Neurocomputing 293 (2018) 55–63

Contents lists available at ScienceDirect

Neurocomputing

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

Semantic invariant cross-domain image generation with generative adversarial networks

Xiaofeng Mao[a], Shuhui Wang[b], Liying Zheng[a], Qingming Huang[b,c,*]

[a] College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[b] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[c] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 23 October 2017
Revised 10 February 2018
Accepted 13 February 2018
Available online 8 March 2018

Communicated by Jiwen Lu

MSC:
41A05
41A10
65D05
65D17

Keywords:
Generative adversarial networks
Image-to-image translation
Semantic invariance

ABSTRACT

Recently, thanks to the state-of-the-art techniques in Generative Adversarial Networks, a lot of work achieves remarkable performance on learning the mapping between an input image and an output image without any paired relation. However, traditional methods on image-to-image translation merely consider the visual appearance properties, they fail to maintain the true semantics of an image during the transfer learning procedure from source to target domain. We propose a new approach that utilizes GAN to translate unpaired images between domains and remain high level semantic abstraction aligned. Our model controls the hierarchical semantics of images by processing semantic information on label level and spatial level respectively by constructing label and attention consistent losses. The experimental results on several benchmark datasets show that generated samples are both visually similar with target images and semantically consistent with their source counterparts. Furthermore, the experiment also suggests that our method can effectively improve the classification performance in unsupervised domain adaptation problem.

1. Introduction

There are a lot of interesting applications on translating images from one domain to another domain [1–5], *e.g.*, facial image to emoji, face swapping, sketch to photos, etc. Isola et al. [6] defines such problem of translating one possible representation of a scene into another as automatic *image-to-image translation*. Years of research in computer vision, image processing, and graphics have produced powerful translation systems in the supervised setting, where example image pairs $\{x, y\}$ are available. Most of them are formulated as per-pixel classification or regression [7–10]. However, even if pixel-wise models can reach a high accuracy and image quality, huge efforts in collecting paired training data is a prerequest. Fortunately, some efficient frameworks were developed to handle this trouble recently [1,3,4,11–13] based on Generative Adversarial Networks (GANs) [14]. Radford et al. [15] found that GAN is prone to build good image representations, as it provides an attractive alternative to maximum likelihood techniques. Because of this excellent characteristic, GAN is widely applied to the field of image synthesis [16–19]. Subsequent works went a step further and combined GAN with unpaired image-to-image translation [1–3,6]. They firstly apply generative models to align different domains or learn a joint distribution without pair information. Since GAN leaves out the tedious process of collecting training data, it gradually becomes an important method for image-to-image translation problem.

Recently, an impressive work CycleGAN [11] exploits the property that translation should be "*cyclically consistent*", in the sense that if we can translate an image to a new one, we can go back to the original image with a similar translation process as well. It uses element-wise metrics between original images and reconstructed images. However, element-wise measures like the squared error are not very suitable for image data, as they do not model the properties of human visual perception. As shown in Fig. 1a, a CycleGAN directly learns the underlying relationship between MNIST and SVHN. Even though the output data distribution was well fitted, the learned relationship could be wrong. After experimental analysis, we think the performance degradation depends on that GAN fails to capture the real semantic concept of images. Therefore, we argue in favor for measuring image similarity using a higher-level and sufficiently invariant representation of the images. Recently, Mirza and Osindero [20] proposed an often cited

* Corresponding author at: School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.
E-mail address: qmhuang@ucas.ac.cn (Q. Huang).

https://doi.org/10.1016/j.neucom.2018.02.092
0925-2312/

(a) CycleGAN (b) CycleGAN with label auxiliary

Fig. 1. A comparison of impact on results between whether to use label information or not. We applied CycleGAN to generate MNIST images based on SVHN in both (a), (b) and the latter was conditioned on labels. We found generated samples in (a) are totally meaningless while for the other one, semantics can be transmitted to output successfully.

technique that generative adversarial nets can be extended to a conditional model if both the generator and discriminator are conditioned on some extra information (such as class labels or data from other modalities). Through performing the conditioning by feeding conditional vectors into the both the discriminator and generator as additional input layer, we can direct the data generation process. For example, a conditioned version of CycleGAN with integration of technologies of [20,21] in Fig. 1b. This trick is widely used in various GAN-based models [1,5,13,20].

More formally, suppose that we are given one set of input images in domain X^c and one set of target images in domain Y^c, respectively, where c specifies the category of the sample. A mapping: $X \rightarrow Y$ is trained to induce an output distribution $P(Y^c|X^c, c)$ that matches the empirical distribution $P(Y^c)$, e.g., we minimize $\sum_c D_{KL}\left(P\left(Y^c\middle|X^c, c\right)\|P(Y^c)\right)$ if Kullback–Leibler divergence is used. It is worth noting that this method can only work on when class labels in both domains are available. However, in practice, label information is always lacking in target domian. We probed further and found out that this formulation is related to the unsupervised domain adaptation problem. Under that condition, both the label $\hat{c}$ of target images and an distribution $P\left(Y^{\hat{c}}\middle|X^c, c\right)$ need to be induced to fit the empirical distribution $P(Y)$. Based on the discussion above, for a specified category c, we abandon the meaningless mappings when $\hat{c} \neq c$. Eventually, the learned distribution can be formulated as $P\left(Y^{\hat{c}}\middle|X^c, \hat{c} = c\right)$. In general, the generated cross-domain image must meet two requirements: appearance similarity with target domain images and consistent semantic concept with source domain counterparts.

In this paper, we purpose a learning model for generating semantic invariant cross-domain images. We utilize GAN to adapt images from the source domain to make them appear as if they were sampled from the target domain. For keeping the semantics of generated image consistent with input image, we add constraints to label level and spatial level. The former is necessary as it makes $\hat{c} = c$ stand. Motivated by [22], we use a classifier to get predict labels on both input and generated samples and construct label consistent losses. In addition, for getting a stable image structure, we align high-level spatial information between domain by applying attention maps of the intermediate layers in the network. The fine grained semantic alignment process is well-distinguished from traditional method in image-to-image translation, and the experimental results shows the stability and quality of generated images achieve a great promotion. Throughout the training process, we only used source domain labels as a priori for semantic extraction. Our model can be viewed as a generative method for resolving unsupervised domain adaptation problems because it directly fits the data distribution in the target domain. Our algorithm consists of two steps: adversarial training and adaptive training. The former optimizes a generation network which applies GAN to produce outputs that similar with images sampled from the target domain. In adaptive training stage, a classification network is trained to extract shared feature which contains all the semantic information between domains. Our model architecture details can be seen in Fig. 2.

In this paper, we make the following contributions:

- We extend the conditioned version of unpaired image-to-image generation task to an unsupervised scenario (target domain labels are unavailable) and propose Semantic Invariant GAN for generating semantic invariant cross-domain images which can be also used in unsupervised domain adaptation.
- Our model controls the hierarchical semantics of images by processing semantic information on label level and spatial level. A series of losses are introduced to retain image semantics and improve quality of generated samples. In addition, we found they are also helpful for other GAN-based models.
- Experimental results on SVHN-MNIST and Office-Home datasets has validated that our model could make use of source domain labels and produced more meaningful samples. The proposed method also achieves state-of-the-art classification accuracy on unsupervised domain adaptation problem.

This paper is organized as follows: We briefly introduce some related works in Section 2. Section 3 presents the problem formulation of cross-domain image generation and a baseline method to address this problem. In Section 4, we describe in detail the formulation of our approach and the iterative training procedure. In Section 5, attention mechanism is introduced to explain how attention loss can help to retain the image semantics. In Section 6, we give a introduction to the experimental setups and compare the running results of our model and previous model. We conclude this paper in Section 7.

2. Related work

Our research is closely related to unpaired image-to-image translation and unsupervised domain adaptation. We briefly review recent representative works as follows.

Unpaired image-to-image translation The goal of unpaired image-to-image translation is to relate two data domains X and Y in unpaired setting. Rosales et al. [23] propose a Bayesian framework that includes a prior based on a patch-based Markov random field computed from a source image, and a likelihood term obtained from multiple style images. More recently, many methods explore about GAN-based joint distribution learning. The most primitive one is CoupledGAN [12], which trains a coupled generative model that learns the joint data distribution across the two

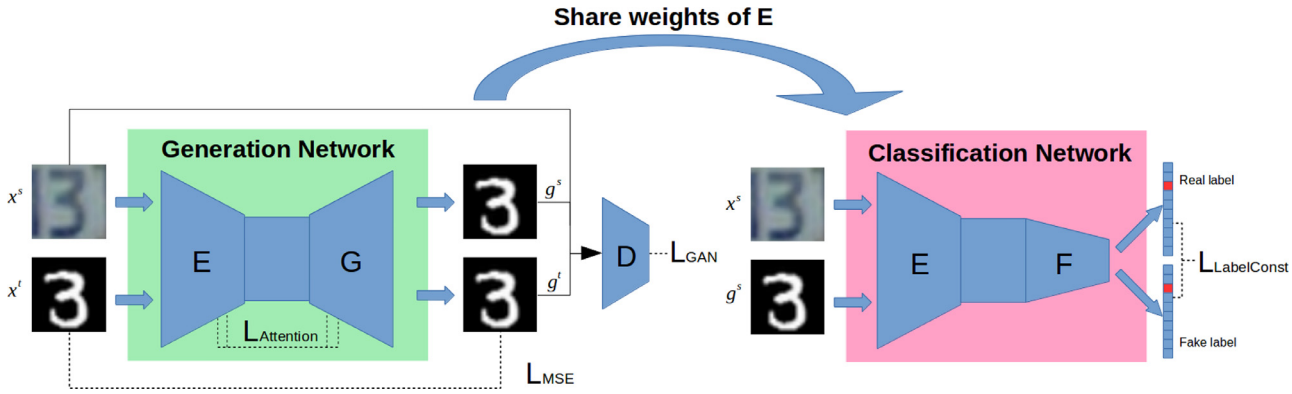

Fig. 2. Our network architecture. Solid line represents data flow in the network, dotted line refers to the loss item that needs to be calculated. The right half was classification network used to extract semantic information from pictures, and the left part was generation network. Inputs of generation network x^s, x^t refer to source and target samples respectively. And g^s, g^t are the outputs of the network with the corresponding inputs. Here g^s is the adapted image from source domain to target domain, and g^t is the reconstruction image from target sample. The weights of encoders were shared. Note that E and F are pretrained using labeled source images $\{\mathbf{x^s}, y\}$ and $\mathbf{h}$ is the representation encoded by E.

domains. However, CoGAN is difficult to converge when training images are more complicated. Another concurrent work CycleGAN [11] exploits the property that translation should be "*cycle consistent*", in the sense that if we translate an image to a new one, we should go back to the original image.

Some other methods learn only one directional mapping (source-to-target). A feature consistency constraint is added in domain transfer network [2] to directly obtain a shared latent representation between domains. Another work called AlignGAN [13] is based on conditional GAN [20]. It learns domain-specific semantics by the conditioned domain vectors and the shared semantics by the other shared latent vectors. Unlike the methods discussed above, Sankaranarayanan et al. (2017) [24] regards image generation as a sub-task for domain adaptation, and forces the encoder to produce the embeddings for the target data in the domain alignment process.

Unsupervised domain adaptation In this section, we focus on visual domain adaptation since it is more relevant to our work. Earlier approaches to domain adaptation focussed on building feature representations that are invariant across domains. This was accomplished either by feature reweighting and selection mechanisms [25], or by learning an explicit feature transformation that aligns source distribution to the target [26,27].

Recently, Deep Neural Networks have been shown to be successful in unsupervised domain adaptation. We group deep learning methods for visual domain adaptation into two categories. One line of work uses discriminative based model. The Deep Adaptation Network (DAN) [28] applied the Maximum Mean Discrepancy (MMD) [29] metric to layers embedded in a reproducing kernel Hilbert space, effectively matching higher order statistics of the two distributions. In contrast, the deep Correlation Alignment (CORAL) [30] method proposed to match the mean and covariance of the two distributions. Ajakan et al. [31] introduced the Domain Adversarial Neural Network (DANN): an architecture trained to extract domain-invariant features. Another class of works are based on generative models. CoGAN [12] approach applied GANs to the domain transfer problem by training two GANs to generate the source and target images respectively. The most latest work [22] learns a transformation in the pixel space from one domain to the other in an unsupervised manner.

3. Problem formulation

We give a problem formulation and a traditional GAN based method on image-to-image translation. Suppose that we have a set

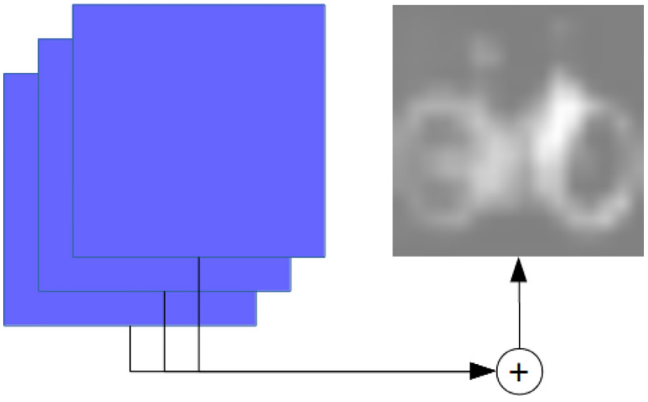

Fig. 3. An illustration of calculating process of attention maps.

of images $\mathbf{x^s}$ in source domain S and a set of images $\mathbf{x^t}$ in target domain T. The problem is learning a mapping $\Phi: S \rightarrow T$. For an input image $x \in \mathbf{x^s}$, a generated image $\phi(x)$ is like a sample in target domain T and has closed semantics with input x.

We use an encoder-decoder network as generator and a classification network as discriminator. Adversarial training procedure is consistent with the traditional GAN: generator G is trained to produce outputs that cannot be distinguished from target images by an adversarially trained discriminator, D, which is trained to do as well as possible at detecting the images are not sampled from target domain. A formalized expression of this method is as follows.

Let $E(\mathbf{x};\ \theta_e)$ be a function defined on source domain which maps an image $\mathbf{x}$ to a hidden representation $\mathbf{h}$ representing features that are inclined to classification. A generator $G(\mathbf{h};\ \theta_g)$ is constructed to decode $\mathbf{h}$ for a new image and a discriminator $D(\mathbf{x};\ \theta_d)$ to detect images were not sampled from target domain. The discriminator, whose output neuron is activated by sigmoid, outputs the likelihood that a given image x has been sampled from the target domain. The generator is updated to produce images which are discriminated to be true by discriminator. Finally, we optimize a standard GAN loss as follows:

$$\begin{aligned} \pounds_{GAN} = & \min_{\theta_g} \max_{\theta_d} \mathbb{E}_{\mathbf{x^t}}\left[\log D\left(\mathbf{x^t};\theta_d\right)\right] \\ & + \mathbb{E}_{\mathbf{x^s}}\left[\log\left(1 - D\left(G\left(E\left(\mathbf{x^s};\theta_e\right);\theta_g\right);\theta_d\right)\right)\right] \end{aligned} \tag{1}$$

where θ_g, θ_d, θ_e represent weights of generator, discriminator and encoder. $\mathbb{E}_{\mathbf{x^t}}$, $\mathbb{E}_{\mathbf{x^s}}$ are the expectations on the set $\mathbf{x^t}$, $\mathbf{x^s}$, here $\mathbf{x^t} \sim T$ and $\mathbf{x^s} \sim S$.

Fig. 4. Sample images from the Office-Home dataset. The dataset consists of images of everyday objects organized into 4 domains; Art: paintings, sketches and/or artistic depictions, Clipart: clipart images, Product: images without background and Real-World: regular images captured with a camera. The figure displays examples from 16 of the 65 categories.

However, this method only transfers the appearance but fails to align high-level semantics. It means generated samples can be pointless. In addition, training GANs is well known for being delicate and unstable, for reasons theoretically investigated in [32].

4. Semantic invariant GAN

We propose a more robust architecture for cross-domain image generation, which contains two sub-networks. The main part is generation network, whose structure remains fixed with the baseline method in Section 3. The other named classification network is used to extract discriminative semantic information. Motivated by Bousmalis et al. [22], if generator has sufficient capacity and eventually converged to the optimum, an output sample will have large probability to be classified into same category as its counterpart in source domain. Therefore, the model is augmented with a classifier $F(\mathbf{h}, \theta_f)$ to get task-specific labels of output samples. Then $\pounds_{LabelConst}$ is constructed as we wish task-specific labels more closer to source labels.

For a detailed description about this modifications. Let $\mathbf{g^s}$ be generated samples given by

$$\mathbf{g^s} = G\left(E\left(\mathbf{x^s}; \theta_e\right); \theta_g\right) \tag{2}$$

where $\mathbf{x^s}$ is the input (source images) and $\mathbf{g^s}$ is the corresponding output (adapted images).

We compute the task-specific labels $\hat{y}$, which stand for the probability of adapted image $\mathbf{g^s}$ that was classified in every category, by feed $\mathbf{g^s}$ to classification network (Note that weights of E in generation network and classification network are shared),

$$\hat{y} = F\left(E\left(\mathbf{g^s}; \theta_e\right); \theta_f\right) \tag{3}$$

Finally, $\pounds_{LabelConst}$ was a cross entropy loss between $\hat{y}$ and ground truth y from source image and label pairs $\{\mathbf{x^s}, y\}$ (in the form of one-hot),

$$\pounds_{LabelConst} = -y \cdot \log \hat{y} \tag{4}$$

$\pounds_{LabelConst}$ controls high level semantic abstraction of generated images. This loss has the same form and purpose as the "*task-specific loss*" in [22]. Apart from $\pounds_{LabelConst}$, DTN [2] puts forward a hypothesis that if generated images $G(x)$ are well adapted, deep features $f(G(x))$ should be identical to $f(x)$, where f is a convolutional network or multilayer perceptron. This formulation is called *f-constancy* term. Even though all of these losses are used to make the adapted images remain invariant semantics, their effect can be distrincted. We consider *f-constancy* being a stronger constraint than others. It can improve the quality and stability of the generated images but fail to generate images with more complex structures. Because a stronger constraint tends to limit the pattern of generated samples, which is a widely discussed problem in GAN training called "mode collapse". The contrast experimental results in Fig. 7 also corroborate the above inference and finally we adopt *label consistent loss* instead of *f-constancy*.

However, we cannot get invariant semantics merely according to consistency on label level. Images classified to the same category always have a distinctive spatial structure. For a two-dimensional image, spatial information is extremely important. A growing number of computer vision models discard the vector representation and use more convolution structures [8]. For this reason, we propose to align high-level spatial information between domain by applying feature maps of the intermediate layers in the network. This practice can be explained as attention mechanism [33]. So, we construct a fine grained semantic alignment loss called attention loss, which we will make a detailed introduction in Section 6.

Constraints mentioned above cannot guarantee that the learned hidden representations contain the whole hinge information in target domain. Based on [34], a hypothesis conveys that a domain-adaptive representation should satisfy two criteria: (i) classify well the source domain labeled data and (ii) reconstruct well the target domain unlabeled data. If target images are encoded by encoder E, we must assure that original image can be restored easily from these encoded representations. So, $\pounds_{MSE}$ between generated and original images is adding to objective function.

Let $\mathbf{g^t}$ be output when feed target images $\mathbf{x^t}$ to generation network, as we wish, $\mathbf{g^t}$ must be resemble with original $\mathbf{x^t}$. $\pounds_{MSE}$ can be formulated as follows:

$$\pounds_{MSE} = d_1\left(G\left(E\left(\mathbf{x^t}; \theta_e\right); \theta_G\right), \mathbf{x^t}\right) \tag{5}$$

where d_1 is similarity metric function. There are many forms of d_1 including Mean Squared Error (MSE) and cosine distance, as well as other variants including hinge and triplet losses. The performance is mostly unchanged, and we report results using the simplest MSE solution.

In order to slightly smooth the resulting image, we added an anisotropic total variation loss [35],

$$\pounds_{TV}(z) = \sum_{i,j}\left(\left(z_{i,j+1} - z_{i,j}\right)^2 + \left(z_{i+1,j} - z_{i,j}\right)^2\right)^{\frac{B}{2}} \tag{6}$$

here $z = \mathbf{g^s}$, i and j represent the specific location of an image pixel. this is an optional loss which only for getting more flatness and beautiful results. In the experiment, we employ $B = 1$.

Our algorithm can be summarized as in Algorithm 1.

Algorithm 1 Semantic invariant GAN.

Input: input $\mathbf{x^s}$, $\mathbf{x^t}$, source label y^s, hyper parameters: multiplier of penalized item α, β, γ, classification network training interval t

Output: adapted samples $\mathbf{g^s}$ with invariant semantics

pretrain parameters of classification network θ_e and θ_f using $\mathbf{x^s}$, y^s
for max iteration i **do**
  feed $\mathbf{x^s}$, $\mathbf{x^t}$ as input updating parameters of generative and discriminative network θ_e, θ_g, θ_d for optimizing adversarial loss $\pounds_{adv} = \pounds_{GAN} + \alpha\pounds_{Attention} + \beta\pounds_{MSE} + \gamma\pounds_{TV}$
  if i *mod* $t == 0$ **then**
    feed $\mathbf{x^s}$, y^s as input adapting classification network parameters θ_e, θ_f for optimizing loss $\pounds_{LabelConst}$
  else
    pass
return output samples $\mathbf{g^s}$

5. Attention loss

Single objects are rendered on black back-grounds and consequently we expect images adapted from these renderings to have similar foregrounds and different backgrounds from the equivalent source images. Z-buffer masks can help to differentiate between foreground and background pixels. By using z-buffer masks as prior knowledge, a content-similarity loss can be calculated that penalizes large differences between source and generated images for foreground pixels only [22]. But z-buffer masks are not available if training data are exclusively RGB images and it is tough to let network know which part of training images will be foreground.

Enlightened by the studies on image saliency [36–38]. We construct a constraint term called attention loss [33] that can determine and distinguish which area of images play a key role in the classification and which is barely irrelevant area. Attention, which is therefore as a key aspect of our visual experience, has recently been demonstrated can play an important role in the context of applying artificial neural networks to a variety of tasks from fields such as computer vision and NLP [33,39–42]. Attention is considered as a set of spatial maps that essentially try to encode on which spatial areas of the input the network focuses most for taking its output decision (e.g., for classifying an image), where, furthermore, these maps can be defined on various layers of the network so that they are able to capture both low-, mid-, and high-level representation information. There are two types of spatial attention maps: activation-based and gradient-based. We used activation-based attention maps due to its simplicity of calculation.

Consider a feature map $T \in R^{C\times H\times W}$, which has size of $H \times W$ and C channels. An activation-based mapping function P takes as input the above tensor and outputs a spatial attention map:

$$P : R^{C\times H\times W} \rightarrow R^{H\times W} \tag{7}$$

The implicit assumption that we make in this section is that the absolute value of a hidden neuron activation (that results when the network is evaluated on given input) can be used as an indication about the importance of that neuron on the specific input. In our network, we use sum of absolute values as mapping function (Fig. 3):

$$P(T) = \sum_{i=1}^{C} |T_i| \tag{8}$$

where T_i is the ith channel in feature map.

It is important to understand why attention map $P(T)$ can be applied in fine grained semantic alignment process on spatial level.

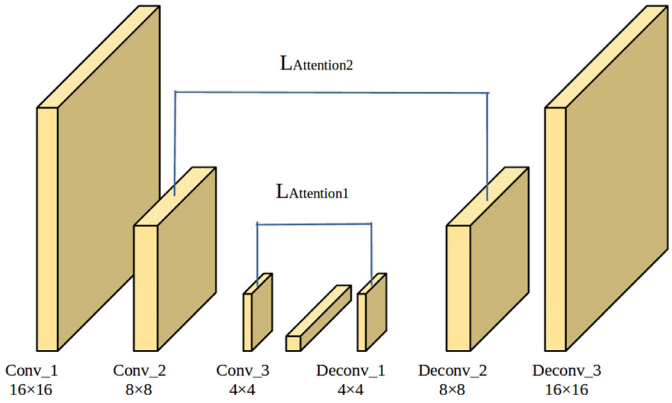

Fig. 5. Attention maps chose for calculating loss in the experiment.

We consider a particular channel in the feature map as the response of a feature filter (convolution kernel). For a specific input, if the activation values in one region of the feature map tends to 0, we think that the region in the image is suppressed, or that it has little effect on the output of the network. Conversely, if a region is often activated (when using a non sparse activation function like *sigmoid*, it means that the activation value is a large positive or small negative number), this region is the critical areas that affect the result. The attention map shows the concerning degree of the model to different regions in an image. Therefore, it represents the semantically significant spatial information learned by the network.

Under unsupervised cross-domain generation settings, we think of attention map as the additional knowledge network learned from the labeled data. For source image and adapted image, we expect them to have similar high-level spatial representation information. That is to say, source images and adapted images are required not only to belong to the same category, but to have the same critical areas that affect the result of classification. Therefore, attention map defined on high layers of shared feature extractor and low layers of generator must be aligned as far as possible.

Let $P(G^l)$ indicate attention map defined on lth layer of generator, $P\left(E^{L-l}\right)$ means attention map defined on last lth layer of shared feature extractor, $\pounds_{Attention}$ can be formulated based 8 as:

$$\pounds_{Attention} = d_2\left(P\left(G^l\right) - P\left(E^{L-l}\right)\right) \tag{9}$$

where d_2 is similarity metric function the same as that in (5). And $L - l$ represents the last lth layer of a network which has layers of L.

6. Experimental results and analysis

In this section, extensive experiments were conducted to evaluate the proposed approach on the two challenging tasks: unpaired image-to-image translation and unsupervised domain adaptation. In first task, we transfer images from the domain of Street View House Number (SVHN) [43] dataset of Netzer et al. (2011) to MNIST [44] dataset by LeCun et al. (2010). Then we generalize from this task to more complex situations, whose objective is to generate clipart images using real world natural images on Office-Home [45] dataset. In next task, we apply the model to unsupervised visual domain adaptation scenario and evaluate the classification performance on the test split of target images.

6.1. Unpaired image-to-image translation

Datasets. We conduct an experiment on object classification benchmarks and try some innovative and challenging cross-domain image generation tasks listed below:

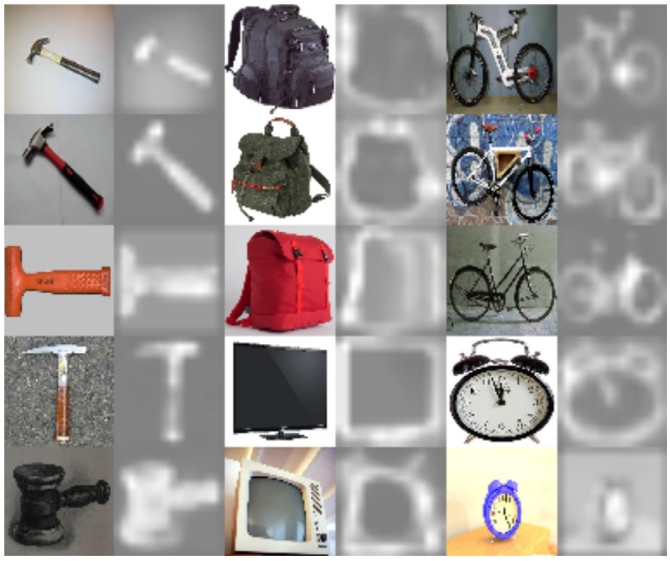

Fig. 6. Attention mechanism visualization.

SVHN to MNIST. We use the popular MNIST dataset of handwritten digits as the source domain and SVHN dataset representing the target domain. MNIST contains 70000 28×28 size images of the 10 digits (0–9) with black background and we use standard splits, comprising of 60,000 training, 10,000 test. Contrast with the former, SVHN with 32×32 size are significantly different in appearance and have diverse profile of background and color. In the experiment, we train classifier on a the original 73,257 SVHN training images and use all of 70,000 MNIST images for fitting data distribution. Then the model was evaluated on 26032 SVHN testing images.

Office-Home Real-World to Clipart. Office-Home [45] dataset is newly created to evaluate domain adaptation algorithms for object recognition. It consists of images from 4 different domains: Artistic images, Clip Art, Product images and Real-World images (See more information in Fig. 4). For each domain, the dataset contains images of 65 object categories found typically in Office and Home settings. We choose Real-World to Clipart direction for evaluating our algorithm.

6.1.1. SVHN to MNIST

In this section, we will describe implementation details of SVHN-MNIST transfer. Our improved model realized by TensorFlow [46], makes some modifications based on DTN [2]. We keep the network architecture fixed with DTN and add some novel loss functions. In order to keep input tensor to have the same shape, MNIST images are resized to 32×32 using Lanczos resample algorithm and then replicated the dim of channel three times to be consistent with SVHN input. In both experiments, multiplier of penalized item α, β, γ are set into 0.01, 1 and 0.0001, classification network training interval t is set into 20.

In preparation phase, source training data are fed into network to pretrain E and F. E is critical part of the model, which determines failure or success of training. Note that parameters of F remained fixed in other stages except for pretraining stage. We set batch size 200 to ensure stability of the adversarial training process and all model parameters are optimized by RMSprop [47]. For enhancing the quality of generated images, a series of advanced loss functions are used. Here we analyze the practical role of these loss in the experiment below:

Attention loss. The symmetry of E and G must be guaranteed when we calculate $\pounds_{Attention}$, that means if there are 8×8 feature maps in encoder, we must ensure generator has 8×8 feature maps too. $\pounds_{Attention}$ here is computed by the formula:

$$\pounds_{Attention} = \pounds_{Attention1} + \pounds_{Attention2} \tag{10}$$

We give an intuitive description of $\pounds_{Attention1}$ and $\pounds_{Attention2}$ in Fig. 5. Attention loss is calculated on both 8×8 and 4×4 feature maps. The former is for capturing the details of the image and the latter is for the sake of getting more consistent semantic information. Indeed, Attention loss contributes to improve the quality of the generated image, highlight the edge of the object. We randomly selected some samples to perform contrast experiment and obtained two groups of data, one taking attention loss into account and another not. As the result shown in Fig. 6a and6c, generated digits using attention loss have clearer edge than those not. In the experiment, we found that 8×8 and 4×4 feature maps were equally essential. Discarding any of them will make a performance degradation.

LabelConst loss. Original images and adapted images having similar high-level semantics can be seen as a strong priori. Previous work adds this priori to networks by using feature consistent loss [2], which constraints deep feature of source images and corresponding adapted images to be as close as possible. Differently, our network aligns high-level semantics using LabelConst loss, due to the latter constraint is more relaxed than feature consistent loss. This endows more freedom to network on image generation.

GAN loss. A lot of work has been done to ameliorate the convergence and stability of traditional GAN [48–51]. We found these improvements are indeed beneficial for getting higher quality pictures in our network. Wasserstein GAN [48], as an ingenious modification to conventional adversarial loss, was widely used in later GAN related work because of its excellent performance and simplicity of implementation. We adopted Wasserstein loss as optimization objective of adversarial training in the experiment, instead of original cross entropy function. The exact form of Wasserstein loss is defined as follows:

$$\pounds_{WGAN} = \min_{\theta_g} \max_{\theta_d} \mathbb{E}_{\mathbf{x}^t}\left[D\left(\mathbf{x}^t;\theta_d\right)\right] - \mathbb{E}_{\mathbf{x}^s}[D(G(E(\mathbf{x}^s;\theta_e);\theta_g);\theta_d)] \tag{11}$$

Note that the activation function of the last layer of the discriminator is removed.

6.1.2. Office-Home Real-World to Clipart

Unsupervised cross-domain image generation require a large volume of training data, especially natural images. Unfortunately, existing datasets for cross-domain task are always single category in one domain. Actually, multi-class transferring is a primary goal of visual domain adaptation [52,53], whose circumstance is similar to ours. So we abandon image-to-image translation datasets and try some standard datasets for vision-based domain adaptation.

We release the Office-Home [45] dataset for our task. As shown in Fig. 4. The Office-Home dataset consists of 4 domains, with each domain containing images from 65 categories of everyday objects and a total of around 15, 500 images. The domains include, Art: artistic depictions of objects in the form of sketches, paintings, ornamentation, etc.; Clipart: collection of clipart images; Product: images of objects without a background, akin to the Amazon category in Office dataset; Real-World: images of objects captured with a regular camera. Some troubles will occur when we use Office-Home to do image transformation task while the one is only for visual domain adaptation, such that Office-Home has numerous categories but each class has a small number of samples. This is a big impediment to GAN learning potential distribution of the data. So this dataset can be considered highly challenging.

The pictures in the dataset have various sizes, so we zoom all the images to the size of 32×32. Network structure and hyper parameters are consistent with which in SVHN-MNIST experiment. We do some visualization about attention mechanism in

Fig. 7. This is the final experimental results. We splited the results into two parts, (a), (b), (c) in SVHN to MNIST experiment and (d), (e), (f) in Real-World to Clipart. For (a), (b), (c): source images and corresponding generated samples are presented in leftcolumn and rightcolumn respectively. For (d), (e), (f): Source images were displayed in first and third columns and second and fourth columns were showing corresponding outputs. Here are detailed description: (a), (d): Our proposed model (Semantic invariant GAN). (b), (e): Primitive method(we used Domain Transfer Network in our experiments). (c), (f): Our model without attention loss, the purpose of joining this comparative experiment is to prove that attention mechanism does improves the performance of algorithm.

our network. Attention maps of network intermediate layer were output and upsampled to same size as original images. As the Fig. 6 shows, our network exactly learned the important part of images and pay more attention to align this area.

The experimental result shows our model has learned the shape of objects and generated particular textures on the object that makes samples look like target images. In addition, generated images hold fixed semantics with their source counterparts even if the appearance has been changed. See detailed results in Fig. 7.

6.2. Unsupervised domain adaptation

In this section, we study the performance of the SIGAN for unsupervised domain adaptation, where labeled data is available only in the source domain and no labeled data is available in the target domain. We compare the SIGAN with state-of-the-art domain adaptation methods: (i) Geodesic Flow Kernel (**GFK**) [54], (ii) Transfer Component Analysis (**TCA**) [26], (iii) Domain Adversarial Neural Network (**DANN**) [31] and (iv) Domain Transfer Network (**DTN**) [2]. For the shallow learning methods **GFK** and **TCA**, we extract deep features from the fc7 layer of the VGG-16 network that was pre-trained on the ImageNet dataset.

Results and discussion. The results are reported for the target classification in each of the transfer tasks in Table 1, where accuracies denote the percentage of correctly classified target data samples in the test split.

We use both training split of SVHN and MNIST to train our model. Then the percentage of correctly classified samples in test split of MNIST was calculated as accuracy. From the results we can see that the GAN based approach greatly surpasses the performance of the traditional shallow learning methods. In addition to the discriminant model, the generative model *DTN* also achieves higher recognition rates in the task. As an advanced version of DTN, our model gets accuracy of 83.41% and has significantly improved the accuracy over DTN. Furthermore, we compare the effect of different losses on the model performance. We found $\pounds_{LabelConst}$ and $\pounds_{GAN}$ was the most fatal part of the objective. They played a major role in the process of producing semantic

Table 1
Recognition accuracies (%) for domain adaptation experiments on the SVHN-MNIST.

Method	Accuracy
GFK	65.42%
TCA	62.71%
DANN	73.90%
DTN	80.23%
SIGAN without $£_{TV}$	83.63%
SIGAN	83.41%
SIGAN without $£_{Attention}$	78.56%
SIGAN without $£_{LabelConst}$	51.77%
SIGAN without $£_{GAN}$	32.60%
SIGAN without $£_{Attention}$ and $£_{LabelConst}$	10.39%
Source only	60.21%

invariant cross-domain images. $£_{Attention}$ can also correct the detail information of the image to a certain extent to improve the recognition rate. SIGAN without $£_{Attention}$ and $£_{LabelConst}$ can be seen as traditional GAN methods which do not model the properties of human visual perception. They failed in unsupervised visual domain adaptation. The last item in the table suggests that the optional loss $£_{TV}$, which is just for slightly smoothing the resulting images, does not affect the model performance.

From what has been discussed above, we can draw the conclusion that in the objective function $£_{LabelConst}$ plays a key role in solving the domain adaptation problem. This shows that the semantics of the generated image is indeed consistent with the original image. It s also critical to the adversarial loss, which makes the model learn the invariant representation between two domains through adversarial training.

7. Conclusion

In this paper, we first introduced the related work about image-to-image translation and stated that traditional image-to-image translation mothods merely consider the visual appearance properties, they failed to maintain the true semantics of an image during the transfer learning procedure from source to target domain. Then, we proposed a new GAN based model called SIGAN, which considered the image semantics, made generated samples both visually similar with target images and semantically consistent with their source counterparts. Our model achieved state-of-the-art results on several benchmark datasets. Based on a series of wonderful characteristic under unsupervised settings, our model can be extended to some more practical scenarios, such as unsupervised visual domain adaptation, data augmentation or image retrieval [55,56]. Although the performance of the algorithm has been greatly improved, the quality of the generated pictures is still not well guaranteed because of significant shortcomings the GAN has. In the future, more effort will be devoted to generate images with higher resolution and higher quality.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China: 61771155, 61672497, 61332016, 61620106009, 61650202 and U1636214, in part by National Basic Research Program of China (973 Program): 2015CB351802 and in part by Key Research Program of Frontier Sciences of CAS: QYZDJ-SSW-SYS013.

References

[1] H. Dong, P. Neekhara, C. Wu, Y. Guo, Unsupervised image-to-image translation with generative adversarial networks, CoRR (2017). http://arxiv.org/abs/1701.02676.
[2] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, CoRR (2016). http://arxiv.org/abs/1611.02200.
[3] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2849–2857.
[4] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 1857–1865.
[5] Y. Lu, Y.-W. Tai, C.-K. Tang, Conditional cycleGAN for attribute guided face image generation, CoRR (2017). http://arxiv.org/abs/1705.09966.
[6] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[7] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[8] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[9] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.
[10] G. Larsson, M. Maire, G. Shakhnarovich, Learning representations for automatic colorization, in: European Conference on Computer Vision, Springer, 2016, pp. 577–593.
[11] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2223–2232.
[12] M.-Y. Liu, O. Tuzel, Coupled generative adversarial networks, in: Advances in neural information processing systems, 2016, pp. 469–477.
[13] X. Mao, Q. Li, H. Xie, AlignGAN: Learning to align cross-domain images with conditional generative adversarial networks, CoRR (2017). http://arxiv.org/abs/1707.01400.
[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[15] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, Comput. Sci. (2015).
[16] X. Wang, A. Gupta, Generative Image Modeling Using Style and Structure Adversarial Networks, Springer International Publishing, 2016.
[17] S. Hartmann, M. Weinmann, R. Wessel, R. Klein, StreetGAN: Towards road network synthesis with generative adversarial networks, in: International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision, 2017.
[18] W.R. Tan, C.S. Chan, H.E. Aguirre, K. Tanaka, Learning a generative adversarial network for high resolution artwork synthesis, CoRR (2017). http://arxiv.org/abs/1708.09533.
[19] H. Zhang, V.M. Patel, B.S. Riggan, S. Hu, Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces, CoRR (2017). http://arxiv.org/abs/1708.02681.
[20] M. Mirza, S. Osindero, Conditional generative adversarial nets, CoRR (2014). http://arxiv.org/abs/1411.1784.
[21] A. Odena, Semi-supervised learning with generative adversarial networks, CoRR (2016). http://arxiv.org/abs/1606.01583.
[22] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, 2017, p. 7.
[23] R. Rosales, K. Achan, B.J. Frey, Unsupervised image translation., in: iccv, 2003, pp. 472–478.
[24] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: Aligning domains using generative adversarial networks, CoRR (2017). http://arxiv.org/abs/1704.01705.
[25] J. Huang, A. Gretton, K.M. Borgwardt, B. Schölkopf, A.J. Smola, Correcting sample selection bias by unlabeled data, in: Advances in neural information processing systems, 2007, pp. 601–608.
[26] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Netw. 22 (2) (2011) 199–210.
[27] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: An unsupervised approach, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 999–1006.
[28] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.
[29] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (Mar) (2012) 723–773.
[30] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in: Computer Vision–ECCV 2016 Workshops, Springer, 2016, pp. 443–450.
[31] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, Domain-adversarial neural networks, stat 1050 (2014) 15.
[32] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, stat 1050 (2017) 17.
[33] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, CoRR (2016). http://arxiv.org/abs/1612.03928.

[34] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation, in: European Conference on Computer Vision, Springer, 2016, pp. 597–613.
[35] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, in: Eleventh International Conference of the Center for Nonlinear Studies on Experimental Mathematics : Computational Issues in Nonlinear Science: Computational Issues in Nonlinear Science, 1992, pp. 259–268.
[36] J. Li, L.-y. Duan, X. Chen, T. Huang, Y. Tian, Finding the secret of image saliency in the frequency domain, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2428–2440, doi:10.1109/TPAMI.2015.2424870.
[37] J. Li, Y. Tian, X. Chen, T. Huang, Measuring visual surprise jointly from intrinsic and extrinsic contexts for image saliency estimation, Int. J. Comput. Vis. 120 (1) (2016) 44–60, doi:10.1007/s11263-016-0892-7.
[38] J. Li, Y. Tian, T. Huang, Visual saliency with statistical priors, Int. J. Comput. Vis. 107 (3) (2014) 239–253, doi:10.1007/s11263-013-0678-0.
[39] W. Yin, H. Schütze, B. Xiang, B. Zhou, Abcnn: Attention-based convolutional neural network for modeling sentence pairs, Trans. Assoc. Comput. Linguist. 4 (2016) 259–272.
[40] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, In ICLR Workshop (2014).
[41] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR (2014). http://arxiv.org/abs/1409.0473.
[42] F. Yang, J. Li, S. Wei, Q. Zheng, T. Liu, Y. Zhao, Two-stream attentive CNNs for image retrieval, in: Proceedings of the 2017 ACM on Multimedia Conference, in: MM '17, New York, NY, USA, 2017, pp. 1513–1521.
[43] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS workshop on deep learning and unsupervised feature learning, volume 2011, 2011, p. 5.
[44] Y. LeCun, The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/ (1998).
[45] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in: (IEEE) Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[46] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, CoRR (2016). http://arxiv.org/abs/1603.04467.
[47] T. Tieleman, G. Hinton, Rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, Technical Report, Technical report, 2012. 31
[48] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, stat 1050 (2017) 9.
[49] X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2813–2821.
[50] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.
[51] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5769–5779.
[52] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International Conference on Machine Learning, 2015, pp. 1180–1189.
[53] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, in: Advances in Neural Information Processing Systems, 2016, pp. 343–351.
[54] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Computer Vision and Pattern Recognition, 2015, pp. 2066–2073.
[55] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with cnn visual features: A new baseline, IEEE Trans. Cybern. 47 (2) (2017) 449–460.
[56] S. Wei, D. Xu, X. Li, Y. Zhao, Joint optimization toward effective and efficient image search, IEEE Trans. Cybern. 43 (6) (2013) 2216–2227.

Xiaofeng Mao received the B.S. degree in Internet of things Engineering from Northeastern University at Qinhuangdao. He is currently studying for a master's degree at the Harbin Engineering University. His research interests include image analysis (visual domain adaptation, object tracking) and machine learning.

Shuhui Wang received the B.S. degree in electronics engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval and large-scale Web multimedia data mining.

Liying Zheng received her Ph.D. degree in control theory and control engineering from Harbin Engineering University (HEU), China, in 2003. She has been a Professor in HEU since 2013. Her current research interests include image analysis and machine learning.

Dr. Qingming Huang is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia computing, image processing, computer vision and pattern recognition He has published more than 300 academic papers in prestigious international journals including IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on CSVT, etc, and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, VLDB, etc. He is the associate editor of IEEE Trans. on CSVT and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, PSIVT, etc.