{v}_t$ only, respectively. We also compare our method with some fine-grained methods, such as pairwise confusion [51], which is denoted as "PC". Furthermore, we compare with the nearest neighbour radial basis function [52] (denoted as "RBF").

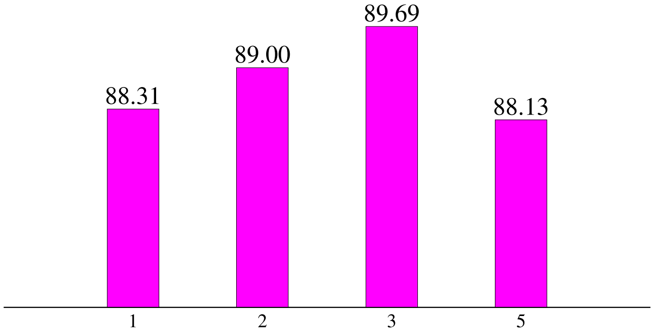
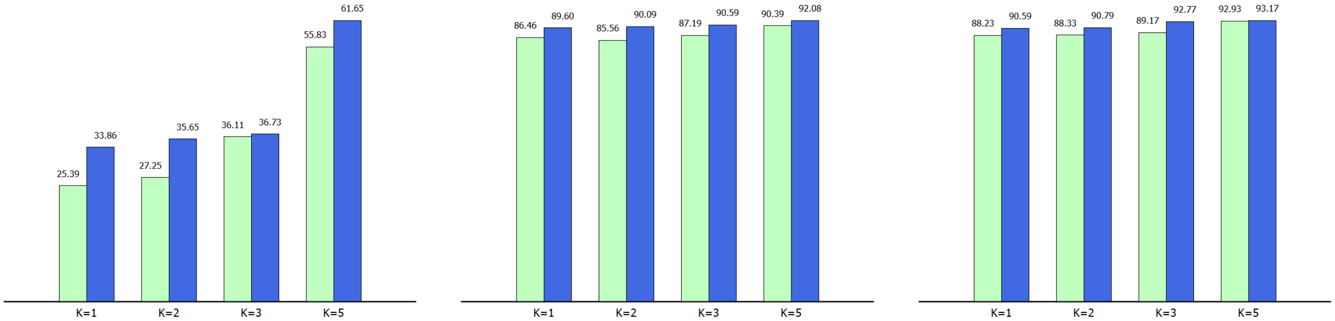


Fig. 8. Performance with accuracy (unit: %) for JCNN-NN with different K nearest neighboring images on the Oxford 102 Category Flower Dataset.

To clarify, we do not compare our DG-VRT to their variants of other text-to-image GANs instead because our focus is how to represent text for image recognition by extending and applying the existing text-to-image GANs. Note that we also run JCNN-NN with five different number of neighbor images on the Oxford 102 Dataset, as shown in Fig. 8. The accuracy of JCNN-NN is 88.13% with $K = 5$ nearest neighboring images, which is lower than $K = 3$ (89.69%). We find $K = 3$ is the best choice for JCNN-NN. Thus we just need to compare our DG-VRT with JCNN-NN ($K = 3$). We also replace DCGAN-A with $K\times$AttnGAN and DCGAN-S with $K\times$StackGAN++ in DG-VRT and get two variants denoted as KG$_A$-VRT and KG$_S$-VRT, respectively.

3) Effectiveness of Visual Feature Representation on K Synthetic Images: As stated in Section III-C, we use the feature combinations $\mathbf{v}_r + \mathbf{v}_s + \mathbf{v}_t$ where $\mathbf{v}_r$ and $\mathbf{v}_s$ indicate the visual feature representation extracted by ResNet [42] on real image, and synthetic images, respectively, and $\mathbf{v}_t$ represents the text-level feature extracted from 3-Text-CNNs [44]. We then develop two different baseline algorithms with two different feature combinations, *i.e.*, $\mathbf{v}_s$ only and $\mathbf{v}_r + \mathbf{v}_s$. For notation simplification, we denote our proposed method to be DG-VRT→ $\mathbf{v}_s+\mathbf{v}_r+\mathbf{v}_t$, which represents the feature combinations $\mathbf{v}_r+\mathbf{v}_s+\mathbf{v}_t$. We further denote the first to second baseline algorithm to be DG-VRT→ $\mathbf{v}_s$ and DG-VRT→ $\mathbf{v}_s + \mathbf{v}_r$, respectively.

We conduct a group of experiments by setting K to be various values, *i.e.*, 1, 2, 3 and 5. The results are summarized in the Fig. 9. From the Fig. 9, we find that the performance of our DG$_S$-VRT (88.23%) is similar to JCNN-NN (88.31%) with $K = 1$. The performance of our DG$_A$-VRT is 90.59% with $K = 1$, which is better than JCNN-NN. Apparently, for any all three algorithms DG-VRT→ $\mathbf{v}_s$, DG-VRT→ $\mathbf{v}_s+\mathbf{v}_r$, and DG-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$, the performance accuracy goes up when the value of K increases. This indicates that multiple synthetic



(a) DG-VRT→ $\mathbf{v}_s$ (b) DG-VRT→ $\mathbf{v}_s + \mathbf{v}_r$ (c) DG-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$

Fig. 9. Performance with accuracy (unit: %) for our VRT→ $\mathbf{v}_s$, VRT→ $\mathbf{v}_s + \mathbf{v}_r$, and VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ using visual feature representation on K synthetic images with different values of K by DG$_S$-VRT (green) and DG$_A$-VRT (blue), *i.e.*, $K = 1, 2, 3,$ and 5 on the 102 Category Flower Dataset.

TABLE III

VISUAL RECOGNITION PERFORMANCE COMPARISON ON THE OXFORD 102 CATEGORY FLOWER DATASET (UNIT: %)

Methods	Accuracy
JCNN-NN [6]	89.16±0.57
PC [51]	91.39±1.32
RBF [52]	86.26±0.82
3-Text-CNNs [44]→ $\mathbf{v}_t$	37.63±0.76
ResNet [42]→ $\mathbf{v}_r$	85.86±1.85
ResNet [42]→ $\mathbf{v}_r + \mathbf{v}_t$	88.39±0.44
KG$_S$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	88.65 ±0.45
KG$_A$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	92.49 ±0.15
DG$_S$-VRT w/o Attn→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	90.56±0.37
DG$_A$-VRT w/o Attn→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	92.72±0.26
DG$_S$-VRT→ $\mathbf{v}_s$	54.27±1.17
DG$_S$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r$	89.68±0.61
DG$_S$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	92.49±0.51
DG$_A$-VRT→ $\mathbf{v}_s$	60.49±0.99
DG$_A$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r$	91.90±0.33
DG$_A$-VRT→ $\mathbf{v}_s$ + $\mathbf{v}_r + \mathbf{v}_t$	**93.02±0.47**

images generated by DCGAN-S and DCGAN-A are complementary to be used for extracting visual concepts embedded in text and boosting the accuracy for image recognition.

We conduct the experiments repeatedly for 10 times with 10 different random data split to form the training and testing set and the results with all six algorithms are summarized in Table III, from which we can observe:

(1) 3-Text-CNNs→ $\mathbf{v}_t$ performs much worse than ResNet→ $\mathbf{v}_r$ by a half and this conveys a clue that text on the dataset is a little weaker when compared with image content.

(2) With single feature representation, our DG$_S$-VRT→ $\mathbf{v}_s$ works much better than 3-Text-CNNs→ $\mathbf{v}_t$ and is close to ResNet→ $\mathbf{v}_r$, which indicates that visual feature extracted on synthetic images is more representative than text-feature.

(3) Combining with real image feature, DG$_S$-VRT→ $\mathbf{v}_s + \mathbf{v}_r$ and DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r$ are able to improve the performance from ResNet→ $\mathbf{v}_r$, and works better than JCNN-NN and $\mathbf{v}_r + \mathbf{v}_t$, (*i.e.*, combination of features without involving DG-VRT), which shows that our visual explanation on text is more effective and robust than using a set of neighboring images defined based on the Jaccard similarity between image metadata especially tags extracted from text.

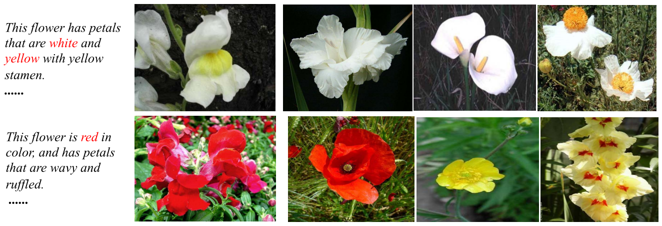

This flower has petals that are white and yellow with yellow stamen.
......
This flower is red in color, and has petals that are wavy and ruffled.
......

Fig. 10. Visualization of JCNN-NN [6]'s $K = 3$ nearest neighboring (NN) images based on the Jaccard similarity of tags extracted from text. From left to right are: text, real image, and 3 nearest-neighbour images, respectively.

(4) Combining with both real image feature and text feature, DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ performs the best.

(5) DG$_S$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ and DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ perform better than using KG$_S$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ and KG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$, which again verify the effectiveness of our DCGAN.

(6) Multi-source feature fusion with attention can effectively improve recognition accuracy. Apparently, the visual representation in our DG-VRT is robust and the visual synthetic image feature is complementary to both visual real image feature and text feature.

To further explain why our proposed DG-VRT works better than JCNN-NN, we conduct analysis on the quality of the nearest neighboring (NN) images used in JCNN-NN. As shown in Fig. 10, the color of neighbor images are not always consistent with real images, and the background of all the neighboring images are more complicated compared with the generated synthetic images in Fig. 6. Moreover, the Jarccard similarity between the query image and the k-th NNs always decreases when the value of k increases.

B. Experiments on Caltech-UCSD Birds-200-2011 Dataset

The Caltech-UCSD Birds-200-2011 Dataset consists of 11,169 bird images from 200 categories and each category has 60 images averagely. We randomly select 9,935 images for training, and use the resting 1,234 images for testing. The dataset is very challenging because it contains images with multiple objects and various backgrounds.

We train our DCGAN-S and DCGAN-A with $K = 5$ and use it to generate synthetic images on the text for each real image to conduct the experimental evaluation. Table IV provides statistics about the diversity of the generated synthetic

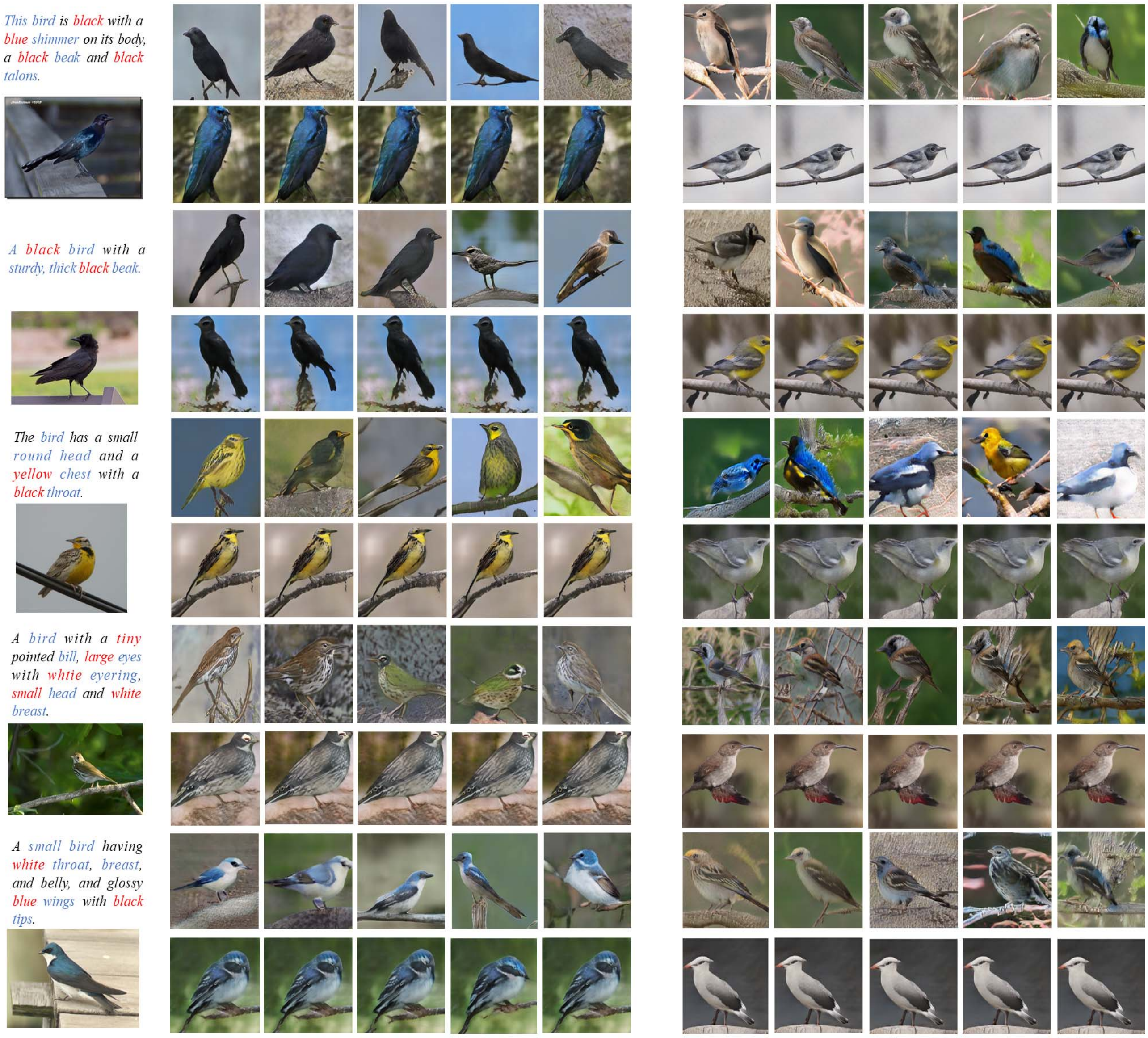

This bird is black with a blue shimmer on its body, a black beak and black talons.
A black bird with a sturdy, thick black beak.
The bird has a small round head and a yellow chest with a black throat.
A bird with a tiny pointed bill, large eyes with whtie eyering, small head and white breast.
A small bird having white throat, breast, and belly, and glossy blue wings with black tips.

Fig. 11. Visualization of $K = 5$ high-resolution and photo-realistic synthetic images conditioned on a text, and compared with the corresponding real images (left region) on the Caltech-UCSD Birds-200-2011 Dataset. The synthetic images generated by DCGAN-A are located in the top of the central region, and the synthetic images generated by AttnGAN are located in the below of the central region. The synthetic images generated by DCGAN-S are located in the top of the central region, and the synthetic images generated by StackGAN++ are located in the below of the central region.

TABLE IV

DIVERSE PERFORMANCE COMPARISON ON THE CALTECH-UCSD BIRDS-200-2011 DATASET

Methods	FID	LPIPS
$K\times$StackGAN++	27.90±0.0281	0.3137 ±0.0317
$K\times$AttnGAN	23.81±0.5265	0.3526±0.0005
DCGAN-S	26.41±0.4812	0.3712 ±0.0197
DCGAN-A	22.61±0.1295	0.3867±0.0049

images and Fig. 11 visualizes a group of synthetic image examples generated by DCGAN-S and DCGAN-A with the corresponding StackGAN++ and AttnGAN.

We repeat the experiments with 10 different random training/testing data splitting and summarize the results in Table V, which can be observed from four aspects.

(1) ResNet→ $\mathbf{v}_r$ performs much better than 3-Text-CNNs→ $\mathbf{v}_t$, which indicates image is more representative than text content.

(2) Using synthetic images only, our DG$_S$-VRT→ $\mathbf{v}_s + \mathbf{v}_r$ and DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r$ still outperform 3-Text-CNNs→ $\mathbf{v}_t$.

(3) DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$ performs a better than DG$_A$-VRT→ $\mathbf{v}_s + \mathbf{v}_r$, which works much better than DG$_A$-VRT→ $\mathbf{v}_s$. There is a similar tend on DG$_S$-VRT. Again, this observation confirms the complementary relationship between three kinds of feature representations.

(4) With either DG$_S$-VRT or DG$_A$-VRT, our DG-VRT→ $\mathbf{v}_s$+$\mathbf{v}_r$, DG-VRT→ $\mathbf{v}_s$+$\mathbf{v}_r$+$\mathbf{v}_t$ performs better than JCNN-NN and $\mathbf{v}_r$+$\mathbf{v}_t$, which suggests that our proposed DG-VRT is good at visual representation on text by extracting visual concepts from the text for boosting the recognition accuracy.

TABLE V

VISUAL RECOGNITION PERFORMANCE COMPARISON ON THE CALTECH-UCSD BIRDS-200-2011 DATASET (UNIT: %)

Methods	Accuracy
JCNN-NN [6]	89.91±0.48
PC [51]	86.87±2.66
RBF [52]	78.98±0.64
3-Text-CNNs [44]$\rightarrow \mathbf{v}_t$	5.86±0.69
ResNet [42]$\rightarrow \mathbf{v}_r$	86.81±1.14
ResNet [42]$\rightarrow \mathbf{v}_r$+$\mathbf{v}_t$	89.38±1.17
KG$_S$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	91.36 ±0.12
KG$_A$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	92.93 ±0.63
DG$_S$-VRT w/o Attn$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	91.25±0.83
DG$_A$-VRT w/o Attn$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	93.36±0.54
DG$_S$-VRT$\rightarrow \mathbf{v}_s$	55.05±0.20
DG$_S$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r$	93.28±1.25
DG$_S$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	94.49± 0.90
DG$_A$-VRT$\rightarrow \mathbf{v}_s$	74.76±2.08
DG$_A$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r$	92.41±0.21
DG$_A$-VRT$\rightarrow \mathbf{v}_s + \mathbf{v}_r + \mathbf{v}_t$	**94.68±0.44**

TABLE VI

INCEPTION SCORE COMPARISON ON THE BIRDS-200-2011 DATASET AND THE OXFORD-102 FLOWER DATASET

Methods	CUB	Oxford
StackGAN++	4.02±0.58	2.49±0.02
AttnGAN	4.31±0.68	3.36±0.02
DCGAN-S	4.29±0.07	3.29±0.08
DCGAN-A	4.51±0.04	3.39±0.02

C. Discussions

Our proposed DCGAN generates K visual plausible synthetic images conditioned on the text of query image only, which makes the generated synthetic images closely correlated to the corresponding real images, displaying key visual elements embedded in text and even providing more details about the underlying background and variations. Therefore, our proposed DG-VRT is more robust to visually represent text for improving the performance of image recognition.

To evaluate the quality of synthetic images generated by our DCGAN, the inception score [53] performances on the Oxford 102 category flowers dataset and Birds-200-2011 dataset are described in Table VI. The StackGAN++ and AttnGAN are used as the baselines. From Table VI, we confirm that the proposed DCGAN can effectively improve the quality of synthetic images.

To evaluate the correspondence between input text and the synthetic images, we also calculate the R-Precision score [7] on the two datasets. The larger the R-Precision score value, the better correspondence between input text and the synthetic image. The R-Precision score performance of our DCGAN is described in Table VII. From Table VII, our DCGAN-A has the best score on the two datasets, which means the synthetic images generated by DCGAN-A are better to match the input text.

We use the Number of Statistically-Different Bins (NDB) and the Jensen-Shannon Divergence (JSD) [54] to evaluate the extent of mode missing of generative models. The NDB determines the relative proportions of samples fallen into clusters predetermined by real data. Lower NDB score and JSD mean the synthetic image distribution approaches the real image distribution better. The NDB scores and JSD performance of our DCGAN on two datasets are described in Table VIII. From Table VIII, we confirm that the proposed DCGAN improves the diversity of synthetic image and maintains visual quality.

TABLE VII

R-PRECISION SCORE COMPARISON ON THE BIRDS-200-2011 DATASET AND THE OXFORD-102 FLOWER DATASET

Methods	CUB	Oxford
StackGAN++	10.57±4.83	13.66±1.44
AttnGAN	67.82±4.43	45.50±1.25
DCGAN-S	17.33±4.85	20.13±0.98
DCGAN-A	68.96±3.17↑	56.88±2.72↑

TABLE VIII

NDB AND JSD COMPARISON ON THE BIRDS-200-2011 DATASET AND THE OXFORD-102 FLOWER DATASET

methods		CUB	Oxford
StackGAN++	NDB	23.16± 0.31	42.53± 0.73
	JSD	0.192± 0.018	0.331± 0.042
AttnGAN	NDB	19.98± 0.21	28.11± 0.51
	JSD	0.164± 0.037	0.228± 0.032
DCGAN-S	NDB	20.53± 0.33	35.17± 0.48
	JSD	0.175± 0.042	0.284± 0.026
DCGAN-A	NDB	19.16± 0.52	26.61± 0.43
	JSD	0.149± 0.022	0.204± 0.026

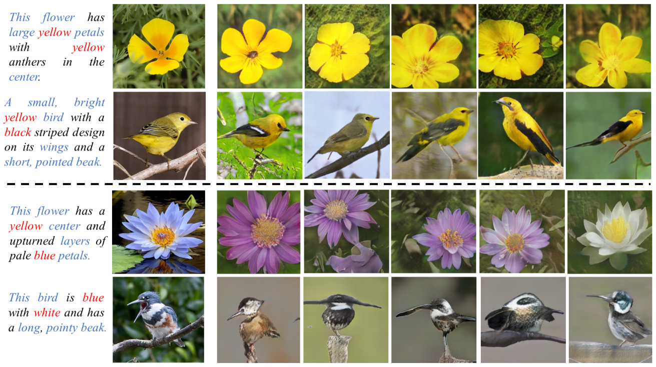

This flower has large yellow petals with yellow anthers in the center.
A small, bright yellow bird with a black striped design on its wings and a short, pointed beak.
This flower has a yellow center and upturned layers of pale blue petals.
This bird is blue with white and has a long, pointy beak.

Fig. 12. Visualization of a successful case (top) and a failure case (bottom) for our proposed VRT. From left to right are: text, real image, and 5 synthetic images, respectively.

We also visualize a successful case and a failure case in Fig. 12. If text describes a flower with detailed information about colors and shapes, our DCGAN is able to generate a bunch of informative synthetic images for leveraging the performance of image recognition. Otherwise, if a text is hard to understand, ambiguous and even misleading, then the generated synthetic images will not be consistent with each other to represent the visual concepts well.

D. Visual Representation on Text for Segmentation

To verify the effectiveness of visual representation on text, we extend it to the application of object segmentation.

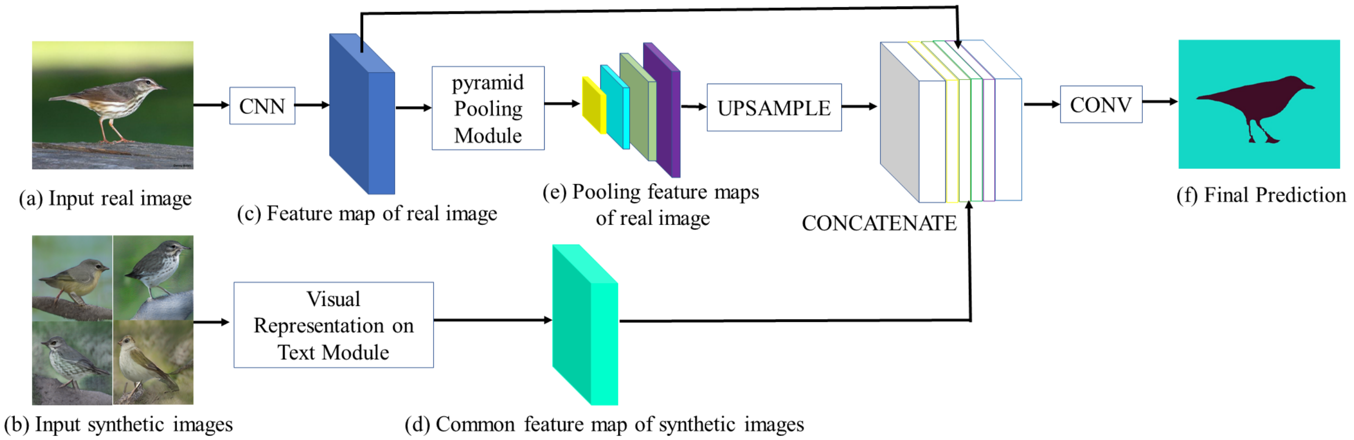

CNN
pyramid Pooling Module
UPSAMPLE
CONV
(a) Input real image
(c) Feature map of real image
(e) Pooling feature maps of real image
CONCATENATE
(f) Final Prediction
Visual Representation on Text Module
(b) Input synthetic images
(d) Common feature map of synthetic images

Fig. 13. The framework of PSPNet+DG$_A$-VRT.

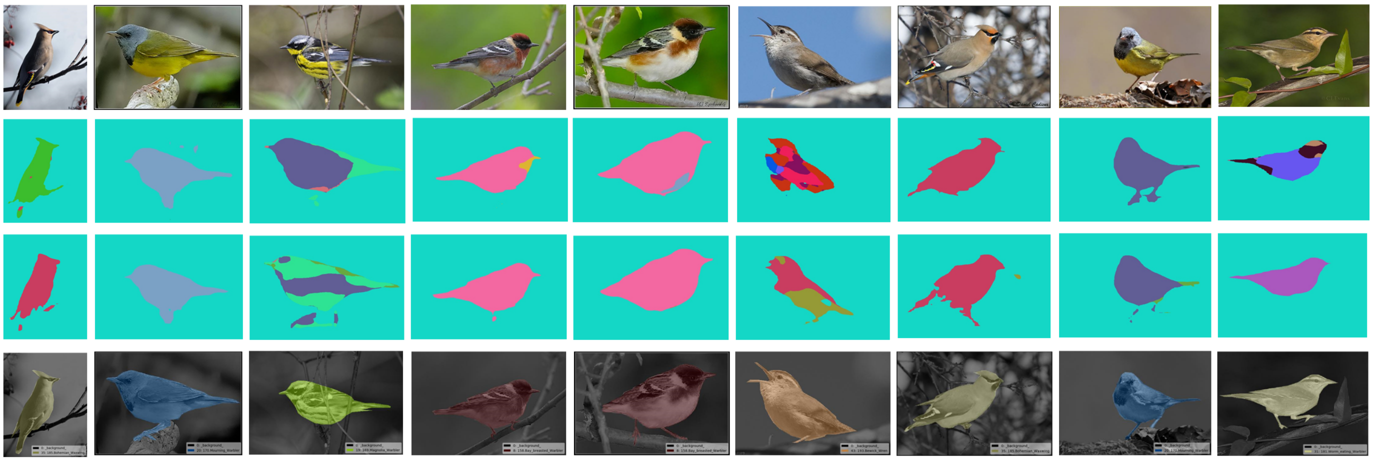


Fig. 14. Visualization of object segmentation results of PSPNet+DG$_A$-VRT (the third row) on the Caltech-UCSD Birds-200-2011 Dataset, compared with the visualization results of PSPNet [26] (the second row). From top to the bottom are real images, the results of PSPNet, the results of PSPNet+DG$_A$-VRT, and the corresponding ground truth of segmentation, respectively.

We use LabelMe [55] to generate the bird's segmentation ground truth on the Caltech-UCSD Birds-200-2011 dataset. The baseline of bird segmentation is PSPNet [26]. We extract the common visual feature $\mathbf{v}_s$ of $K = 3$ synthetic images based on the Resnet module in PSPNet, and the synthetic images are generated by DCGAN-A. Before full connection layer, we fuse the visual representation feature $\mathbf{v}_s$ with the result of pyramid pooling in global feature space.

To better understand the architecture of visual representation on text for segmentation, we first describe the framework of PSPNet+DG$_A$-VRT, as shown in Fig. 13. We use the image-level CNN (ResNet101) to extract the visual feature of synthetic images, which are generated by DCGAN-A conditioned on the same text. Then, we use the proposed visual representation on text module to compute the common visual feature of these synthetic images.

As far as we know, the architecture of PSPNet [26] consists of three parts: the image-level CNN, the pyramid Pooling Module and the final convolution layer. In order to ensure the role of the original image features, we just combine the common visual feature and the result of pyramid pooling module in global feature. Finally, we feed the combined feature to a convolution layer to predict the segmentation result, while the kernel size of convolution layer equals the number of segmentation classes.

TABLE IX

OBJECT SEGMENTATION APPLICATION ON THE CALTECH-UCSD BIRDS-200-2011 DATASET

Methods	Accuracy (%)	Loss
PSPNet	84.83±0.19	0.955 ±0.099
PSPNet+DG$_A$-VRT	84.99±0.38	0.939±0.046

The comparison results on testing dataset are summarized in Table IX, from which we can observe from two aspects. (1) PSPNet+DG$_A$-VRT improves the object segmentation accuracy compared with the PSPNet. (2) The loss of PSPNet+DG$_A$-VRT is lower than the PSPNet. Apparently, the visual representation in PSPNet+DG$_A$-VRT is robust and the visual synthetic image feature is complementary to both visual real image feature for object segmentation.

To better understand the efficiency of visual representation on text for segmentation, we visualize the segmentation results of PSPNet+DG$_A$-VRT on the Caltech-USCD Birds-200-2011 Dataset in Fig. 14. We also visualize the segmentation results of PSPNet in Fig. 14. As we can observe, the segmentation results of PSPNet+DG$_A$-VRT cover more segmentation regions in the image than those of the PSPNet. The PSPNet segments the same object into multiple objects with a higher probability than the PSPNet+DG$_A$-VRT, such as the columns 6 and 9 in Fig 14.

V. Conclusion

In this paper, we have proposed a novel DG-VRT to visually represent text with adversarial learning for image recognition. With our DCGAN, K images are generated conditioned on a text to represent the visual concepts. The visual synthetic image feature has been proved to be able to improve accuracy for image recognition, and it is complementary to real image features and text features. The experimental results conducted on two datasets well support our claims in the paper.

Our future work includes further exploring our DCGAN to produce better quality of synthetic images, and extending our DG-VRT to solve more general multimedia problems. The proposed DG-VRT cannot effectively handle visual recognition in complex scenes, such as there are lots of object in the image. We will explore the generative adversarial networks with DG-VRT under multi-object and multi-classification tasks to improve the visual recognition performance in complex datasets. We use four Nvidia 1080Ti GPUs to train the proposed DG-VRT with Pytorch. We also verify the proposed DG-VRT on Huawei MindSpore platform, and will explore more complex DCGAN and visual recognition models in MindSpore platform.

Acknowledgment

This work was co-supervised by Chengjiang Long and Chunxia Xiao.

References

[1] J. J. McAuley and J. Leskovec, "Image labeling on a network: Using social-network metadata for image classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 828–841.

[2] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2960–2968.

[3] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2872–2880.

[4] F. Fang, M. Yi, H. Feng, S. Hu, and C. Xiao, "Narrative collage of image collections by scene graph recombination," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 9, pp. 2559–2572, Sep. 2018.

[5] C. Long, R. Collins, E. Swears, and A. Hoogs, "Deep neural networks in fully connected CRF for image labeling with social network metadata," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1607–1615.

[6] J. Johnson, L. Ballan, and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4624–4632.

[7] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[8] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[9] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[10] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.

[11] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: Manipulating images with natural language," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 42–51

[12] F. Fang, F. Luo, H. P. Zhang, H. J. Zhou, A. L. Chow, and C. X. Xiao, "A comprehensive pipeline for complex text-to-image synthesis," *J. Comput. Sci. Technol.*, vol. 35, no. 3, pp. 522–536, 2020.

[13] F. Huang, X. Zhang, Z. Li, T. Mei, Y. He, and Z. Zhao, "Learning social image embedding with deep multimodal attention networks," in *Proc. Thematic Workshops ACM Multimedia (Thematic Workshops)*, 2017, pp. 460–468.

[14] J. Johnson, L. Ballan, and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4624–4632.

[15] Y. S. Rawat and M. S. Kankanhalli, "ConTagNet: Exploiting user context for image tag recommendation," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1102–1106.

[16] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 562–570.

[17] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.

[18] R. Hu, M. Rohrbach, S. Venugopalan, and T. Darrell, "Utilizing large scale vision and text datasets for image segmentation from referring expressions," 2016, *arXiv:1608.08305*. [Online]. Available: http://arxiv.org/abs/1608.08305

[19] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7454–7463.

[20] C. Long, G. Hua, and A. Kapoor, "Active visual recognition with expertise estimation in crowdsourcing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 3000–3007.

[21] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active learning of a kernel machine ensemble for recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1209–1216.

[22] C. Long and G. Hua, "Multi-class multi-annotator active learning with robust Gaussian process for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2839–2847.

[23] C. Long, G. Hua, and A. Kapoor, "A joint Gaussian process model for active visual recognition with expertise estimation in crowdsourcing," *Int. J. Comput. Vis.*, vol. 116, no. 2, pp. 136–160, Jan. 2016.

[24] C. Long and G. Hua, "Correlational Gaussian processes for cross-domain visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 118–126.

[25] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active visual recognition from crowds: A distributed ensemble approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 582–594, Mar. 2018.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.

[28] S. Zhao and Z. Zhang, "Attention-via-attention neural machine translation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[29] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4385–4395.

[30] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," 2021, *arXiv:2101.00545*. [Online]. Available: https://arxiv.org/abs/2101.00545

[31] J. H. Tan, C. S. Chan, and J. H. Chuah, "COMIC: Toward a compact image captioning model with attention," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2686–2696, Oct. 2019.

[32] W. Zhang and C. Xiao, "PCAN: 3D attention map learning using contextual information for point cloud based retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12436–12445.

[33] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10213–10222.

[34] A. Islam, C. Long, A. Basharat, and A. Hoogs, "DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4676–4685.

[35] D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, and C. Xiao, "ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8139–8148.

[36] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A fully-trained generative adversarial networks for text to face generation," 2019, *arXiv:1904.05729*. [Online]. Available: https://arxiv.org/abs/1904.05729

[37] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7986–7994.

[38] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1060–1069.

[39] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1429–1437.

[40] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6199–6208.

[41] Q. Hoang, T. D. Nguyen, T. Le, and D. Q. Phung, "MGAN: Training generative adversarial nets with multiple generators," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, p. 24.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[44] W. Yin and H. Schütze, "Multichannel variable-size convolution for sentence classification," 2016, *arXiv:1603.04513*. [Online]. Available: http://arxiv.org/abs/1603.04513

[45] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[46] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process. (ICVGIP)*, Dec. 2008, pp. 722–729.

[47] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6629–6640.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[51] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 70–86.

[52] B. J. Meyer, B. Harwood, and T. Drummond, "Nearest neighbour radial basis function solvers for deep neural networks," 2018, *arXiv:1705.09780v3*. [Online]. Available: https://arxiv.org/abs/1705.09780

[53] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 2234–2242.

[54] E. Richardson and Y. Weiss, "On GANs and GMMs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 5852–5863.

[55] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, pp. 157–173, 2008.

Tao Hu (Member, IEEE) received the B.S. degree in software engineering from the School of Computer, Wuhan University of Technology, Wuhan, in 2006, the M.S. degree in software engineering from the School of Software and Microelectronics, Peking University, Beijing, in 2009, and the Ph.D. degree in computer science from the School of Computer Science, Wuhan University, Wuhan, in 2020. He currently works as an Associate Professor with the School of Information Engineering, Hubei Minzu University, Enshi, China. His current research interests include deep learning, computer animation, and image processing.

Chengjiang Long (Member, IEEE) received the B.S. degree in computer science and technology and the M.S. degree in computer science from Wuhan University in 2009 and 2011, respectively, and the Ph.D. degree in computer science from the Stevens Institute of Technology in 2015. He has been working as a Principal Scientist with JD Tech at Silicon Valley R&D Center (a part of JD.COM) since June 2020. Prior to working at JD.COM, he worked as a Computer Vision Researcher/Senior Research and Development Engineer at Kitware from February 2016 to April 2020. He also worked as an Adjunct Professor with University at Albany, SUNY, from August 2018 to May 2020, and was an Adjunct Professor with Rensselaer Polytechnic Institute (RPI) from January 2018 to May 2018. During his Ph.D. study, he worked at NEC Labs America and GE Global Research, as a Research Intern, in 2013 and 2015, respectively. He has published over 45 papers, including top journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*, top international conferences, such as CVPR, ICCV, and AAAI, and holds one patent. His research interests include the various areas of computer vision, machine learning, artificial intelligence, and computer graphics. He is a member of AAAI. He is also a reviewer of more than 20 top international journals and conferences.

Chunxia Xiao (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Mathematics, Hunan Normal University, in 1999 and 2002, respectively, and the Ph.D. degree from the State Key Laboratory of CAD & CG, Zhejiang University, in 2006, China. He became an Assistant Professor at Wuhan University in 2006, where he became a Professor in 2011. During October 2006 to April 2007, he held a postdoctoral position with The Hong Kong University of Science and Technology, and he visited the University of California at Davis from February 2012 to February 2013. He is currently a Professor with the School of Computer Science with Wuhan University, China. He has conducted a considerable amount of research in geometry reconstruction, image editing, computational photography, image analysis, and synthesis, and he has published more than 110 papers in journals and conferences. His research areas include computer graphics, computer vision, virtual reality, and augmented reality. He is a member of ACM.