Freestyle Object Localization

Jia Liu ¹

¹Nanjing University of Science and Technology

December 7, 2023

Abstract

In this paper, we define the freestyle object localization problem where a model is expected to autonomically learn and locate an arbitrarily specified object given only the name. However, learning the semantic features of an object requires corresponding training data which are not always available in existing datasets. Obviously, Internet provides massive off-the-rack data for learning given only the keyword while few existing methods are able to directly use them. To directly use them, a localization model should firstly overcome two difficulties: (1) learning from single-class annotated images and (2) from images with unpredictable sizes. For single-class annotation, we establish a novel probabilistic model to capture the distribution of the object in the background data space. Based on a deep hierarchical architecture with alterable depth that is designed for unpredictable sizes, we propose an association module and design an energy function to drive the probabilistic model. By maximizing the log-likelihood, the proposed model is able to highlight the semantic features of object that has the highest probability in training images, i.e., the specified object. In experiments, the proposed model achieves state-of-the-art performance among weakly-supervised localization methods on public datasets and freestyle localization of objects that are rare in public datasets.
Freestyle Object Localization

Jia Liu, Donghui Li, Wenhua Zhang, Fang Liu, Zhen Cui, Maoguo Gong, Fellow, IEEE, and Licheng Jiao, Fellow, IEEE

Abstract—In this paper, we define the freestyle object localization problem where a model is expected to autonomically learn and locate an arbitrarily specified object given only the name. However, learning the semantic features of an object requires corresponding training data which are not always available in existing datasets. Obviously, Internet provides massive off-the-rack data for learning given only the keyword while few existing methods are able to directly use them. To directly use them, a localization model should firstly overcome two difficulties: (1) learning from single-class annotated images and (2) from images with unpredictable sizes. For single-class annotation, we establish a novel probabilistic model to capture the distribution of the object in the background data space. Based on a deep hierarchical architecture with alterable depth that is designed for unpredictable sizes, we propose an association module and design an energy function to drive the probabilistic model. By maximizing the log-likelihood, the proposed model is able to highlight the semantic features of object that has the highest probability in training images, i.e., the specified object. In experiments, the proposed model achieves state-of-the-art performance among weakly-supervised localization methods on public datasets and freestyle localization of objects that are rare in public datasets.

Index Terms—Object localization, freestyle learning, deep neural networks, probabilistic model.

1 INTRODUCTION

O BJECT localization aims to locate a specified object from each given image scene [1]. It is a hot topic in computer vision community and widely applied. To recognize and locate the corresponding object, fully-supervised models such as RCNN series [2], [3], YOLO series [4], [5], SSD series [6], [7], and anchor-free based ones [8], [9] require massive fully annotated data (boxes and classes) to learn the semantic features of the object.

Above methods can achieve high detection and localization accuracy in real applications but rely on finely annotated datasets where the location, size, and classes of objects should be manually annotated in each training image. For example, the widely used VOC2007 dataset [10] contains 9963 images and there are 24640 annotated objects within 20 classes. Preparing this dataset including collecting images, pre-processing, and annotating requires a mass of human efforts and time costs. Even though existing datasets have provided significant convenience for many applications, with the rapid update of information, they cannot satisfy the requirement of many new objects and scenarios.

As a consequence, researchers have tried many methods to reduce the efforts of human in annotating the training data. Recently weakly-supervised object localization has attracted much attention. It aims to train a model via image-level annotated images [1] where only the class of the objects in the training images is required to annotate. Multiple-instance learning (MIL) [11], [12], [13] is popular to deal with such a problem. To construct end-to-end deep learning model, Bolei Zhou et al. proposed the class activation mapping (CAM) [14] which locates the corresponding objects with class activation maps based on a discrimination model. Due to its concise learning process, many weakly-supervised methods follow such a framework [15], [16], [17], [18], [19]. Moreover, many methods use multiple branches [20], [21] or training multiple networks [22], [23] to sufficiently use limited annotated information. With the limited supervision information, weakly-supervised methods usually achieve less accuracy but can satisfy the requirement in some applications. Weakly-supervised methods significantly free human from dataset preparation but still need human to prepare image-level annotated dataset with various classes of objects. Unsupervised methods avoid the annotation of data such as co-segmentation [24]. It segments the common object in a set of images. But without annotation, the semantic class of the object is difficult to be learned and applied in a new test scene.

One of the goals in artificial intelligence is to reduce the efforts of human to the maximum. So in this paper, we aim to achieve the freestyle goal that the localization model is able to locate an arbitrarily specified object given only the object name. But the localization model requires data to learn semantic features. Undoubtedly, Internet provides sufficient raw training samples and searching engines are convenient tools to collect images of corresponding objects. However, different from pre-processed images in existing datasets, as in Fig. 1 which shows examples of images in VOC2007 dataset and searched images given an object name (“tiger” for instance), there are two most critical challenges to learn from the raw searched images: (1) In VOC2007 as shown in Fig. 1, most objects in images are annotated by human and the learning model is able to learn the discriminative information for recognition. While for searched images, given the name of the object, the only annotation information is that most of searched images contain the object. But
We define the freestyle object localization problem where contributions are summarised as follows.

To overcome the two challenges as discussed above, we propose a novel architecture with alterable depth and an association module. The architecture can adapt to the unpredictable sizes and the association module is used to highlight the semantic features via learning.

Following the architecture, we design an energy function and derive a novel probabilistic model to learn the distribution of the object for memory. We overcome the difficulty of distribution estimation in probabilistic models via a ingenious trick. The optimization and localization process are also derived for learning and association.

On public datasets, the proposed method achieves equivalent performance to recent weakly-supervised localization methods. Moreover, we provide several objects’ name that are difficult to find in public datasets to demonstrate the freestyle localization capability.

The innovative problem and learning model open up new opportunities for using massive raw images in Internet for specific tasks and may inspire new solutions for large image models.

The rest of this paper is organized as follows. Motivation, related work, and probabilistic models are introduced in Section II. The detailed architecture, modeling, and localization are described in Section III. Experiments on both public and searched images are implemented in Section IV. Finally we summarise the paper and discuss the future work in Section V.

2 Motivation and Preliminaries

In this paper, we consider a new problem that is more significant but more challenging than existing localization problems. Weakly-supervised object localization problem is a recently popular and challenging problem which aims to reduce the efforts of human. Here we first discuss the mechanisms of weakly-supervised object localization methods and then pose the challenges of the proposed problem.

2.1 Weakly-Supervised Object Localization

Due to they use less annotation information which significantly reduce the efforts of human, weakly-supervised object localization and detection methods have been widely researched. Among them, MIL [12] and CAM [14] based ones are two types of typical frameworks. MIL based ones take each training image as a bag that contains the multiple instances, i.e., the objects and background. MIL models are composed of two modules including instance selector and detector. Instance selector uses detector to compute the score of each instance and selects a highest one. Then the instance detector is trained via selected instances. The two modules are optimized alternately to generate the optimal detector.

CAM based ones train a deep discriminative model and the corresponding object is localized via the map from the output class information. To classify different objects, the key features of each object is learned and extracted in a
deep architecture. Then via class activation mapping, the corresponding objects will be highlighted in the activated maps.

Even though weakly-supervised object localization methods require much less annotation information, existing ones still need annotation of multiple classes in the images which are difficult to achieve freestyle learning. Given the object name, a large volume of corresponding images can be obtained from the Internet via a searching engine. Such data can be used for learning to achieve freestyle localization. But two challenges should be first overcome which are not considered by existing localization methods.

### 2.2 Challenges

As discussed in Introduction, the raw searched images in Internet are different from those in datasets from two major aspects.

#### 2.2.1 Single-Class Annotated Images

The annotation information in prepared datasets can be guaranteed for existing learners to learn semantic information. For example, fully supervised localization or detection methods require the annotations of boxes and classes of corresponding objects. Weakly-supervised ones require annotations of multiple classes to learn the discriminative information for localization. While given an object name, the only annotation information in the searched images is that most of them contains such an object. Both fully supervised and weakly supervised object localization methods are difficult to learn with the limited annotation information.

From the methodology perspective, learning from single-class annotated data has been researched for decades which is called one-class classification (OCC) [33]. There are various solutions to learn from the single-class annotated data. Statistical methods such as one class SVM (OCSVM) [34], one class convolutional neural network (CNN) [35], and deep SVDD [36] aim to find a hyper-plane that encircles the training data. Encoder-decoder based ones [37] distinguish out of distributed ones via the reconstruction error. Generative adversarial network (GAN) can also be designed to deal with OCC [38] and end-to-end network can be trained for discrimination [39]. Recently, some new learning paradigms are applied to OCC such as the knowledge distillation [40] where the variance between teacher and student networks can be used for discrimination. However, most OCC methods are designed for distinguishing positive and negative samples via learning from positive training data. As shown in Fig. 1 (b), there may be positive and negative objects in one image. Moreover, it is also a great challenge to derive the detection method to distinguish positive and negative objects in one image.

The recently popular continual learning [41], zero-shot learning [42], or knowledge distillation [43] may be employed to solve such a problem by taking the single-class annotated images as those of a incremental class or unknown class. But trained with regularized data, for raw data of a new class, those methods are faced with another challenge of the unpredictable image sizes.

#### 2.2.2 Unpredictable Image Sizes

To adapt to the fixed architecture of many learning models, the input images are usually resized into the same size by downsampling, upsampling, cropping, and padding. But such pre-processing operators may generate distorted objects. To deal with such a problem, spatial pyramid pooling [44] and global pooling [45] were proposed by fixing the output size of pooling layers. They can deal with small size variance. There are also networks that can directly deal with the unpredictable image size such as the fully convolutional net [46] which abandons the dense connections. But it is mainly designed for image segmentation or transformation that requires the same input and output sizes.

Possible solutions may be inspired from models for sequence data due to their unpredictable length. Recurrent neural networks [47] can naturally deal with the unpredictable length with adaptive depth to the sequence length. Transformer [48] uses point-wise feed forward layer to deal with each step of a sequence independently.

Inspired from the above architectures, we propose an architecture with alterable depth which is able to adapt to the input image size. Then with such an architecture, a probabilistic model is designed to capture the distribution of the searched object.

### 2.3 Probabilistic Models

In this paper, the distribution of searched object is learned and associated via a probabilistic model. Probabilistic models play important roles in the development of deep learning [49] among which restricted Boltzmann machine (RBM) [50] is a popular one. Probabilistic models learn to capture the distribution of input data $x$ via modeling the parameterized probability of it:

$$P(x; \theta) = \frac{\exp[-E(x, \theta)]}{\sum_{x'} \exp[-E(x', \theta)]}$$

where $\theta$ is the parameter set of the model, for instance, connecting weights and biases in neural networks. $E(x, \theta)$ is the energy function that drives the probabilistic model. It forms an energy field in the data space of $x$ and so $x'$ is one of all possible data in the space which is called fantasy data in [51]. The denominator is a partition function that guarantees the probability sum over the data space to be 1. The model is optimized by maximizing the log-likelihood of it which means to assign more probability to training data $x$ and less probability to all the other data $x'$. Then the energy field can well follow the distribution of training data. A toy example is shown in Fig. 2 where the relationship among training data $x$, fantasy data $x'$ in a two-dimensional data space, and the probabilistic model is illustrated. With randomly initialized parameters, i.e., the $\theta$, sampled data from the probability $P(x; \theta)$ seem irregular. The aim of optimization is to capture the distribution of training data and the sampled data will follow this distribution. The training of probabilistic models can be unsupervised where only data are available. By designing different energy functions based on different architectures, probabilistic model can model different tasks, such as image classification [31], inpainting [32], image fusion [52], sketch synthesis [53], and even change detection [54], [55]. As a consequence, in this
paper, we try to model the distribution of searched objects via a probabilistic model for localization.

However, it is a great challenge to optimize the probabilistic model which impedes its wide application. To maximize the log-likelihood, usually the gradient is first derived:

$$\frac{\partial \log P(x; \theta)}{\partial \theta} = \frac{\partial [-E(x, \theta)]}{\partial \theta} - P(x'; \theta) \frac{\partial [-E(x', \theta)]}{\partial \theta}$$

(2)

where the first term is the derivative of the energy given training data $x$ which can be easily derived. The second term is the gradient expectation over all $x'$, i.e., the whole data space which is a great challenge to estimate. In RBM, the gradient expectation is estimated by numerous Gibbs sampling. With the single layer network of RBM, the status probability of visible and hidden unites can be derived from the probability $P(x; \theta)$. However, in this paper, the complexity of searched images and the hierarchical backbone we use pose great challenges to estimate the whole data space.

With the above discussed challenges and difficulty in using probabilistic models, a new architecture and a new learning paradigm is required to achieve the freestyle object localization. Then the whole story is described in detail as follows.

3 Probabilistic Learning and Object Association

To achieve freestyle object localization in virtue of Internet, as discussed above, we have confirmed the preliminary scheme with an architecture that adapts to unpredictable image sizes and a probabilistic model that captures the distribution of corresponding object. In this section, we further detail the learning and association schemes of the new learning paradigm.

3.1 Architecture

The story begins from the architecture which is the basic of many learning models. With the single-class annotation, it is difficult to use end-to-end architectures that learn mapping from image to label. But hierarchical architectures such as CNN provide excellent learnable feature extractors. With hierarchical convolution and pooling layers, CNN is able to extract the main contents. So we first use a CNN to construct the backbone of our network. To achieve the goal of association and establish the probabilistic model, we remove fully connecting layers in CNN and propose an association module with the input of feature maps in CNN. The energy function and probabilistic model is defined based on the association module.

As a whole, we construct the network architecture with a backbone and an association module. Then we consider the unpredictable sizes of input images. Without fully connecting layer, convolutional and pooling layers in backbone can adapt to any image sizes. We also define the association module as several convolutional layers and it is compatible to size of searched images. But the required depth to extract abstract features is different with different input sizes and the number of channels of output feature maps is different with different depth. Then a channel regularization layer is used between backbone and association module to regularize the number of channels via channel-wise pooling or upsampling.

As a consequence, the network architecture is designed as shown in Fig. 3 with a backbone (bluish blocks), an association module (yellow blocks), and the bond between them, i.e., the channel regularization layer (purple layer). The depth of the backbone is adaptive to the size of input images as shown in Fig. 3. For large images, more layers are necessary to extract the abstract features. While for small images, less layers are enough. For wide images, the output feature maps may be also wide ($3 \times 2$ pixels) and for high images, the output feature maps keep the similar aspect ratio ($2 \times 3$ pixels). The input image propagates through the backbone network until the size of feature maps is smaller than the size of convolutional kernel in the next layer (for instance, $2 \times 3$ or $3 \times 2$ pixels in Fig. 3). The feature maps are then fed into the channel regularization layer followed by the association module.

With an input image $I$, we formulate the architecture as: $F^L(I) = f(I, \theta_B^{(1-L)}, \theta_A)$. $F^L$ denotes the output of the association module and $L$ denotes the alterable depth of the backbone network. $\theta_B^{(1-L)}$ is the parameter set of the layers from 1 to $L$ in the backbone network. $\theta_A$ is the parameter set of the association module which are identical with different depth of the backbone. The proposed learning model is established based on the $F^L$.

There are other obstacles and note the architecture will be further derived as the story goes on.

3.2 Probabilistic Model and Energy Function

The challenge of unpredictable size is easily solved via the network architecture. Here the story goes well as planned. Next is to establish a probabilistic model in order to capture the distribution of searched object. Following the probabilistic model, we first define the probability of input image $I$:

$$P(I; \theta_B^{(1-L)}, \theta_A) = \frac{\exp[-E(F^L(I))]}{\sum_I \exp[-E(F^L(I))]}$$

(3)

where the numerator is the exponential energy of $F^L(I)$ and the denominator is the sum of that over all the possible data. Here $I'$ is similar to $x'$ in Eq. (1) which is one of the image in the whole space. For example, if $I$ is a $2 \times 2$ binary image, $I'$ is one of the data in the set $\{[0, 0, 0, 0], [0, 0, 0, 1], [0, 0, 1, 0], [0, 1, 0, 0], [1, 0, 0, 0], [1, 0, 0, 1], [1, 0, 1, 0], [1, 1, 0, 0], [1, 1, 0, 1], [1, 1, 1, 0], [1, 1, 1, 1]\}$. The denominator is the exponential energy sum over all the $I'$s in the set.
With the probabilistic model, it is critical to define the energy function. Since we aim to highlight the corresponding object, we directly use the average over all components in the output maps as the energy:

\[ E(L(I)) = -\sum_{(i,j,k)} P(L(I), (i,j,k))/N \tag{4} \]

where \( N \) is the number of components. With the hierarchical architecture and this energy function, the output can be taken as the response of a feature extractor to corresponding objects. Optimizing this model aims to increase the output of training images while decrease those of all the other data. Then the features of main contents in the training images can be learned and highlighted.

To optimize the model, following Eq. (2), we first derive the gradient of the log-likelihood:

\[ \Delta \theta = \frac{\partial E_L(I)}{\partial \theta} - \sum_{I'} P(I'; \theta) \frac{\partial E_L(I')}{\partial \theta} \tag{5} \]

where \( \theta \) denotes the parameter set: \( \theta = \{\theta^{1-L}_B, \theta_A\} \). The first term is easy to compute via back-propagation. However, on large images which results in almost an infinite data space, it is difficult to estimate the distribution of fantasy data via a hierarchical network, i.e., the \( P(I'; \theta) \) in the second term. Then the story reaches an impasse.

### 3.3 Executable Learning Model

There is always light at the end of the tunnel. There are many redundant and irrelevant data in the whole data space such as random noise and impossible scenes that the object belongs to. So we return to the searched images as shown in Fig. 1 and find that searched images usually contain not only the searched object itself, tiger for instance, but also background objects and stuffs, such as trees, grasses, and even other animals. With various online sources of the searched images, such a phenomenon is common in most searched objects as examples shown in Fig. 4. Meanwhile, the aim of object localization is to distinguish the corresponding object from the other objects or stuffs. It is not necessary to model the object in the whole data space and the background information in the searched images can be used to learn the discriminative features between the specific object and background.

Some examples are illustrated in Fig. 4. With an arbitrary keyword, most of the searched images contain corresponding object while the background is not always the same. As a consequence, we can directly use the training images to model the distribution of the object in the visible scenes instead of in the whole data space. By maximizing the log-likelihood, the features of object that has the largest probability can be legitimately learned. With such a mechanism, even abstract objects can be learned such as a specific expression. For example, as shown in Fig. 4, a specific expression can be made in different faces which lets the model learn the abstract expression instead of the tangible faces.

But there is another obstruction in this way: how to use the background information? The top abstract features of CNN contain the main abstract contents of the input image. Then the response of feature extractors to all the contents in images can be found in lower layers. Compared with abstract features in top layer where redundant information is removed, the feature maps in lower layers contain more information about various objects because the filters are applied more locally. Consequently, to use the information of whole scene in the image, we equip the association module for each layer in backbone network as shown in Fig. 5 and establish the final executable learning model:

\[ P_I(I; \theta_B, \theta_A) = \frac{\exp[-E(F^l(I))]}{\sum_I \exp[-E(F^l(I))]} \tag{6} \]

where the denominator denotes the exponential energy sum over the feature maps of all the hierarchical layers in backbone network with the input of \( I \).

The computation architecture of the denominator is shown in Fig. 5 where the channel regularizer layer and association module are applied to each layer of the backbone to compute \( F^l(I) \) with \( l = 1, 2, \ldots, L \) (yellow feature maps in each layer). Here the parameters in association modules are shared for each layer in order to capture the discriminative information between object and background. By maximizing the log-likelihood, the network is expected to extract the features of the object that has the highest probability in the visible scenes.

### 3.4 Feasibility Analysis

There are so many objects in each searched image and why the network is able to learn and highlight the features of
Channel regularization and association module highlighted with the above instanced network.

If there is an tiger in input image, the output can be fully output because many redundant information, i.e., restrained those irrelevant objects have much less influence on the value in lower pixel values in feature maps. While in top layer, objects such as trees are responded negatively which results may extract detail features such as colors of orange and extract features of an object such as tiger, the lower layers value in lower layers. For example, if the CNN is learned to denote the response of feature extractors to various objects. The output of the lower layers are maps and they can be integrated into the model. The channel regularization layer and association module (shared parameter) are equipped in each layer of the backbone network. Then the information of all objects in training images can be integrated into the model. Then we discuss the learning model. Maximizing the denominator of Eq. (6). The channel regularization layer and association module (shared parameter) are equipped in each layer of the backbone network. Then the information of all objects in training images can be integrated into the model.

Fig. 4. Examples of some searched images. With enormous images sources in Internet, most of the searched images contains the searched object but the background objects vary a lot. Such a phenomenon can be used to model the distribution of the searched object within the visible scene.

Keyword: “tiger”

Objects: tiger, phacochoerus, bough, root pile, ...
Objects: tiger, snowfield, grove, twig, ...
Objects: tiger, lake, canopy, jungle, ...
Objects: B-2, person, hangar, ladder, ...
Objects: B-2, building, wood, lake, ...
Objects: B-2, runway, vehicle, grassland, ...

Fig. 5. Architecture to model the whole scene. It is used to compute the denominator of Eq. (6). The channel regularization layer and association module (shared parameter) are equipped in each layer of the backbone network. Then the information of all objects in training images can be integrated into the model.

searched object, e.g., the tiger, by optimizing Eq. (6)?

First about the architecture, as visualized features in CNN [56], [57], lower layers extract local detail features such as colors, lines, edges, and et al. Higher layers combine the features in lower layers and the features in them represent one or more objects in input image. The output of association module can be taken as the response of feature extractors. The output of the lower layers are maps and they can denote the response of feature extractors to various objects. All objects in an input image are able to influence the output value in lower layers. For example, if the CNN is learned to extract features of an object such as tiger, the lower layers may extract detail features such as colors of orange and black. Then in the response maps of lower layers, only the pixels in the position of tigers can be highlighted. Other objects such as trees are responded negatively which results in low pixel values in feature maps. While in top layer, those irrelevant objects have much less influence on the output because many redundant information, i.e., restrained features in lower layers, are removed via the pooling layers. If there is an tiger in input image, the output can be fully highlighted with the above instance network.

Then we discuss the learning model. Maximizing the log-likelihood of the probability model in Eq. (6) amounts to increasing the numerator while decreasing the denominator. There are many objects in an image and the features of them are also different. Since we search images using the name of the object, e.g., the tiger, it can be guaranteed that most images contain the tiger. By maximizing the log-likelihood, the network can be learned to extract and highlight the features of only tigers. If not, the log-likelihood cannot be maximized as discussed in the following two cases.

(1) Since the energy is defined as the output average value of association module, if features of other objects instead of tigers are learned and highlighted, e.g. trees, then the values of output in top layer is significantly reduced for many training images because many images do not contain trees. While the local detail features of trees such as lines, edges, and so on can also be found in other objects. This will not significantly reduce the output values in lower layers. Then the log-likelihood cannot be maximized for all training images.

(2) If the features of not only tigers but also other objects are learned and highlighted, then for all the training images, the numerator can achieve high values. But the denominator is computed by the average of output over all positions in the feature maps of all layers. Since there are also some special detail features of the objects in hierarchical layers, such as the different colors and textures of tigers and trees, the denominator will be larger than the case where only features of tigers are learned due to the local property of lower layers as the above discussion about the architecture. Then the log-likelihood is smaller in this case.

Only when the network is tuned to extract and highlight the features of only tigers, the numerator can achieve high values on most training images and the value of denominator is as small as possible which is the optimal case of the model. We exhibit the output maps of the association module of different layers in Fig. 6. They are computed via both randomly initialized and trained network parameters. We use the searched tiger images to train the network. Before training, the output maps show highlights of dif-
different parts of the input images. After training, almost all
the background objects are restrained in the output maps
and the textures of the tiger are highlighted. Lower layers
highlight the detail information and higher layers highlight
the holistic object. Follows are optimization and object lo-
calization that are easy to be derived.

3.5 Optimization, Localization, and Limitations

With the accessible probabilistic model in Eq. (6), the in-
calculable gradient in Eq. (5) can be modified as:

$$
\Delta \theta = \frac{\partial F^l(I)}{\partial \theta} - \sum_l P_l(F^l(I); \theta) \frac{\partial F^l(I)}{\partial \theta} \tag{7}
$$

where $P_l(F^l(I); \theta) = \frac{\exp[-E(F^l(I))]}{\sum \exp[-E(F^l(I))]}$ as derived from
Eq. (6). Then all the terms are computable with back-
propagation.

For convenience, we update the parameters separately
according to the two terms. Given a training image $I$, in Eq.
(7), the first term is computed following the architecture in
Fig. 3. The network parameter set $\theta$ is updated via back-
propagation. In the second term, the association module is
applied to all layers as in Eq. (6) and Fig. 5. The output
of the association module will be minimized in Eq. (7)
and the negative gradient of output layer is computed and
back-propagated through the network. Even though there
are no reference output, both the two terms follow the
basic back-propagation algorithm which include forward,
backward, and updating processes. So the whole model can
be conveniently optimized. The whole training process is
summarised in Algorithm 1. The training is straightforward
in spite of the intricate derivations above.

After optimization, the association module is able to
generate highlighted signal when the input image contains
the corresponding object. It is also critical to recognize which
part of the image contributes most to the highlighted signal
in order to locate the object. Evidences can be found in
the back-propagated gradients. Given a test image $I^t$, the
gradient of energy function is first computed:

$$
\Delta I^t = \sum_l P_l(F^l(I^t); \theta) \frac{\partial E(F^l(I^t))}{\partial I^t} \tag{8}
$$

In $\Delta I^t$, the value in each pixel denotes the influence of
the corresponding pixel in $I^t$ to the output of association mod-
ule. Then a heat map can be generated via: $M(i, j) = (\Delta I^t(i, j))^2$ where $(i, j)$ is the position of each pixel, to
highlight the region that actives the association module, i.e.,
the corresponding object. By boxing the heat map as in [14],
the location of corresponding object can be obtained. The
detection process in summarised in Algorithm 2.

However, neural network is a black box which is difficult
to interpret. So we can only guarantee that prominent and
common features of searched objects can be highlighted as
discussed above. For example, the tigers usually have the
prominent feature of unique texture. So the proposed learn-
ning model is able to highlight the region that contains the
unique texture. But for some other animals such as birds that
are with various textures, shapes, and so on, it is difficult
to learn a prominent and common feature. Meanwhile, the
whole objects may not be completely detected. For example,
maybe only the body of the tiger where the texture is obvi-
ous can be highlighted. Moreover, due to the uncertain data
volume, it may also suffer from overfitting and robustness
problems as many learning methods. So in the following
experiments, we evaluate the proposed learning model by
using both public datasets and raw searched images.

4 EXPERIMENTAL STUDY

To the best of our knowledge, none of existing object local-
ization or detection methods are able to achieve freestyle
object localization. But without any comparison, it is diffi-
cult to verify the feasibility of the new learning model and
evaluate its performance in real cases. So we conduct the


Algorithm 1 Training process of the proposed model.

**Input:**
Object name.

**Initialization:**
Search and download images from a searching engine using the object name as keyword.
Randomly initialize network parameters $\theta = \{\theta_A, \theta_B\}$.

**Repeat:**

- **for each image:**
  - Propagate forward with the architecture in Fig. 3.
  - Compute $\Delta \theta = \partial F^L(I)/\partial \theta$ in Eq. (7).
  - Update $\theta = \theta + \alpha \Delta \theta$ with the learning rate $\alpha$.

- **for each image:**
  - Propagate forward with the architecture in Fig. 5.
  - Compute $\Delta \theta = \sum P_l(I^l; \theta) \partial F^l(I)/\partial \theta$ in Eq. (7).
  - Update $\theta = \theta - \alpha \Delta \theta$.

**Output:**
The trained network.

Algorithm 2 Object detection process of the proposed model.

**Input:**
A test image.
The trained network.

**Detection:**

- Propagate forward with the architecture in Fig. 5.
- Compute $\Delta I^l = \sum P_l(I^l; \theta) \partial E(F^l(I^l))/\partial I^l$ in Eq. (8).
- Generate heat map $M(i, j) = (\Delta I^l(i, j))^2$ and corresponding box.

**Output:**
The box.

Experiments in two folds including feasibility verification by using public datasets with comparison to state-of-the-art localization methods and freestyle object localization by providing names of objects that are not available in those public datasets.

In the first fold of experiments, we use test sets in three public datasets, the PASCAL VOC 2007, VOC 2012 [10] and MS COCO 2017, to evaluate the proposed method quantitatively by comparing with existing state-of-the-art methods. The three datasets are widely used in evaluating object localization and detection methods. For VOC datasets, there are 20 classes in them and the proposed learning model is trained respectively via the training sets in two datasets (proposed/VOC2007 and proposed/VOC2012) and searched images using the names of 20 classes from Baidu Image (proposed/Baidu). For proposed/Baidu, it is supposed that no training data are available where the test sets in VOC datasets are only used for evaluation. Fig. 7 shows some images in the VOC 2007 dataset and searched images from Baidu Image. It can be found that images in VOC 2007 are with similar sizes while that of the searched images is varied. Moreover, the proportion of width and height is also different (from squares to strips). Another difference is that images in VOC datasets usually contain two or more objects while there is mainly the specific object in searched images which can help the model to learn semantic information of the corresponding objects. Note that the superiority of the proposed method is not to improve localization performance but to achieve the freestyle object localization which is a great challenge for existing methods. So we conduct the second fold of experiments.

In the second fold of experiments, to demonstrate the freestyle localization capability, we impromptu provide 10 names including Tiger, Panda, Trump, Yun Ma, Iron Man, WALL-E, B-2 Stealth and Strategic Bomber, aircraft carrier (AC), happy animals, and emoji of Jacky Cheung. They belong to different types of objects or persons, including animals, famous persons, well-known movie roles, weapons, and even abstract expressions. They are infrequent in public datasets. We type the Chinese name of them in Baidu Image and obtain the searched images. They are sorted via relevance and the first 400 images are downloaded via a browser helper object named ImageAssistant3. For each object, 100 images are randomly selected as testing images and the rest images are used for training. The 10 types of images are with various sizes, complex backgrounds, and without any annotation information except the name. Freestyle object localization makes it possible to free human from data annotation and preparation. Note that the searched images are public online provided by Baidu (the searched persons are both public figures and the photos are circulating all over the world) without conflict of right. The result images are only used to exhibit the performance without any other purposes.

4.1 Implementation Details
Since the training paradigm of the proposed model is different from general end-to-end architectures, we directly use CUDNN3 with CUDA 12.1 and Visual Studio 2019 (Windows 11 operating system with GPU of NVIDIA RTX 4090) to construct the forward, backward, and updating processes. Since we propose a new learning paradigm instead of complex architectures, an ordinary CNN is used as the backbone network. The CNN is composed of hierarchical

---

1. https://cocodataset.org/
2. https://image.baidu.com/
3. https://www.pullywood.com/ImageAssistant/
blocks and each block is composed of two convolutional layers with 5 × 5 convolutional kernels without padding and one pooling layer with 2 × 2 average pooling. The numbers of channels in the blocks are 3 – 32 – 64 – 64 – 128 – 128 – 256 – 256 – 256 – 512 – 512 – 512 – 512 – 512. The number of layers is not fixed with diverse image sizes. We set 128 channels in the regularization layer where the feature maps are regularized via channel-wise average pooling or nearest neighbor upsampling. There are 4 convolutional layers in the association module with the first layer generated via 3 × 3 convolutional kernels. Then the other layers are generated via 1 × 1 convolutional kernels. The channel numbers are 128 – 128 – 64 – 64. ReLu activation function is used except the output layer of association module (Sigmoid). We use basic back-propagation algorithm to optimize the model and the learning rate is set as 0.00001. In addition to the architecture and learning rate, there are not other hyper-parameters as described in Section III. The code is provided in GitHub5.

Weakly-supervised methods use image-level annotated data for training which require the least annotation information among existing methods. So, on public datasets, we compare the proposed model with some classical and recent weakly-supervised object localization and detection methods, including WSDNN [58], DT+PL [59], ZLDN [22], ML-LocNet [60], WCCN [61], Ts2C [17], WSOD [62], PG-PS [18], MIST [63], CASD [64], C-MIL [13], and CPE [65].

4.2 Feasibility Verification
We evaluate the performance of proposed method and compared methods with average precision (AP) of each class and the mean AP (mAP) over all the classes. The evaluation results on the test sets of VOC datasets are listed in Tables 1 and 2, respectively. Weakly-supervised methods use image-level labels to train the learning models and locate the position of corresponding objects. With the influence of various objects in background given only the image-level labels, it is a great challenge to correctly locate the objects. Among the compared methods, CAM based ones such as WCCN, TS2C, WSOD, and PG-PS are popular recently. CAM uses classification information to generate a heat map to highlight the objects that contribute most to the classification. The background information influences the result of CAM a lot. For example, objects of vehicle (e.g., bus and car) are within similar backgrounds. So, by classification, the objects can be well located due to the background of them is not distinguishable. While for objects of boat, bottle, and plant, the backgrounds of them are unique. So the background objects may also contributes to classification and the AP values of them are relatively lower.

Similarly, the proposed method also generates a heat map but highlight the object that has the largest probability. So the background variance between objects of different classes has less influence on the results. As a consequence, the AP values of difference objects are relatively equilib-rium. For example, AP values on most objects are above 40% except bird. While for compared methods, AP values of more objects are lower than 40%. Meanwhile, the proposed method achieves better performance on objects of bottle, chair, person, and plant than most compared methods. Even though the backbone of proposed model is a general CNN without any additional network modules and learning tricks, in terms of mean AP in the two tables, it achieves the

5. https://github.com/liusiqinqin/HAAI. Here we directly use CUDNN for coding. There are much less personal users as well as templates and tutorials than those of popular deep learning platforms such as PyTorch. So there may be some problems for reproduction. You could contact the first author without hesitation for solutions.

### Table 1
Detection Average Precision (%) on the VOC 2007 Test Set

<table>
<thead>
<tr>
<th>Method</th>
<th>acco</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>car</th>
<th>bus</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dog</th>
<th>horse</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSDNN [58]</td>
<td>48.4</td>
<td>38.3</td>
<td>33.5</td>
<td>25.9</td>
<td>21.4</td>
<td>48.8</td>
<td>33.2</td>
<td>39.2</td>
<td>8.9</td>
<td>41.8</td>
<td>26.6</td>
<td>38.8</td>
<td>44.7</td>
</tr>
<tr>
<td>DT+PL [59]</td>
<td>35.7</td>
<td>30.8</td>
<td>28.3</td>
<td>20.5</td>
<td>15.3</td>
<td>33.3</td>
<td>24.2</td>
<td>33.3</td>
<td>10.8</td>
<td>17.3</td>
<td>10.7</td>
<td>14.7</td>
<td>15.8</td>
</tr>
<tr>
<td>ZLDN [22]</td>
<td>35.4</td>
<td>30.5</td>
<td>30.1</td>
<td>18.8</td>
<td>20.8</td>
<td>62.7</td>
<td>36.3</td>
<td>31.1</td>
<td>21.9</td>
<td>37.8</td>
<td>4.5</td>
<td>40.1</td>
<td>69.7</td>
</tr>
<tr>
<td>ML-LocNet [60]</td>
<td>60.8</td>
<td>70.6</td>
<td>47.8</td>
<td>30.2</td>
<td>24.8</td>
<td>64.9</td>
<td>38.4</td>
<td>37.9</td>
<td>11.0</td>
<td>51.3</td>
<td>55.3</td>
<td>48.1</td>
<td>68.9</td>
</tr>
<tr>
<td>WCCN [61]</td>
<td>48.5</td>
<td>60.8</td>
<td>36.5</td>
<td>28.8</td>
<td>16.2</td>
<td>10.2</td>
<td>45.6</td>
<td>45.6</td>
<td>10.9</td>
<td>41.1</td>
<td>20.9</td>
<td>44.2</td>
<td>64.1</td>
</tr>
<tr>
<td>TS2C [17]</td>
<td>59.3</td>
<td>57.5</td>
<td>43.7</td>
<td>24.3</td>
<td>13.5</td>
<td>63.9</td>
<td>61.7</td>
<td>57.9</td>
<td>24.1</td>
<td>46.9</td>
<td>36.7</td>
<td>45.8</td>
<td>39.9</td>
</tr>
<tr>
<td>WSOD [62]</td>
<td>58.4</td>
<td>64.7</td>
<td>38.5</td>
<td>30.3</td>
<td>16.4</td>
<td>66.5</td>
<td>38.1</td>
<td>43.1</td>
<td>27.6</td>
<td>40.0</td>
<td>19.3</td>
<td>48.9</td>
<td>58.2</td>
</tr>
<tr>
<td>proposed/Baidu</td>
<td>63.8</td>
<td>64.4</td>
<td>40.1</td>
<td>27.7</td>
<td>17.7</td>
<td>31.8</td>
<td>40.1</td>
<td>52.9</td>
<td>42.2</td>
<td>62.7</td>
<td>63.7</td>
<td>44.1</td>
<td>64.1</td>
</tr>
<tr>
<td>MIST [63]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CASD [64]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-MIL [13]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PG-PS [18]</td>
<td>56.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 2
Detection Average Precision (%) on the VOC 2012 Test Set

<table>
<thead>
<tr>
<th>Method</th>
<th>acco</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>car</th>
<th>bus</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dog</th>
<th>horse</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZLDN [22]</td>
<td>54.3</td>
<td>63.7</td>
<td>48.1</td>
<td>16.9</td>
<td>21.5</td>
<td>70.8</td>
<td>60.4</td>
<td>59.7</td>
<td>1.2</td>
<td>53.5</td>
<td>44.4</td>
<td>36.6</td>
<td>63.6</td>
</tr>
<tr>
<td>ML-LocNet [60]</td>
<td>37.9</td>
<td>60.4</td>
<td>40.4</td>
<td>24.3</td>
<td>18.7</td>
<td>38.7</td>
<td>63.3</td>
<td>52.5</td>
<td>41.3</td>
<td>47.1</td>
<td>46.8</td>
<td>31.5</td>
<td>61.0</td>
</tr>
<tr>
<td>WCCN [61]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TS2C [17]</td>
<td>46.4</td>
<td>76.4</td>
<td>39.7</td>
<td>33.8</td>
<td>28.7</td>
<td>71.4</td>
<td>68.2</td>
<td>35.9</td>
<td>4.6</td>
<td>67.4</td>
<td>60.0</td>
<td>21.3</td>
<td>58.0</td>
</tr>
<tr>
<td>WSOD [62]</td>
<td>65.3</td>
<td>69.1</td>
<td>47.4</td>
<td>26.4</td>
<td>20.6</td>
<td>61.1</td>
<td>59.9</td>
<td>62.3</td>
<td>23.7</td>
<td>50.4</td>
<td>20.1</td>
<td>78.8</td>
<td>52.7</td>
</tr>
<tr>
<td>proposed/Baidu</td>
<td>53.4</td>
<td>56.4</td>
<td>47.2</td>
<td>24.9</td>
<td>35.4</td>
<td>38.4</td>
<td>53.9</td>
<td>63.0</td>
<td>49.2</td>
<td>52.5</td>
<td>77.9</td>
<td>55.9</td>
<td>62.0</td>
</tr>
</tbody>
</table>

5. https://github.com/liusiqinqin/HAAI. Here we directly use CUDNN for coding. There are much less personal users as well as templates and tutorials than those of popular deep learning platforms such as PyTorch. So there may be some problems for reproduction. You could contact the first author without hesitation for solutions.
state-of-the-art results which demonstrates the feasibility. However, as the discussed limitations, it only highlights the regions with the common prominent features instead of the semantic objects. Then, for object of bird, the features are diverse with different kinds of birds and the proposed method achieves lower performance than compared methods. As a consequence, it cannot outperform recent methods overall such as CASD and CPE which improve the baselines by proposing new modules and techniques such as comprehensive attention distillation in [64] and contrastive proposal extension [65]. However, the aim of this paper is not to improve localization accuracy but to detect corresponding objects given only the name.

Suppose that we want to localize the 20 types of objects but we have no datasets. Then freestyle localization is able to directly learn from raw searched data and for proposed/Baidu, the VOC datasets are only used for evaluation to demonstrate the feasibility of the localization mechanism. The results of proposed/Baidu are also provided in Tables 1 and 2. The AP values of proposed/Baidu is similar to that of proposed/VOC but for some classes, the performance is significantly degraded. On the one hand, the searched images may contain many noise. For example, in Fig. 8, the searched sheep images may not only contain the sheep but also other types of images, such as cartoons, paper-cutting, and even mutton. Even though they are all related to sheep, but are totally different from the object of sheep in test images. On the other hand, there may many processed images which may be far from the real cases. For example, we search for monitor in Baidu Image and there are many images from online shops where the images of monitor are processed to remove the background as shown in Fig. 8. Then the background information cannot be well modeled which results in miss detection of monitor. Even so, from the mAP, proposed/Baidu is able to outperform many state-of-the-art methods which demonstrates the learning capability of the proposed method. Moreover, the compared methods can only learn from provided datasets while freestyle localization method can learn from either datasets or searched images. Compared with existing methods that requires days or even months to prepare and annotate the datasets, freestyle localization significantly reduces the efforts of human.

The visible results obtained by proposed/VOC and proposed/Baidu are shown as heat maps in Fig. 9. In the visualized images, both of them is able to highlight the corresponding objects and the performance of them seems similar due to the same localization mechanism. There are also some differences. There are more fragmented regions in those by proposed/Baidu. This is because the background information in searched images may not cover all the cases in VOC datasets. Then it is difficult for it to distinguish background objects. Even though the location of corresponding objects can be highlighted, due to the lack of semantic information, the completeness of the highlighted objects cannot be guaranteed such as the localization of bus and horse.

The results on COCO datasets are listed in Table 3 where the proposed method achieves the best overall result among state-of-the-art weakly-supervised methods. Experiments on public datasets demonstrate that the proposed method is able to generate acceptable localization results and the

<table>
<thead>
<tr>
<th>method</th>
<th>mAP</th>
<th>mAP_{0.5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCCN [61]</td>
<td>-</td>
<td>12.3</td>
</tr>
<tr>
<td>PG-PS [18]</td>
<td>-</td>
<td>20.7</td>
</tr>
<tr>
<td>MIST [63]</td>
<td>13.0</td>
<td>26.3</td>
</tr>
<tr>
<td>CASD [64]</td>
<td>13.9</td>
<td>27.8</td>
</tr>
<tr>
<td>C-MIL [13]</td>
<td>7.8</td>
<td>18.8</td>
</tr>
<tr>
<td>CPE [65]</td>
<td>13.7</td>
<td>27.1</td>
</tr>
<tr>
<td>proposed/COCO</td>
<td>14.2</td>
<td>29.1</td>
</tr>
</tbody>
</table>
objects within the 80 classes. For new objects as tested in this subsection, we should collect, pre-process, and annotate sufficient data for them. While for the proposed method, we just tell the name and the data preparing and learning can be automatically achieved without any efforts of humans. So we impromptu choose 10 objects which are infrequent in public datasets to test the freestyle localization capability. Localization of the 10 objects are widely required in different applications, such as commerce, entertainment, and military. But we fail to find an existing method that can conveniently deal with those objects without the efforts of humans.

Locating animals can be used for wildlife conservation but requires professional workers to prepare and annotate related dataset. Then freestyle localization can be utilized in this case. Recognizing a specified person may find their applications in entertainment and commerce. But there are few datasets for training localization model of specific persons which limit the applications of existing methods. Here we use the names of Trump and Yun Ma for test, respectively. There are various movie roles and it is difficult to prepare the dataset to learn from them. Then we focus on the localization of Iron Man and WALL-E. Localizing new types of weapons is significant in military defense while the datasets are usually not public. But there are images of various weapons online. Finally, we test the proposed method by abstract expressions on different faces to demonstrate the information mining capability.

The fused heat maps of some test images are exhibited in Fig. 10 and the localization precision is listed in Table 4 where AP on test sets and CorLoc [18] on training sets are exhibited. Here we directly show the fused heat maps with the original images. Among the 10 cases, tigers have the most prominent features with the special texture. So they can be well recognized from different backgrounds. Moreover, a cartoon tiger is also recognized but another one is missed due to its inconspicuous texture. Panda also has prominent features with black and white fur. But many local regions are highlighted such as neck and belly which are the regions with color change. Even so, different pandas can be well recognized. Persons are more difficult to localize due to the similar color and geometry of different faces. But the proposed method is able to highlight specified face after learning. From the results on Trump and Yun Ma, the proposed method is able to distinguish the learned person from other people and background objects. For Trump, the prominent feature is the golden hair. So in many images, the hair is specially highlighted. For Yun Ma, the proposed method is able to recognize him in various backgrounds and scenarios. Even more surprising, it recognizes the Yun Ma in childhood while assigns much less confidence to an online celebrity named “little Ma Yun”. Iron Man is in more complex backgrounds due to the movie scenarios. But the proposed method model is able to well recognize it. The prominent feature of the suit is the mechanical texture. So, the results specially highlight this part. When there are more than one suits in the scenarios, both suits can be recognized and the proposed method also distinguishes Iron Man and Captain America or other roles. Moreover, WALL-E is also well localized in the movie scenarios. By learning from the searched weapons, the proposed method is able to distinguish B-2 from other airplanes and recognize the key features of aircraft carrier. Most interestingly, the proposed method can not only localize concrete objects but also abstract expressions. We search for images of “happy animals” and the proposed method is able to locate the expression of “happiness” on different animals after learning.
In most images, the mouth is located due to that it shows the “happiness”. Moreover, the popular emoji of Jacky Cheung in chatting applications (a very funny expression) can also be learned and located on various faces. The proposed method can not only locate such an expression of Jacky Cheung himself, but also on other people or synthetic faces.

From Table 4, the localization accuracy is even higher than that on public test sets which demonstrates that the proposed method is able to achieve acceptable localization accuracy in searched images. However, since the heat map is generated via the gradient back-propagated from association layers, as the discussed limitations, only local regions with prominent features are highlighted. For example, only part of Iron Man and aircraft carriers are highlighted in some heat maps. Moreover, the highlighted regions sometimes are scattered, i.e., the heat map highlights different parts of the object. The integrality of objects cannot be guaranteed.

Similar to general learning methods, the generalization and robustness of the proposed method also depend largely on the abundance of training data. The mechanism is to highlight the consistent components in searched images such as the special textures of tigers. So, as long as such components are not occluded, it can be recognized in test images. As a consequence, the proposed method is robust to variance of object sizes, postures, and even occlusions with abundant training samples. We show some cases where the objects are not normally presented in Fig. 11. With the special texture, the proposed method is able to recognize the occluded or partially visible tigers. With the black and white fur, pandas of different postures and in different backgrounds can be recognized. The variances of movie roles are much larger than nature objects. While the proposed method is also able to recognize them in different scenarios, even in those out of movies such as games.

As one of the difficulties in dealing with searched images, the severe size variance is a great challenge for existing learning architectures. Fig. 12 shows the image sizes of all the searched images. With various online sources, the image sizes vary a lot not only within one class but also in different classes. For example, many images of Emoji are small while many B-2 images are with large size. We can simply select images of which the sizes are within a range (for example, width ∈ [400, 600] and height ∈ [400, 600]) as in many datasets to deal with the size variance but the volume of training data will be significantly reduced and the generalization of the proposed method will be degraded. We show the percentage of image sizes that are within such a range in Fig. 12 and the results trained by the selected images are exhibited in Fig. 13. The performance of the proposed method is significantly reduced on all test images when it is only trained by the selected images. The results also demonstrate that the volume of training data influence the generative and robustness a lot, especially for objects in complex scenarios such as movie roles. Moreover, the image size may also be unpredictable in test data and practical applications. So it is necessary to deal with the severe size variance in the freestyle localization scenario.

4.4 Discussion
One of the aim of artificial intelligence is to reduce the burden of humans in various applications. Many methods have been proposed to further reduce the burden of humans during learning process, such as weakly-supervised and unsupervised ones. We define a freestyle localization problem
which requires the learner to autonomically learn via any available data and we focus on images in Internet in this paper. From the above experiments, the proposed model has significant application potentials with three reasons. (1) Compared with existing weakly-supervised methods, it achieves equivalent performance to them which guarantees the localization accuracy in some applications. Moreover, the performance can be further improved via data augmentation, more complex architecture and learning tricks, and finely designed inference algorithms. (2) Most importantly, the proposed method is able to learn any objects as long as they are sufficiently available in Internet. The localization scope is not limited by existing public datasets which is the core of freestyle learning and is specially important in some urgent applications. (3) The proposed method is established based on the output of association module which is compatible to many multi-layer backbone network architectures. This makes it possible to deal with other types of data such as voice, languages, and even videos.

However, as a totally new problem and new method, there are many apparent and latent problems. For example, as shown in experiments, one of the apparent problems is that the proposed method cannot guarantee the integrity of the objects, only part of the objects with common features can be recognized. This can be alleviated by designing new inference methods as many improved CAMs. The latent problems may include potential overfitting and robustness. In practice, objects may have a wide range of appearances and variations. As analysed above, the overfitting can be avoided by abundant training samples. With the diversity of searched images, the abundant images can be guaranteed for most objects and the proposed method is able to detect objects with different postures and appearances as shown in Fig. 11. For some objects with scarce images in Internet, in future work, we intend to explore some training tricks such as data augmentation and network parameter regularization. For the robustness, the diversity of scenarios of most objects can also be guaranteed in searched images. But for some special objects such as the monitor in experiments, the background information cannot fully represent the real cases. As a consequence, we will develop new optimization methods to directly optimize the probabilistic model in Eq. (1) in order to estimate all possible scenarios.

5 Conclusion
This paper presents a freestyle object localization problem which is able to learn freely and localize an arbitrarily specified object given only the name. Moreover, we propose a novel learning model to deal with such a challenging problem with the assistance of Internet and searching engines. To directly learn from searched images, an association module is constructed based on a hierarchical network to define the energy that drives a probabilistic model. During the optimization, with the difficulty in estimating distribution of the whole data space, we propose to use training images to estimate that of background instead of whole data space. With freestyle object localization, we only need to tell the system what we want it to learn without preparing labeled data or any other assistant information for it. The freestyle learning makes it possible to learn everything in web without the limitation of existing datasets. In some applications, it significantly reduces the efforts of human in preparing dataset for training. Experiments on three public datasets demonstrate the effectiveness and state-of-the-art performance. Moreover, experiments on searched images demonstrate its freestyle object localization capability which is a great challenge for existing learning methods.

However, as a totally new framework, the defect is also obvious and there are also many other challenges to accurately learn from searched images. Therefore, in the future work, we will consider but not limit to the following orientations. (1) We will refer to CAM derived methods due to the similar detection process and improve the detection integrality. (2) We will focus on the noise in searched images, such as irrelevant objects, and improve the robustness to them. (3) We will try to overcome the difficulties in distribution estimation to reduce the dependence on background information. (4) We will explore intelligent learning models for more freestyle applications such as freestyle classification, freestyle semantic segmentation, and freestyle image generation.

References

Jia Liu (member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Xidian University, in 2013 and 2018, respectively. Now, he is an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include computational intelligence and image understanding.

Donghui Li received the B.S. in software engineering from Shandong University of Technology, Zibo, China, in 2021. He is pursuing the master’s degree in computer science with the Nanjing University of Science and Technology, Nanjing, China. His research interests include few-shot learning and image segmentation.

Wenhuai Zhang (member, IEEE) received the B.S. degree from North University of China, Taiyuan, China, in 2015, and the Ph.D. degree from Xidian University, Xi’an, China, in 2021. She is currently a Post-Doctoral Researcher with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include object tracking and image processing.

Fang Liu (member, IEEE) received the B.S. degree in information and computing science from Henan University, Kaifeng, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi’an, China, in 2018. She is an Associate Professor with the Nanjing University of Science and Technology, Nanjing, China. Her research interests include deep learning, object detection, image classification, and change detection.

Zhen Cui (member, IEEE) received the Ph.D. degree from Institute of Computing Technology (ICT), Chinese Academy of Sciences in 2014. He was a Research Fellow in National University of Singapore (NUS) from Sep 2014 to Nov 2015. Currently, he is a Professor of Nanjing University of Science and Technology, China. His research interests cover computer vision, pattern recognition and machine learning, especially focusing on deep learning, graph neural network, etc.
Licheng Jiao (Fellow, IEEE) received the B.S. degree in high-voltage engineering from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in theoretical electrical engineering from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. Since 1992, he has been a distinguished professor with the school of Electronic Engineering, Xidian University, Xi’an, where he is currently the Director of Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He has been a foreign member of the academia European and the Russian academy of natural sciences. His research interests include machine learning, deep learning, natural computation, remote sensing, image processing, and intelligent information processing. Prof. Jiao is the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the fellow of IEEE/IET/CAAI/CIE/CCF/CAA, a Councilor of the Chinese Institute of Electronics, a committee member of the Chinese Committee of Neural Networks, and an expert of the Academic Degrees Committee of the State Council.