

# ***IET Computer Vision***

## **Special issue** **Call for Papers**

---



**Be Seen. Be Cited.**  
**Submit your work to a new  
IET special issue**

Connect with researchers and  
experts in your field and share  
knowledge.

Be part of the latest research  
trends, faster.


**Read more**



**The Institution of  
Engineering and Technology**

## ORIGINAL RESEARCH

# Unsupervised detection of contrast enhanced highlight landmarks

Tao Wu<sup>1,2,3</sup>  | Wenzhuo Fan<sup>2,3</sup> | Shuxian Li<sup>1,2,3</sup> | Qingqing Li<sup>3</sup> | Jianlin Zhang<sup>3</sup> | Meihui Li<sup>3</sup>

<sup>1</sup>Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu, China

<sup>2</sup>School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup>Institute of Optics and Electronics Chinese Academy of Sciences, Chengdu, China

## Correspondence

Jianlin Zhang  
Email: [jlz@ioe.ac.cn](mailto:jlz@ioe.ac.cn)

## Abstract

In the field of landmark detection based on deep learning, most of the research utilise convolutional neural networks to represent landmarks, and rarely adopt Transformer to represent and encode landmarks. Meanwhile, many works focus on modifying the network structure to improve network performance, and there is little research on the distribution of landmarks. In this article, the authors propose an unsupervised model to extract landmarks of objects in images. First, Transformer structure is combined with the convolutional neural network structure to represent and encode the landmarks; next, positive and negative sample pairs between landmarks are constructed, so that the semantically consistent landmarks on the image are pulled closer in the feature space and the semantically inconsistent landmarks are pushed farther in the feature space; then the authors concentrate their attention on the most active points to distinguish the landmarks of an object from the background; finally, based on the new contrastive loss, the network reconstructs the image by the landmarks of the object that are continuously learnt during training. Experiments show that the proposed model achieves better performance than other unsupervised methods on the CelebA, Annotated Facial Landmarks in the Wild, 300W datasets.

## KEYWORDS

computer vision, image processing, unsupervised learning

## 1 | INTRODUCTION

Landmark detection, as one of the important tasks in computer vision, is widely applied in pose estimation, motion capture, behaviour analysis, and virtual games. With the development of deep learning, the landmark detection of objects has gradually progressed from the traditional method of abstract representation of objects [1, 2] to the method based on deep learning.

At present, landmark detection methods based on deep learning are mainly divided into supervised methods [3–6] and unsupervised methods [7–10]. The supervised methods demand to add the ground truth of the images to supervise the network during training. However, manual annotation of the labels of a large amount of image data is slow, expensive, and difficult to achieve. The unsupervised way does not require the real labels of the data, which has obvious advantages in the face of large datasets.

The landmark detection task is a typical regression problem. The early application of deep convolutional neural network for landmark detection is achieved by directly regressing the position coordinates of landmarks. It means that the features extracted by the network are finally returned to the landmark coordinates and confidence information of the object directly through a fully connected layer, such as refs. [13, 14]. After that, predicting landmark heatmaps [15–18] became the mainstream method. That is, each landmark heatmap is obtained by convolution or encoding of the final output feature layer, wherein the position of the most active point in the heatmap represents the position of the corresponding landmark.

These unsupervised methods [5–8, 15–18] all use the method based on landmark heatmap to extract landmarks. They all finally train a landmark extraction network, and determine the location coordinates of the landmarks by converting the landmark feature maps into landmark heatmaps. Furthermore,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *IET Computer Vision* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

these methods all employ the convolutional neural network (CNN) structure to represent the landmarks of the object. However, the representation ability of the CNN structure for pixel-level landmarks is not very strong. In addition, in the final loss function, most of the studies utilise the difference between the reconstructed image and the original image as a judgement of network effectiveness. We hope that the landmarks of the extracted target are closest to the true value of the landmarks, not that the reconstructed image is close to the original image. If the loss function only has the difference between the reconstructed image and the original image, the whole problem becomes a problem of reconstructing an image, not a problem of extracting landmarks.

To solve the above problems, this paper proposes an unsupervised algorithm to extract the landmarks of objects. The algorithm applies the Transformer [19] structure combined with CNN to capture and encode landmarks. At the same time, two parallel networks are employed to extract the landmarks and background of the object, forming a separation effect between the landmarks and the background. What's more, positive and negative sample pairs between landmarks are constructed, which can make semantically consistent landmarks aggregate, and semantically inconsistent landmarks mutually exclusive. In the process of continuous training, the network will focus on the changing points in the two images, and the changing points indicate that the most active position in the image is the landmark position of the object. The contributions of the paper are as follows:

1. Based on the insight that landmarks could be highlight position of information to represent object, we propose a new unsupervised landmark detection network which extracts object information and aggregates them via Transformer structure. By focussing on active position of the heatmap, the landmarks can be robustly extracted and precisely located.
2. Constructing positive and negative sample pairs between landmarks and employing the contrastive loss to make semantically consistent landmarks closer and semantically inconsistent landmarks to be separated.
3. Our proposed model performs well on CelebA, Annotated Facial Landmarks in the Wild (AFLW), 300W datasets, and has better results than previous unsupervised models.

The rest of this article is organised as follows. Section 2 introduces Transformers, unsupervised related work and active position acquisition. The proposed model is elaborated in Section 3. Experimental results and analysis are presented in Section 4. Finally, conclusions are presented in Section 5.

## 2 | RELATED WORK

### 2.1 | Transformer

Transformer is the earliest model proposed by Google to solve natural language processing tasks. Now, many

researchers have applied Transformer in the field of computer vision. The Vision Transformer (ViT) [21] model structure is consistent with the original Transformer, and finally a Multilayer Perceptron Head is used to implement image classification. DeiT [22] introduces a knowledge distillation structure that requires less data and less computing resources to achieve better image classification results. Transformer in transformer (TNT) [23] extracts local and global features jointly by two Transformer structures inside and outside, and makes it outperform ViT and DeiT by stacking TNT modules. To predict object instances, Detection Transformer [24] gives a fixed set of learnable target query sets, and finally achieves binary maximum matching through the Hungarian algorithm. Segmentation Transformer [25] employs a convolution feature to stack multiple Transformer structures to achieve the purpose of object segmentation. There are many tasks to apply Transformer in computer vision as well, such as refs. [26, 27].

In this paper, we combine Transformer with unsupervised landmark detection for the first time and utilise Transformer's self-attention mechanism to capture and encode landmarks of objects. At the same time, combined with the landmark and background information of the object, the detection of landmarks is realised by reconstructing the image.

### 2.2 | Unsupervised learning of object landmarks

There are many studies on object landmark detection. However, most methods need to add real labels as supervision signals. The unsupervised method of extracting landmarks is mainly achieved by learning the landmark representation network in the reconstructed process.

In recent years, many researchers have made great contributions in the field of unsupervised extraction of object landmarks. Jakab et al. [17] reconstructed the object image based on the generated landmark heatmaps and learnt the landmarks of the object from the synthesised image. Although it is simple of his network structure, his method does not distinguish the background and foreground and it is easy to mark the background points as the landmarks of the whole image. The authors in ref. [18] propose a transporter structure, which adopts two branches and blocks the gradient of one branch during training to achieve the retention of landmarks information. The method of ref. [28] adds a variety of constraints on landmarks during training to make the distribution of landmarks in the image more reasonable.

Compared with their work, our algorithm extracts the landmark information and background information in the image through a parallel structure and completes the separation of landmarks and backgrounds. At the same time, in order to make the entire network focus on the most active points in the landmark representation map, this paper have performed a Gaussian process on the landmark representation map, which provides clear foreground information for the subsequent reconstructed images.



## 2.3 | Active position acquisition

In most studies on landmark extraction, the perceptual loss [29] is employed as a final measure of whether the network is effective. The location of the final landmark is determined by the most active location in the landmark representation map.

We want to strengthen self-supervision between landmarks. Contrastive learning has recently emerged as a very effective method in self-supervised learning in computer vision. In these studies [30, 31], they adopted the method of contrastive learning in specific downstream tasks. The key idea of contrastive learning is to bring similar data closer and push data with large differences in similarity farther.

Our downstream task is landmark extraction. Similar to most contrastive learning methods, we also need to construct positive and negative sample pairs between landmarks. The positive sample pairs of landmarks are extracted by using the Gaussian heatmap after data augmentation. The points with the same semantic information in two different instance images containing the same category of objects are the positive sample pairs of landmarks. For a certain landmark, all the remaining landmarks are its negative sample pairs.

After constructing the positive and negative samples between landmarks, the contrastive loss is combined with the perceptual loss to form the loss function of the entire network. At the same time, in the landmark representation maps, the position of the most active point is the landmark position of the object.

Compared with the work of refs. [15–18], we add a contrastive loss, which not only enables the restriction of positive landmark pairs, but also prevents negative landmarks from having a certain influence on the result. Like most contrastive learning methods, positive pairs are generated by data augmentation as well, except that sample pairs are not different views of the same object. Contrastive loss is also trained to minimise the Info Noise Contrastive Estimation loss.

## 3 | METHOD

Let  $x \in R^{H \times W \times 3}$  be an image containing objects and complex backgrounds, and  $xw \in R^{H \times W \times 3}$  be another instance containing objects of the same class. The difference from image  $x$  to image  $xw$  may be caused by many reasons, such as changes in colour, viewing angle, shape, lighting, and other factors. The change is generally divided into two types, the change in appearance and the change in shape. In order to better extract the features of the object and distinguish the difference between the object and the background, we perform a colour transformation on the image  $x$ , denoted as  $xa \in R^{H \times W \times 3}$ . The image  $xa$  is generated by the colour change of the image  $x$ , and the image  $xw$  can be selected in a variety of ways. For continuous video, the image  $xw$  can be adjacent video frames. For regular datasets, the image  $xw$  can be generated using spatial affine transformation or thin plate spline (TPS).

The overall structure of the entire network is shown in Figure 1, which mainly includes three parts: landmark information extraction sub-network, background separation sub-network, and landmark aggregation and exclusion part. The

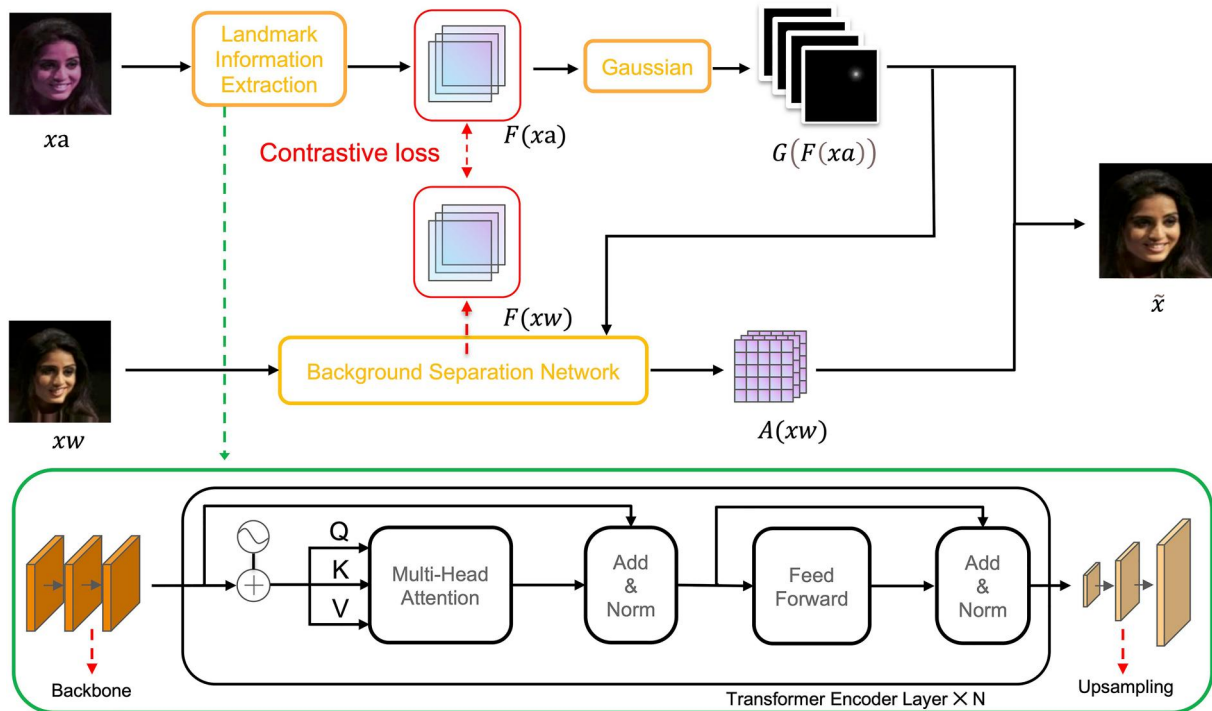


FIGURE 1 The overall framework structure of the network proposed.

reconstructed image obtains the landmark information of image  $xa$  through landmark information extraction sub-network; the reconstructed image obtains the background information of image  $xw$  through background separation sub-network; finally, the landmarks of the objects in the image are extracted more accurate through landmark aggregation and exclusion part.

### 3.1 | Landmark information extraction

The landmark information extraction network which is also applied in Section 3.2 aims to capture and encode the landmarks in the image  $xa$  to control the shape of the reconstructed image. The sub-network is divided into two steps, which are the information extraction part and the landmark heatmap generation part.

In the first step, the information extraction part takes the image  $xa \in R^{H \times W \times 3}$  as input and obtains the feature map  $F(xa) \in R^{H \times W \times K}$ , where  $K$  is the number of landmarks you want to extract. Since the self-attention mechanism in Transformer can make the network pay more attention to the semantically consistent landmarks, we adopt the Transformer structure to capture the landmarks of objects and encode them. In the task of landmark detection, the purpose is to capture the landmarks of the object and encode the landmarks. Therefore, we only need to adopt the encoding part in the Transformer and rely on the attention mechanism to obtain dependencies from the sequence without decoding operations. Here, we use a standard Transformer encoding structure.

The detailed structure of the information extraction part is shown in the green box in the Figure 1. It shows that the image  $xa$  first goes through a backbone network to obtain a two-dimensional feature map of size  $C \times H/16 \times W/16$ . This backbone network is a lightweight network mainly composed of multiple downsampling modules. Then, the feature map is flattened into a  $L \times C$  sequence, where  $L \in R^{H/16 \times W/16}$ , and sent to the encoder structure of the Transformer. Finally, the sequence through the encoder is restored to the two-dimensional data, and the data is restored to the images of size  $H \times W$  through the upsampling network.

The core of the Transformer structure is the multi-head attention mechanism. The input sequence  $X \in R^{L \times C}$  is mapped to three new subspace representations  $Q, K, V \in R^{L \times C}$  through the weight matrix  $W_Q, W_K, W_V \in R^{C \times C}$ . Compute the attention matrix from  $Q, K$ , and  $V$ :

$$A = \text{softmax}(Q \times K^T) \times V \quad (1)$$

Attention matrix expounds that the query vector  $Q_i \in 1 \times C$  at each  $X_i$  (the feature vector at position  $i$ ) is operated on with all key vectors  $K$ , and the correlation weight  $W_i \in R^{1 \times L}$  of  $Q_i$  and  $K$  is obtained. Then, a Softmax function is passed to normalise the data and widen the difference between the data. Finally, the attention value is achieved by the

weighted summation of the weight coefficients  $W$  and  $V$ . By doing so, the attention mechanism captures landmark-to-landmark correlations in image  $X$ , and reveals how much of the predicted contribution of each landmark is aggregated from various locations in the image. The number of encoders is the same as standard Transformer, and the number is six. What's more, the  $1 \times 1$  convolution which changes the dimension of  $C$  to  $K$  will be applied to make the number of channels of the feature map equal to the number of extracted landmarks.

The ultimate purpose of landmark information extraction sub-network is to extract the landmark information that is more of local information in the image  $xa$ . In the traditional way of using the fully connected layer to regress the landmark coordinates, all the landmarks are calculated at the same time. And they share the same feature information, which is quite different from heatmap-based methods. The method of calculating landmarks based on heatmap is to do feature matching in the spatial dimension. It causes the convolution kernel to slide on the feature map plane, and more attention is paid to and utilises local information. Therefore, it is necessary to employ heatmap-based method to complete the extraction of landmarks.

In the second step, the landmark heatmap generation part takes the result of the information extraction part  $F(xa) \in R^{H \times W \times K}$  as input, and outputs the heatmap  $G(F(xa)) \in R^{H \times W \times K}$  containing the landmark information, where  $K$  is the number of landmarks. For each landmark feature map  $F_i = H \times W$ , ( $i = 1, 2, \dots, K$ ), it is re-normalised to a probability distribution by the Softmax function:

$$P_i = \frac{\sum_{u \in \Omega} u e^{F_i}}{\sum_{u \in \Omega} e^{F_i}} \quad (2)$$

where,  $u$  is the value of each pixel in  $F_i$ , and  $\Omega$  is the definition domain of the image. The final landmark Gaussian heatmaps are all replaced by a Gaussian function centred on  $P_i$ :

$$\text{Gauss}(k) = \exp\left(-\frac{1}{2\sigma^2}\|u - P_i\|^2\right) \quad (3)$$

where  $\text{Gauss}(k) \in H \times W$ , in the range 0–1, and  $\sigma$  is a fixed standard deviation.

### 3.2 | Background separation

For the landmark detection task, if the landmarks are regarded as foreground information, the information other than the landmarks can be regarded as background information. And in addition to the inconsistency of landmark information between image  $xa$  and image  $xs$ , only the appearance information is inconsistent. Therefore, in this section, the background separation sub-network is constructed to obtain the background information from image  $x$  to image  $xw$ . The background

information is appearance information. The structure diagram of the background separation sub-network is shown in Figure 2.

The background separation sub-network takes the image  $xw$  as input, and first passes through landmark information extraction sub-network to obtain the landmark Gaussian heatmap  $G(F(xw)) \in R^{H \times W \times K}$ . Next, we stack the information of the first convolutional layer in  $F(xw)$  with the image  $xw$  and send it to the appearance encoder to get the appearance features representation map  $\alpha(xw) \in R^{C \times W \times K}$  of the image  $xw$ . Then, the information representation vector  $V(xw)$  of image  $xw$  was achieved by using the landmark Gaussian heatmap  $G(F(xw))$  of image  $xw$  to perform a spatial pooling on the appearance feature representation map  $\alpha(xw)$  of image  $xw$ . Finally, the information representation vector  $V(xw)$  is projected onto the landmark heatmap of image  $xa$ , and the background information representation map  $A(xw)$  of image  $xw$  can be obtained. The background information representation map  $A(xw)$  of image  $xw$  is equal to the background information representation map  $A(x)$  of the image  $x$ .

The background separation sub-network can effectively extract the background information of the image  $xw$ , and minimise the influence of the landmark information of the image  $xw$  on the reconstructed image when reconstructing the image  $x$ .

### 3.3 | Landmark aggregation and exclusion

The reconstruction of the image  $x$  is composed of the landmark information extracted from the image  $xa$  and the background information of the image  $xw$ . The landmark information of the image  $x$  is the same as the image  $xa$ , and the background information of the image  $x$  is the same as the image  $xw$ . However, if there is only the loss between the reconstructed image and the original image, it is easy to cause overfitting during training. Moreover, our goal is to extract the landmarks of the target, rather than expecting the reconstructed image to be more consistent with the original image. Therefore, we introduce a contrastive loss to realise semantic consistency and achieve more accurate positioning of landmarks by controlling the landmarks of the extracted objects. Of course, this is a completely unsupervised method.

The loss of the entire network consists of two parts. The first part is the perceptual loss between the reconstructed image and the original image [29], which makes the features of the reconstructed image and the original image between different network layers tend to be consistent; the second part is the contrastive loss. Contrastive learning [20] has just begun to be applied to image classification. Its core idea is to shorten the distance between positive sample images and increase the distance between negative sample images. Positive sample images are similar images and negative sample images are different images. Inspired by the idea of contrastive learning, we also construct positive samples and negative samples between landmarks, making the distance between the landmarks of the positive samples (semantically consistent landmarks)

closer and the distance between the landmarks of the negative samples (semantically inconsistent landmarks) farther.

The reconstruction network  $\Gamma(A(xw), G(F(xa)))$  is a conventional convolutional network. It stacks the appearance information  $A(xw)$  of the image  $xw$  and the Gaussian heatmap  $G(F(x))$  containing the landmark information of the image  $xa$  as input and outputs  $\tilde{x} \in R^{H \times W \times K}$ , where  $\tilde{x}$  is the reconstructed image of image  $x$ .

Our idea is that the difference between image  $x$  and image  $xa$  is only in background, but the positions of their landmarks are the same, and the change between image  $x$  and image  $xw$  is only in the shape of the object. Therefore, a landmark  $(x_i, y_i)$  of the object in the image  $xa$  has a landmark  $(xw_i, yw_i)$  on the image  $xw$  corresponding to it, and the Gaussian heatmap corresponding to the landmark should also be highly similar. At the same time, the landmark  $(x_i, y_i)$  in the image  $xa$  have only one corresponding landmark  $(xw_i, yw_i)$  on the image  $xw$ , that is, they constitute a positive sample pair. While the landmark  $(x_i, y_i)$  of the image  $xa$  and all other landmarks on image  $xw$  except  $(xw_i, yw_i)$  constitute negative sample pairs, and their corresponding landmark Gaussian heatmaps should also make a big difference. For the similarity between each landmark in the image  $xa$ , we use the Hadamard product of the heatmap corresponding to the landmark to obtain, namely:

$$S = G(F(x_i)) \odot G(F(x_j)) \quad (4)$$

Among them,  $i = 1, 2, \dots, K$  and  $i \neq j$ . For the similarity between the landmark  $(x_i, y_i)$  of the image  $x$  and the landmark  $(xw_i, yw_i)$  of the image  $xw$ , use:

$$S = G(F(x_i)) \odot G(F(xw_i)) \quad (5)$$

The similarity matrix calculated by the similarity formula reflects the comparison of landmarks in the same batch and the comparison of landmarks in different batches. Its schematic diagram is shown in Figure 3. Then,  $S$  is normalised, and the contrastive loss, as before, can be written as:

Contrastive loss

$$= - \sum_{n=1}^N \sum_{k=1}^K \frac{\exp\left(S_k^{(n)} \hat{S}_k^{(n)} / \tau\right)}{\exp\left(\frac{S_k^{(n)} \hat{S}_k^{(n)}}{\tau}\right) + \sum_{j \neq k} \sum_{i \neq n} \exp\left(S_k^{(n)} \hat{S}_j^{(i)} / \tau\right)} \quad (6)$$

where  $\tau$  is a temperature hyperparameter that controls the 'kurtosis' of the similarity measure. This formula indicates two optimisation conditions, one that maximises the similarity between the same landmarks and the other that maximises the dissimilarity between landmarks within different batches.

Through the above method, we extracted the landmarks of the object, and through contrastive learning, the network can learn semantically consistent landmarks in the changes of the two images.

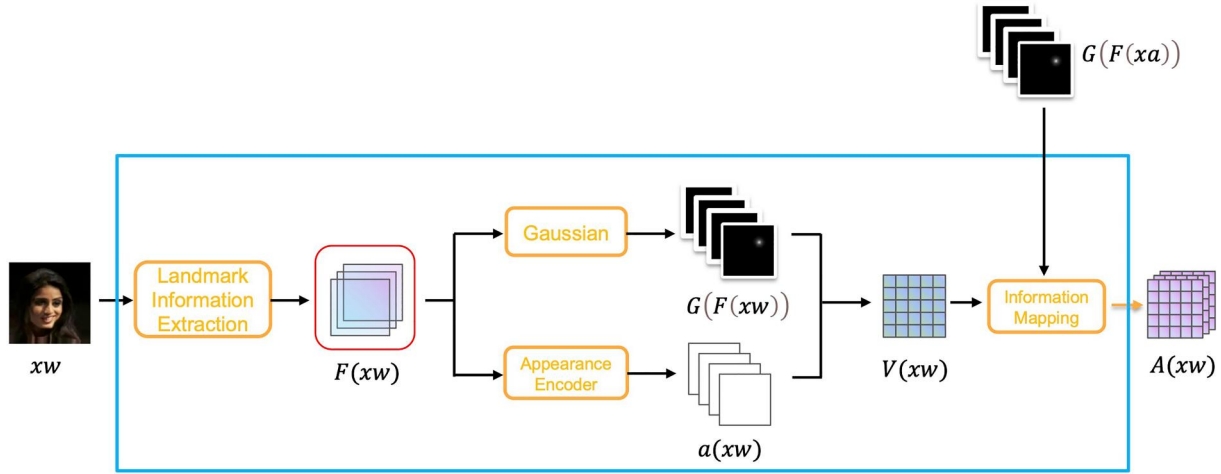
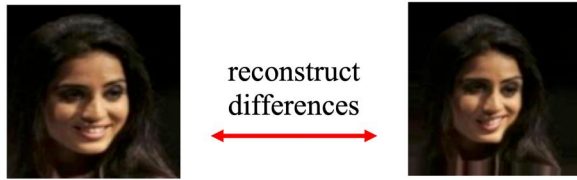
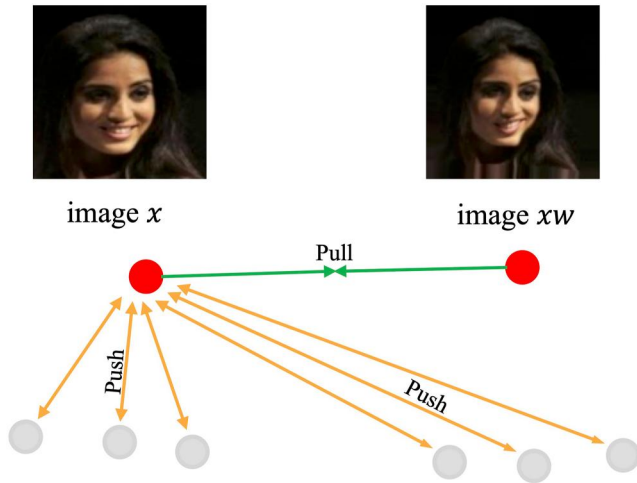


FIGURE 2 The structure of the background separation sub-network.



(a) perceptual loss



(b) contrastive loss

FIGURE 3 The loss function component of the network.

## 4 | EXPERIMENTS

In Section 4.1, we introduce the dataset, experimental details, and evaluation indicator. Next, we describe in detail the results of our algorithm on the datasets and the comparison with other algorithms in Section 4.2. In Section 4.3, ablation experiments are done to verify the effectiveness of the introduction module and the rationality of some super parameters.

### 4.1 | Setup

#### 4.1.1 | Dataset

**CelebA:** CelebA has a total of more than 20w images. Multi-Attribute Labelled Faces (MAFL) (a subset of CelebA dataset) was adopted as the test set, and the part of CelebA excluding MAFL as the training set. Like the majority of methods, we train on the training set and regress to five landmark ground truths, and finally test on the test set.

**AFLW:** The AFLW dataset consists of 21,997 web photos, including a total of 25,993 faces. Each face is marked with 21 landmarks. Like the CelebA dataset, we finally train and return to the five landmarks of eyes, nose, and mouth. And 90% of the data set is used as the training set, and the remaining 10% is used as the test set.

**300W:** The 300W dataset is a very general face alignment dataset. The dataset has a total of 3837 images, each image contains more than one face, but only one face is annotated for each image. The dataset has a total of 3148 images in the training set and 689 images in the test set.

#### 4.1.2 | Model detail

On the dataset described in Section 4.1.1, we conducted training and evaluation. In the process of training and testing, because the size of data is not uniform, all the images are adjusted to  $128 \times 128$  in the data pre-processing stage. For the information extraction and background information extraction, Xavier init is used for initialisation, and pre-training weights are not applied. All models are trained with a batch size of 64 on four graphics processing unit (GPUs). We trained our algorithm using Adam optimiser with the initial learning rate as  $2 \times 10^{-4}$  and the weight decay as  $1 \times 10^{-4}$ . The total number of epochs is 150. At 30 and 90 epochs, the learning rate is adjusted to  $2 \times 10^{-5}$  and  $2 \times 10^{-6}$  respectively.



### 4.1.3 | Evaluation indicator

A fast evaluation of network performance first requires an evaluation indicator. Since the unsupervised learning method does not have real labels during training, it is different from the evaluation indicator of the supervised method. We map the extracted  $K$  landmarks to the real labels of the dataset by regression and compare them with the real labels of the dataset. Similar to these methods [7–9, 15–18], we adopt the error between the predicted value and the true value as a percentage of the distance between the eyes as the evaluation metric.

## 4.2 | Main results

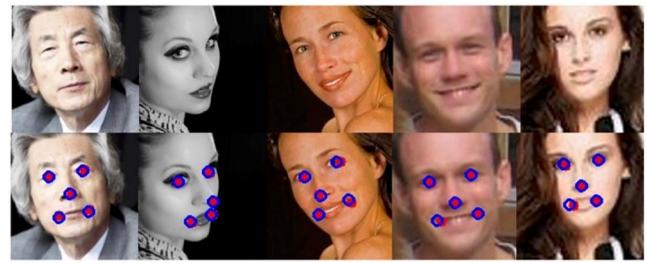
For each instance image  $x$ , we utilise TPS (thin film spline interpolation) to generate image  $xw$  and colour transformation to generate image  $xa$ . Then, image  $xw$  and image  $xa$  are sent to the network to learn. We set the number of landmarks to 10, and finally return to the five landmarks of the left eye, right eye, nose, left corner of the mouth, and right corner of the mouth, and calculate the percentage of error.

Figures 4–6 shows the result of mapping the five landmarks obtained by our network regression together with the ground-truth values of the five landmarks on image  $x$  on the test sets of CelebA, AFLW, and 300W respectively. The red solid points are the landmarks predicted by the network, and the blue hollow points are the real labels of the image. From the effect in the figures, our method can perform well on the three datasets, and the extracted landmarks can well capture the facial features when the image changes. Besides, the difference between the landmarks obtained by regression and the true value of the landmarks is very small.

Tables 1–3 show the overall mean error percentages of our network on the test set of the three datasets CelebA, AFLW, 300W. The backgrounds of characters in the AFLW and 300W datasets are more complex than those in the CelebA dataset. Therefore, it is also more difficult to train the network when faced with the AFLW and 300W datasets. However, the data in the tables shows that our method cannot only achieve good performance on datasets with relatively simple character backgrounds, but also accurately extract landmarks of objects in the case of complex backgrounds, which further confirms that our method is robust.



**FIGURE 4** The performance of our algorithm on the CelebA dataset.



**FIGURE 5** The performance of our algorithm on the Annotated Facial Landmarks in the Wild dataset.



**FIGURE 6** The performance of our algorithm on the 300W dataset.

**TABLE 1** Results of testing on the CelebA dataset.

Method	Unsupervised	Mean error
MTCNN [3]	No	5.39
TCDCN [4]	No	7.95
Cascaded CNN [5]	No	9.73
CFAN [6]	No	15.84
BRULÈ [32]	Yes	6.8
Thewlis [7]	Yes	5.83
Shu [8]	Yes	5.45
Dense 3D [9]	Yes	4.02
Mallis D [10]	Yes	3.32
Ours	Yes	2.49

Abbreviation: CNN, convolutional neural network.

**TABLE 2** Results of testing on the AFLW dataset.

Method	Unsupervised	Mean error
MTCNN [3]	No	6.90
TCDCN [4]	No	7.65
Cascaded CNN [5]	No	8.97
RCPR [11]	No	11.6
Sparse [12]	Yes	10.53
Dense 3D [9]	Yes	10.99
Ours	Yes	7.19

Abbreviations: AFLW, Annotated Facial Landmarks in the Wild; CNN, convolutional neural network.



**TABLE 3** Results of testing on the 300W dataset.

Method	Unsupervised	Mean error
TCDCN [4]	No	5.54
Wing loss [33]	No	4.04
BRULÈ [32]	Yes	8.9
Sparse [12]	Yes	7.97
Dense 3D [9]	Yes	8.23
Ours	Yes	4.93

Since the current research on extracting landmarks using unsupervised methods mainly focuses on CNN-based models, we compare our method with CNN-based models. Our method has only 2.49% and 7.19% errors on the CelebA dataset and AFLW dataset respectively, which has surpassed many important unsupervised algorithm models and supervised algorithm models. On the 300W dataset, it also achieved an error of 4.93%, surpassing these unsupervised algorithm models and approaching the effect of supervised models. The experimental results show that our proposed algorithm for unsupervised landmark extraction with the introduction of the transformer structure has great advantages over the unsupervised landmark extraction algorithm based on CNN.

### 4.3 | Ablation

The internal self-attention mechanism of Transformer can capture and encode landmarks well. The introduction of contrastive loss can make the distance between semantically consistent landmarks closer, and the distance between semantically inconsistent landmarks is more distant. Furthermore, it can prevent the collapse of the model. In order to verify the effectiveness and stability of our proposed algorithm, we replaced the components in the new algorithm and carried out ablation experiments. Meanwhile, experiments are made to make some hyperparameters more reasonable and effective as well.

On the CelebA dataset, we replaced the structure of the landmark extraction part with the Unet network structure in these methods [16, 18] and calculate the mean error of the corresponding algorithm to prove the effectiveness of the Transformer structure. In order to verify the rationality of the contrastive loss, we removed the contrastive loss during ablation experiments, and only retained the perceptual loss between the image and the reconstructed image.

It can be seen from the data in Table 4 that the error of the landmarks extracted by the whole network reaches 5.03% without introducing the Encoder structure of the Transformer and the contrastive loss. When the Transformer structure is introduced and the contrastive loss is not introduced, the error reaches 4.45%. The error is 4.42% when the contrastive loss is introduced, and the structure of Transformer is not introduced. This is far from our result of 2.49%, and it also shows that Transformer structure and contrastive loss can play a positive role in the convergence of the final network.

**TABLE 4** Ablation experiments on CelebA dataset.

	Transformer structure	Contrastive loss	Error
Network	×	×	5.03
	×	✓	4.45
	✓	×	4.42
	✓	✓	2.49

**TABLE 5** Params of models.

Model	Params	FLOPs	Error
Unet	73.2M	352.8G	4.45
Mixed network	139.1M	472.2G	2.49

Abbreviation: FLOPs, floating point operations.

Our proposed model experiment results show that the internal self-attention mechanism in the Transformer structure can more effectively represent the landmarks of the object, but it is worth discussing whether the better performance of the new algorithm will lead to the increase of model parameters and the slower reference speed. Therefore, we calculate the parameters and floating point operations (FLOPs) of the model using the mixed network with Transformer structure and Unet for information extraction on single GPU. From the results in Table 5, the Transformer structure brings better landmark extraction ability, but the parameters and FLOPs also increase.

The loss of the whole algorithm is composed of perception loss and contrastive loss. The overall expression is  $\text{loss} = \lambda \times \text{perceptual loss} + \mu \times \text{contrastive loss}$ .

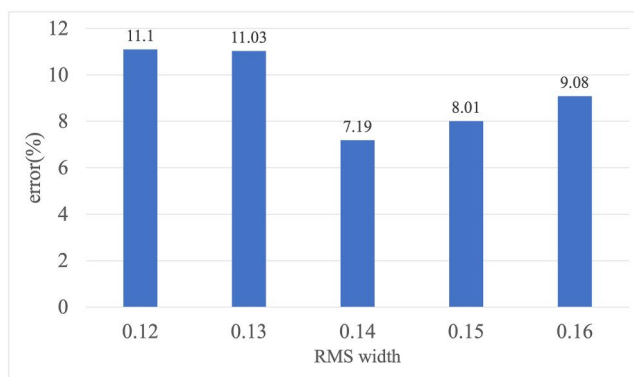
The introduction of contrastive loss makes the whole algorithm become multi-task training. Because of the different magnitude of the two losses, directly adding the perceived loss and the comparative loss to get the overall loss may lead to the whole training process being dominated by a loss or learning bias. Therefore, in order to balance the two losses, we need to design experiments to determine the number of the two hyperparameters. During the experiment, we set the hyperparameters  $\alpha$  and  $\beta$  of the loss function to the data shown in Table 6 and calculate the mean errors. From the experimental results in Table 6, we can see that when  $\alpha$  is 0.33 and  $\beta$  is 0.67, the performance of the whole algorithm is optimal, so the loss function of the whole algorithm is  $\text{loss} = 0.33 \times \text{perceptual loss} + 0.67 \times \text{contrastive loss}$ .

During the experiment, we found that the root mean square (RMS) width has a great influence on the final contrastive loss function when constructing the landmark heatmap. The size of the RMS width reflects how much the reconstructed image contains the neighbourhood information of the landmarks of the image  $xa$ . If the RMS width value is too small, it means that the landmarks information of the image  $xa$  does not help the reconstructed image  $\tilde{x}$ . If the RMS width value is too large, each landmark neighbourhood of the image  $xa$  may contain information of other landmarks, which

**TABLE 6** Results of different loss function hyperparameters.

Reconstruct loss	Contrastive loss	Error
0.83	0.17	2.70
0.8	0.2	2.89
0.75	0.25	2.78
0.67	0.33	2.80
0.5	0.5	2.69
0.33	0.67	<b>2.49</b>
0.25	0.75	2.76
0.2	0.8	2.86
0.17	0.83	2.79

Note: The bold value means the value with the smallest error.

**FIGURE 7** The result of error changing with RMS width. RMS, root mean square.

will affect the reconstruction. Therefore, we conducted an ablation experiment on the AFLW dataset to discuss the optimal RMS width value. First, we keep the network structure and hyperparameters unchanged, only modify the size of the RMS width value, then restart the training, and finally plot the error obtained for each RMS width value as a line figure. It can be seen from the results in Figure 7 that the RMS width value of 0.14 is the most suitable, and the RMS width value continues to increase or decrease, which will lead to the increase of the final experimental error.

## 5 | CONCLUSION

In this paper, we propose an unsupervised algorithm to extract landmarks of objects. The algorithm introduces the Transformer structure that can capture the location of landmarks well and encode the landmarks. At the same time, the proposed structure constructs positive and negative sample pairs between landmarks, which maximises the similarity between the same landmarks and the dissimilarity between landmarks within different batches. The results show that our algorithm performs well on many datasets, and it performs better than many unsupervised algorithms and even exceeds many

supervised algorithms. Finally, we have made relevant experiments to verify the effectiveness of the proposed algorithm and obtain reasonable values of the hyperparameters.

## AUTHOR CONTRIBUTIONS

**Tao Wu:** Writing – original draft; writing – review and editing. **Wenzhuo Fan:** Investigation. **Shuxian Li:** Investigation. **Qingqing Li:** Investigation. **Jianlin Zhang:** Project administration. **Meihui Li:** Project administration.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>, <https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/>, and <https://ibug.doc.ic.ac.uk/resources/300-W/>.

## ORCID

Tao Wu  <https://orcid.org/0000-0001-5999-9064>

## REFERENCES

1. Cootes, T., Baldock, E.R., Graham, J.: An introduction to active shape models. *Image Process. Anal.* 328, 223–248 (2000)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(6), 681–685 (2001). <https://doi.org/10.1109/34.927467>
3. Zhang, Z., et al.: Facial landmark detection by deep multi-task learning. In: *European Conference on Computer Vision*, Zurich, Switzerland, pp. 94–108 (2014)
4. Zhang, Z., et al.: Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(5), 918–930 (2015). <https://doi.org/10.1109/tpami.2015.2469286>
5. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 3476–3483 (2013)
6. Zhang, J., et al.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: *European Conference on Computer Vision*, Zurich, Switzerland, pp. 1–16 (2014)
7. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised object learning from dense equivariant image labelling. In: *NIPS*, Long Beach (2017)
8. Shu, Z., et al.: Deforming autoencoders: unsupervised disentangling of shape and appearance. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 650–665 (2018)
9. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object frames by dense equivariant image labelling. In: *Neural Information Processing Systems*, Long Beach, USA (2017)
10. Mallis, D., et al.: From keypoints to object landmarks via self-training correspondence: a novel approach to unsupervised landmark discovery. *arXiv preprint arXiv: 2205.15895* (2022)
11. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, pp. 1513–1520 (2013)
12. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 5916–5925 (2017)
13. Guo, X., et al.: PFLD: a practical facial landmark detector. *arXiv preprint arXiv: 1902.10859* (2019)

14. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1653–1660 (2014)
15. Minderer, M., et al.: Unsupervised learning of object structure and dynamics from videos. In: *Neural Information Processing Systems*, Vancouver, Canada (2019)
16. Daniel, T., Tamar, A.: Unsupervised image representation learning with deep latent particles. *arXiv preprint arXiv: 2205.15821* (2022)
17. Jakab, T., et al.: Unsupervised learning of object landmarks through conditional image generation. In: *Neural Information Processing Systems*, Montréal, Canada (2018)
18. Kulkarni, T.D., et al.: Unsupervised learning of object keypoints for perception and control. In: *Neural Information Processing Systems*, Vancouver, Canada (2019)
19. Vaswani, A., et al.: Attention is all you need. In: *Neural Information Processing Systems*, Long Beach, USA (2017)
20. Chen, T., et al.: A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, Vienna, Austria, pp. 1597–1607 (2020)
21. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929* (2020)
22. Touvron, H., et al.: Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning*, pp. 10347–10357 (2021)
23. Han, K., et al.: Transformer in transformer. *Adv. Neural Inf. Process. Syst.* 34, 15908–15919 (2021)
24. Carion, N., et al.: End-to-end object detection with transformers. In: *European Conference on Computer Vision*, Glasgow, KY, USA, pp. 213–229 (2020)
25. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890 (2021)
26. Snower, M., et al.: 15 keypoints is all you need. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6738–6748, (2020)
27. Yang, S., et al.: Transpose: keypoint localization via transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 11802–11812 (2021)
28. Zhang, Y., et al.: Unsupervised discovery of object landmarks as structural representations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2694–2703 (2018)
29. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *European Conference on Computer Vision*, Amsterdam, Netherlands, pp. 694–711 (2016)
30. Choudhury, S., et al.: Unsupervised part discovery from contrastive reconstruction. *Adv. Neural Inf. Process. Syst.* 34, 28104–28118 (2021)
31. Huang, C., et al.: Self-supervised masking for unsupervised anomaly detection and localization. *IEEE Trans. Multimedia*, 1–1 (2022). <https://doi.org/10.1109/TMM.2022.3175611>
32. Bupalov, I., Buzun, N., Dyllov, D.V.: BRULÉ: barycenter-regularized unsupervised landmark extraction. *Pattern Recogn.* 131, 108816 (2022). <https://doi.org/10.1016/j.patcog.2022.108816>
33. Feng, Z.H., et al.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2235–2245 (2018)

**How to cite this article:** Wu, T., et al.: Unsupervised detection of contrast enhanced highlight landmarks. *IET Comput. Vis.* 1–10 (2023). <https://doi.org/10.1049/cvi2.12197>