MEnet: A Metric Expression Network for Salient Object Segmentation

Shulian Cai[1*], Jiabin Huang[1*], Delu Zeng[2†], Xinghao Ding[1], John Paisley[3]
[1] Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, China
[2] School of Mathematics, South China University of Technology, China
[3] Department of Electrical Engineering, Columbia University, USA
{liuxicaicai,huangjiabin}@stu.xmu.edu.cn, dlzeng@scut.edu.cn, dxh@xmu.edu.cn, jpaisley@columbia.edu

Abstract

Recent CNN-based saliency models have achieved excellent performance on public datasets, but most are sensitive to distortions from noise or compression. In this paper, we propose an end-to-end generic salient object segmentation model called Metric Expression Network (MEnet) to overcome this drawback. We construct a topological metric space where the implicit metric is determined by a deep network. In this latent space, we can group pixels within an observed image semantically into two regions, based on whether they are in a salient region or a non-salient region in the image. We carry out all feature extractions at the pixel level, which makes the output boundaries of the salient object finely-grained. Experimental results show that the proposed metric can generate robust salient maps that allow for object segmentation. By testing the method on several public benchmarks, we show that the performance of MEnet achieves excellent results. We also demonstrate that the proposed method outperforms previous CNN-based methods on distorted images.

1 Introduction

Image saliency detection and segmentation is of significant interest in the fields of computer vision and pattern recognition. Recent saliency detection studies can be divided into two categories: those based on *hand-crafted features* and *learning-based* approaches. In previous literature, the majority of saliency detection methods have used hand-crafted features. Traditional low-level features for such saliency detection models mainly consist of color, intensity, texture and structure [Yang *et al.*, 2013; Cheng *et al.*, 2015; Borji and Itti, 2012]. Though hand-crafted features with heuristic priors perform well in simple scenes, they are not robust to more challenging cases, such as when salient regions have similar color to background.

Learning-based methods, in particular using convolutional neural networks (CNNs) [LeCun *et al.*, 1998] have been proposed to address the shortcomings of using hand-crafted features for saliency detection. For example, [Wang *et al.*, 2017] uses a multi-stage refinement mechanism to effectively combine high-level object semantics with low-level image features to produce high-resolution saliency maps, while [Luo *et al.*, 2017; Liu and Han, 2016; Zhang *et al.*, 2017a] exploit multi-level and multi-scale convolutional features for object segmentation. But even though they obtain good performance, CNN-based approaches also have room for improvement in their robustness to distorted scenes and to other common distortions such as noise [Chen *et al.*, 2017].

Metric learning is an area receiving much attention in computer vision, such for image segmentation [Fathi *et al.*, 2017], face recognition [Hu *et al.*, 2014] and human identification [Yi *et al.*, 2014], as a way for measuring similarity between objects. Inspired by the metric learning framework, we propose a deep metric learning architecture for image saliency segmentation that is also robust to potential distortions within an image. Our goal is to learn a metric space containing semantic features using a deep CNN such that two homogeneous sections of this space are learned for the salient and non-salient regions of the image space.

These features are learned at the pixel level and allow for distinguishing between salient regions and background using a distance measure. Simultaneously, we introduce a metric loss function based on metric learning and cross entropy. We also use multi-level information for feature extraction, similar to other approaches, such as Hypercolumns [Hariharan *et al.*, 2015] and U-net [Ronneberger *et al.*, 2015].

We experiment on several benchmark data sets and show how our proposed approach achieves results at state-of-art level. Moreover, we show how the proposed model is robust to distortions within an image.

2 A Metric Expression Network (MEnet)

We illustrate our proposed model architecture MEnet in Figure 1. As shown there, an encoder-decoder CNN first generates feature maps at different scales (blocks), which through convolution and up-sampling gives a feature vector for each pixel of an image according to how it maps through the layers. These extracted features are then used in a combined metric loss and cross entropy function to learn the salient regions as described below. We first discuss the encoder-decoder CNN

*The co-first authors contributed equally.
†Corresponding author: dlzeng@scut.edu.cn

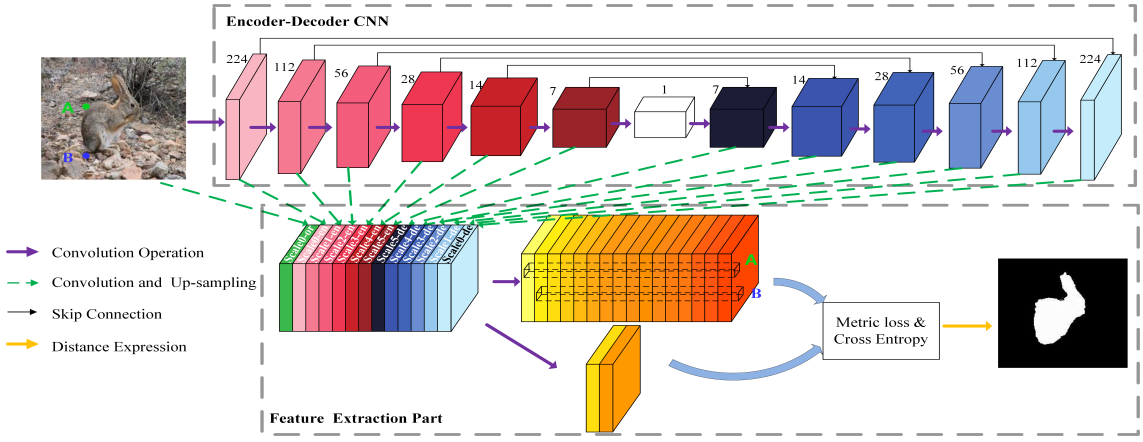


Figure 1: The proposed framework for saliency segmentation.

followed by our loss function and finally our semantic distance measure.

2.1 Encoder-Decoder CNN for Feature Extraction

Saliency segmentation usually requires global information about the image [Wang *et al.*, 2015], and thus multi-scale information is beneficial for more precise image segmentation. To learn this global information with deep learning, we use convolutions and pooling layers to increase the receptive field of the model and compress all feature information into feature maps whose size are 1×1, as shown as the white box in Figure 1. For multi-scale information, in previous approaches such as SegNet [Badrinarayanan *et al.*, 2017] and U-net, the encoder-decoder is used to extract multi-scale features. Here we use a similar structure. Through the decoder module, we up-sample these feature maps and view the feature map at each scale as representing information at a certain semantic level. We propose a symmetric encoder-decoder CNN architecture to extract global and multi-scale feature maps.

The encoder-decoder network of Figure 1 uses a deep symmetric CNN architecture with skip connections as indicated by black arrows. It consists of an encoder half and a decoder half, each block of which contains one of the two basic blocks shown in Figure 2. For encoding, at each down-sampling step we double the number of feature channels using a convolution with stride 2. For decoding, each step in the decoder path consists of an up-sampling of the feature map by a deconvolution after concatenating the input with the skip connection, also with stride 2. This part is similar to U-Net, but the difference is that U-net is designed for image segmentation, which is objective and works well even with cropped feature maps. For saliency segmentation, it is subjective and easily affected in different scenarios. Thus, global information is of significant importance to salient object segmentation. We maintain the size of the feature map to make full use of all the information in the larger receptive field. Our goal in using a symmetric CNN is to generate different scales of feature maps, which are concatenated to obtain feature vectors for each corresponding pixel in the input image that contain multi-scale information across the channel dimension. For instance, previous work in this direction showed that deep CNNs can learn such a feature representation that captures local and global context information for saliency segmentation [Zhao *et al.*, 2015].

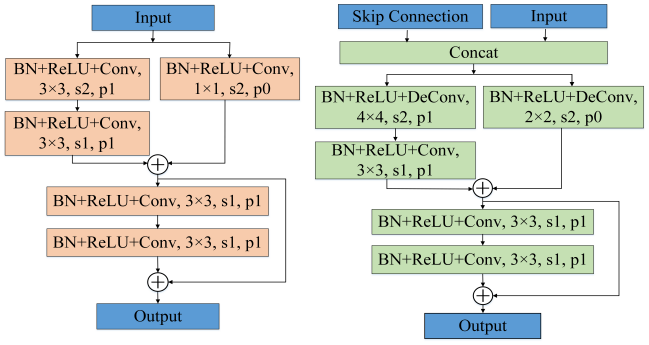


Figure 2: Basic encoder (left) and decoder (right) blocks.

We ultimately want to distinguish salient objects from background and so want to map image pixels into a feature space where that distance across salient and background regions is large, but within regions is small. Therefore, as shown in Figure 1, we can convert the 13 different scales of the encoder-decoder network into a set of feature vectors as indicated by the green dashed lines. That is, in the feature extraction part, each scale generates one output feature map of the same size via a single convolution and up-sampling; while the first "feature map" is simply obtained from convolving the original image across its RGB channels. Though the proposed algorithm is similar to the Hypercolumns model, one difference is that when training, the Hypercolumns model predicts heatmaps from feature maps of different scales by stacking additional convolutional layers. Hypercolumns is more like DHSNet [Liu and Han, 2016] which uses multi-scale saliency labels for segmentation. Instead, MEnet up-samples each scale of feature map to the same size during training. Another difference is that, where Hypercolumns classifies each category at separate layers, MEnet integrates the multi-scale feature maps for these tasks. As these 13 features may have unequal information value, learn this with another convolutional filter of these 13 feature maps. After concatenating the feature maps at each level we further use

convolutional operations with 16 kernels to generate the final feature maps. We incorporate cross entropy to help with this task, as described in the following section. In this case, the final feature vector is in $\mathbb{R}^{16}$.

2.2 Loss Function

Many previous works on saliency detection based on deep learning uses cross entropy (CE) to optimize the network [Li and Yu, 2016; Luo *et al.*, 2017]. This loss function is written as follows:

$$L_{CE}(l|\theta_1) = -\frac{1}{N \times |\Omega|} \times \sum_{n=1}^{N} \sum_{i=1}^{|\Omega|} \sum_{y=0}^{1} 1\{l_i^{(n)} = y\} \ln P(l_i^{(n)} = y|\theta_1), \quad (1)$$

where θ_1 is the set of learnable parameters of network, Ω is the pixel domain of the image, L_{CE} denotes the loss of the entire training set, N is the number of training data, $\mathbf{1}\{\cdot\}$ is the indicator function, and $y \in \{0, 1\}$, where $y = 1$ denotes the salient pixel and $y = 0$ denotes the non-salient pixel. $P(l_i^{(n)} = y|\theta_1)$ is the label probability of the i-th pixel predicted by network. In MEnet, we generate $P(l_i^{(n)} = y|\theta_1)$ via a convolution with 2 kernels from feature extraction part as shown in Figure 1.

Metric learning has been widely used in computer vision tasks. For instance, in [Fathi *et al.*, 2017; Harley *et al.*, 2017], the idea of metric learning is applied to segmentation. However, in [Fathi *et al.*, 2017], only one scale of the input corresponds to one corresponding size feature map, while we propose to use more feature maps from different scales to generate the final saliency map. In [Harley *et al.*, 2017], local attention masks are constructed by pairwise distance computations from a neighborhood around each pixel, which may not be suitable for saliency segmentation. Therefore, we instead use the triplet loss to compute the global information. Our metric loss function (ML) is defined as in Equation 2. In our network, the input is an RGB image whose size is $H \times W \times 3$, and all the images are resized to $224 \times 224 \times 3$, hence $H = W = 224$ here. The output is a feature metric space which is generated by 16 kernel convolutions in Figure 1, and the size is $H \times W \times C$ (we set $C = 16$). Each pixel in the $H \times W$ image corresponds to a C-dimension vector in the salient feature map. The metric loss function is defined as following:

$$L_{ML}(f|\theta_2) = \frac{1}{N \times |\Omega|} \times \sum_{n=1}^{N} \sum_{i=1}^{|\Omega|} \left[\sum_{k \in set^+} \frac{\|f_i^{(n)} - f_k^{(n)}\|_2^2}{|set^+|} - \sum_{k \in set^-} \frac{\|f_i^{(n)} - f_k^{(n)}\|_2^2}{|set^-|} \right], \quad (2)$$

where θ_2 is the set of learnable parameters of network and $f_i^{(n)}$ denotes the feature vectors corresponding to the pixel in the n-th image of the training set. We denote $k \in set^+$ (or $k \in set^-$), with $\Omega = set^+ \cup set^-$, meaning that $f_k^{(n)}$ is the positive or negative feature vector of $f_i^{(n)}$, respectively. That is, either $f_i^{(n)}$ and $f_k^{(n)}$ are from the same region (salient or non-salient), otherwise, $f_k^{(n)}$ is from a different region from $f_i^{(n)}$. We use Euclidean distance to calculate the distance between two feature vectors.

This loss function in (2) encourages an encoder-decoder network that enlarges the distance between any pair of feature vectors having different saliency, and reduces the distance for those with the same saliency. This is equivalent to

$$L_{ML}^*(f|\theta_2) = \frac{1}{N \times |\Omega|} \sum_{n=1}^{N} \sum_{i=1}^{|\Omega|} \big(\|f_i^{(n)} - \bar{f}_+^{(n)}\|_2^2 - \|f_i^{(n)} - \bar{f}_-^{(n)}\|_2^2 \big), \quad (3)$$

where we average all $f_k^{(n)}$ in Equation 3 to get $\bar{f}_+^{(n)}$ and $\bar{f}_-^{(n)}$. That is $\bar{f}_+^{(n)}$ is the mean of all positive pixels from a single image, while $\bar{f}_-^{(n)}$ corresponds to all negative pixels. Intuitively, Equation 3 enforces that the feature vectors extracted from the same region be close to the center of that region while keeping away from the center of the other region in salient feature space. In this case, we can obtain a more robust distance evaluation between the salient object and background. We also add a second cross entropy loss function as a constraint which shares the same network architecture with the objective function and empirically we have observed that the combined results were significantly better than only using either the metric loss or the cross entropy loss alone. Therefore, our final loss function is defined as below:

$$L_{MEnet}(f, l|\theta) = L_{ML}^*(f|\theta_2) + \lambda L_{CE}(l|\theta_1), \quad (4)$$

where $\theta = \theta_1 \cup \theta_2$ and λ is set to 1 in our experiments.

2.3 Semantic Distance Expression

If we train the proposed MEnet to minimize the loss function $L_{MEnet}(\cdot)$, we will obtain a network T_{θ^*}, where θ^* is converged value of θ. Given an observed input image for testing, where the pixel domain is Ω, we usually describe pixel $i \in \Omega$ by its intensities I_i across the channels. But it is difficult to define the semantic distance by $d_{ij}^{I_\Omega} = d(I_i, I_j)$, e.g., by Euclidean distance $d_{ij} = \|I_i - I_j\|_2$. However, through transformation of T_{θ^*}, we will obtain the corresponding feature vectors $\{f_i\}_{i \in \Omega}$ to represent the input. Then the distance can be expressed as $d'_{ij} = d_{ij}^{T_{\theta^*}(I_\Omega)} = \|f_i - f_j\|_2$, and finally the saliency map S for saliency segmentation is obtained by:

$$S_i = \|f_i - E_{f_j \sim P_B(\cdot)} f_j\|_2 = \|f_i - \sum_{j \in \Omega_B} P_B(f_j) f_j\|_2, \quad (5)$$

where $P_B(\cdot)$ is the probability distribution of the feature vector $f_j \in \Omega_B$, and $\Omega = \Omega_B \cup \Omega_S$, where Ω_B and Ω_S denote the background region and salient region only computed from the component of L_{CE} in the loss function (4) within the converged network T_{θ^*}. We note that, Ω_B and Ω_S are not accurate segmentations and they are to be further investigated in the experimental section. To conclude, by network transformation we can express $d_{ij}^{I_\Omega}$ as $d_{ij}^{T_{\theta^*}(I_\Omega)}$. As illustrated in Figure 3, we anticipate that through this space transformation, the intra-class distance will be smaller than the inter-class distance.

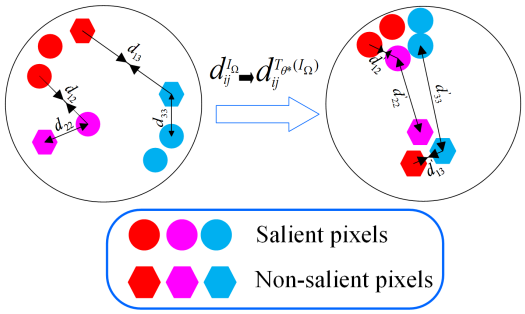


Figure 3: Idealized semantic distance expression with MEnet.

3 Experiments

We test our proposed MEnet on several public saliency datasets and distorted images, comparing with state-of-the-art saliency detection methods. We use the Caffe software package to train our model [Jia *et al.*, 2014].

3.1 Datasets

The datasets we consider are: MSRA10K [Cheng *et al.*, 2015], DUT-OMRON (DUT-O) [Yang *et al.*, 2013], HKU-IS [Li and Yu, 2015], ECSSD [Yan *et al.*, 2013], MSRA1K [Liu *et al.*, 2011] and SOD [Martin *et al.*, 2001]. MSRA10K contains 10000 images. It is the largest dataset and covers a large variety of content. HKU-IS contains 4447 images, most images containing two salient objects or multiple objects. ECSSD dataset contains 1000 images. DUT-OMRON contains 5168 images, which was originally designed for image segmentation. This dataset is very challenging since most of the images contain complex scenes. Existing saliency detection models have yet to achieve high accuracy on this dataset. MSRA1K including 1000 images, all belongs to the MSRA10K. SOD contains 300 images.

3.2 Training

We use stochastic gradient descent (SGD) for optimization, and the MSRA10K and HKU-IS are selected for training. For MSRA10K, 8500 images for training, 500 images for validation and the MSRA1K for testing; HKU-IS was divided into approximately 80/5/15 training-validation-testing splits. To prevent overfitting, all of our models use cropping and flipping images randomly as data augmentation. We use batch normalization [Ioffe and Szegedy, 2015] to speed up convergence.

All experiments are performed on a PC with Intel(R) Xeon(R) CPU I7-6900k, 96GB RAM and GTX TITAN X Pascal (12G). We use a 4 convolutional layer block in the upsample and downsample operations. Therefore the depth of our MEnet is 52 layers. The parameter sizes are shown in Figure 1 and Figure 2. We set the learning rate to 0.1 with weight decay of 10^{-8}, a momentum of 0.9 and a mini-batch size of 5. We train for 110,000 iterations. Since salient pixels and non-salient pixels are very imbalanced, network convergence to a good local optimum is challenging. Inspired by object detection methods such as SSD [Liu *et al.*, 2016], we adopt hard negative mining to address this problem. This sampling scheme ensures salient and non-salient sample ratio equal to 1, eliminating label bias.

3.3 Performance Comparison

We compare MEnet with 10 state-of-the-art models for saliency detection: MC [Zhao *et al.*, 2015], ELD [Wang *et al.*, 2015], DCL [Li and Yu, 2016], DHSNet [Liu and Han, 2016], DS [Li *et al.*, 2016], UCF [Zhang *et al.*, 2017b], Amulet [Zhang *et al.*, 2017a], SRM [Wang *et al.*, 2017], NLDF [Luo *et al.*, 2017], MSRNet [Li *et al.*, 2017] and 2 traditional metric learning methods: AML [Li *et al.*, 2015] and Lu's method [You *et al.*, 2016].

A visual comparison is shown in Figure 4 along with other state-of-the-art methods. MEnet performs better in these challenging scenes, e.g., when the salient region is similar to background. In addition, F-measure scores and MAE are shown in Table 1. We note that the better models (e.g., DHSNet, NLDF, Amulet, SRM, MSRNet and etc.) need pre-training and the conditional random field (CRF) method [Krähenbühl and Koltun, 2011] is used as post-processing in DCL and MSRNet. MEnet is trained from scratch and does not require pre/post-processing. It is still competitive with state-of-the-art models, particularly on the challenging datasets DUT-O and HKU-IS.

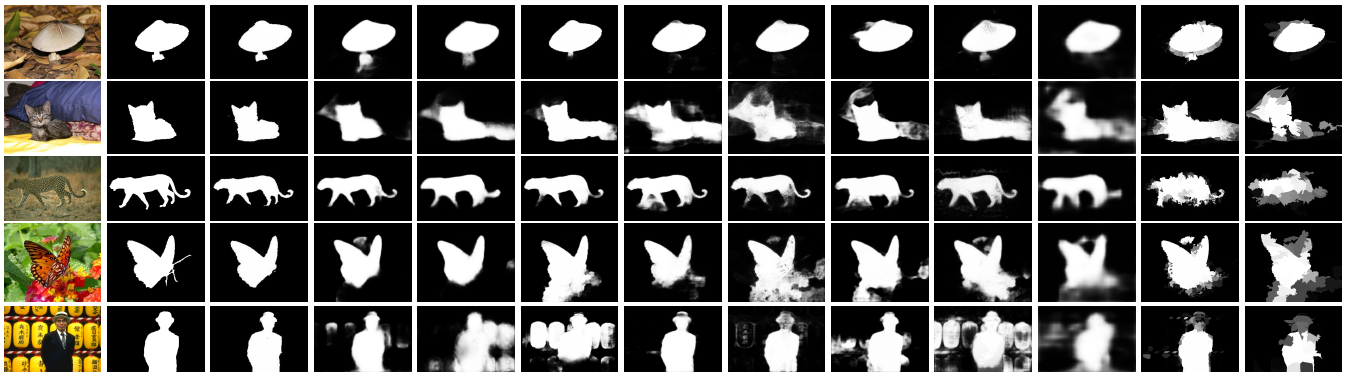
Table 2 shows the running times of the compared methods. For fair evaluation, the time efficiency of all models are performed on the same PC described above. It takes 86ms for our model to generate each saliency map with a GPU. Though our model is deeper, our test time is comparable with the fast models.

Evaluation on Distorted Images

We also test the models on distorted images. We note that MEnet does not train on distorted images for this case, as similar to previous works. During testing, the trained models are then directly tested on distorted images. To show the robustness of MEnet in this setting, we work with public datasets corrupted by Additive White Gaussian Noise (AWGN) and JPEG compression (with random strengths). For AWGN, we let the variance vary from 0.07 to 0.29, while for JPEG compression, we vary the quality factor from 3 to 6. We compare F-measure scores in Table 3. We can see that MEnet clearly outperforms other methods. Additionally, we show PR curves of our approach in Figure 5. Since the saliency maps generated by metric loss prediction tend to be binary, it is difficult to draw PR curves which need continuous salient values. Therefore, we select saliency maps generated by CE prediction to draw PR curves. In Figure 5, we observe that the performance of the proposed method is a little better than others on distorted datasets. As shown in Figure 7, the performance of other methods degrade rapidly with increasing noise, while MEnet still achieves robust performance. We believe reason for the robustness of MEnet owes to the fact that multi-scale features and metric loss are integrated into this structure, where features from either low or high levels can be fully utilized. In particular, we can see some evidence in [Du *et al.*, 2017] for denoising which uses an auto-encoder (similar to our Encoder-Decoder module) to obtain more robust features. A similar metric loss idea was shown to be ro-

	DUT-O		HKU-IS		ECSSD		MSRA1K		SOD	
	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$
Ours	0.732 •	0.074 ◦	0.879 •	0.044 •	0.880 ◦	0.060 ◦	0.928 •	0.028 •	0.594 ◦	0.139
SRM	0.718 ◦	0.071 •	0.877 ◦	0.046 ◦	0.892 •	0.056 •	0.894	0.045	0.617 •	0.120 •
MSRNet	0.695	0.074 ◦	–	–	0.868	0.056 •	0.903	0.036	0.579	0.124 ◦
NLDF	0.691	0.080	0.873	0.048	0.880 ◦	0.063	–	–	0.591	0.130
Amulet	0.654	0.098	0.841	0.052	0.873	0.060 ◦	–	–	0.550	0.160
UCF	0.645	0.132	0.820	0.072	0.854	0.078	–	–	0.557	0.186
DCL	0.660	0.095	0.844	0.063	0.857	0.078	0.922 ◦	0.035 ◦	0.573	0.147
DS	0.646	0.084	0.790	0.079	0.834	0.079	0.858	0.059	0.552	0.141
DHSNet	–	–	0.859	0.053	0.877	0.060 ◦	–	–	0.594	0.124 ◦
ELD	0.618	0.092	0.779	0.072	0.810	0.080	0.882	0.037	0.540	0.150
MC	0.622	0.094	0.733	0.099	0.779	0.106	0.885	0.044	0.497	0.160

Table 1: Comparison of quantitative results including F-measure (lager is better) and MAE (smaller is better). The top two results are indicated by • and ◦, respectively. DHSNet is trained on MSRA-B and DUT-O, MSRNet is trained on HKU-IS and MSRA-B, and UCF, Amulet and NLDF are all trained on MSRA-B dataset which contains MSRA1K. Therefore, we do not compare our model with these four models on these datasets.



(a) Images (b) GT (c) Ours (d) SRM (e) MSRNet (f) NLDF (g) Amulet (h) UCF (i) DHSNet (j) DCL (k) DS (l) ELD (m) MC

Figure 4: Visual comparisons with nine methods. MEnet can obtain detailed and accurate saliency maps.

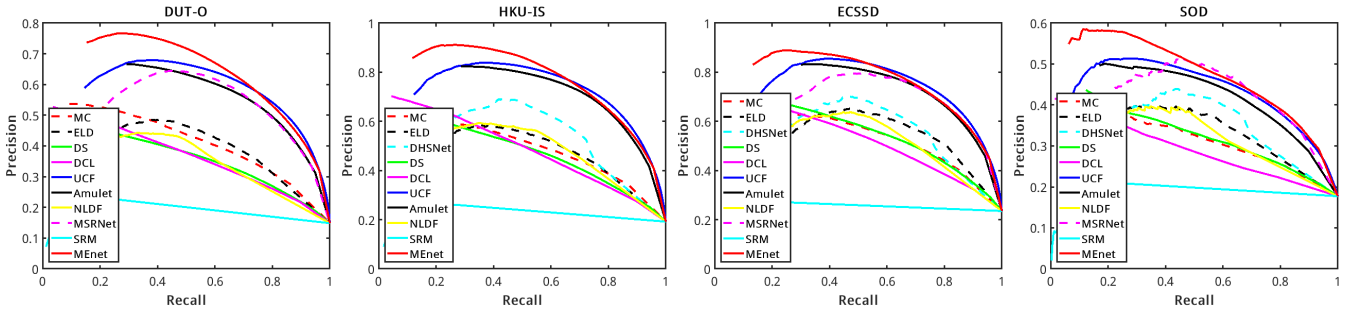


Figure 5: Comparison of precision-recall curves of other CNN-based methods on four datasets corrupted by AWGN (with random strengths).

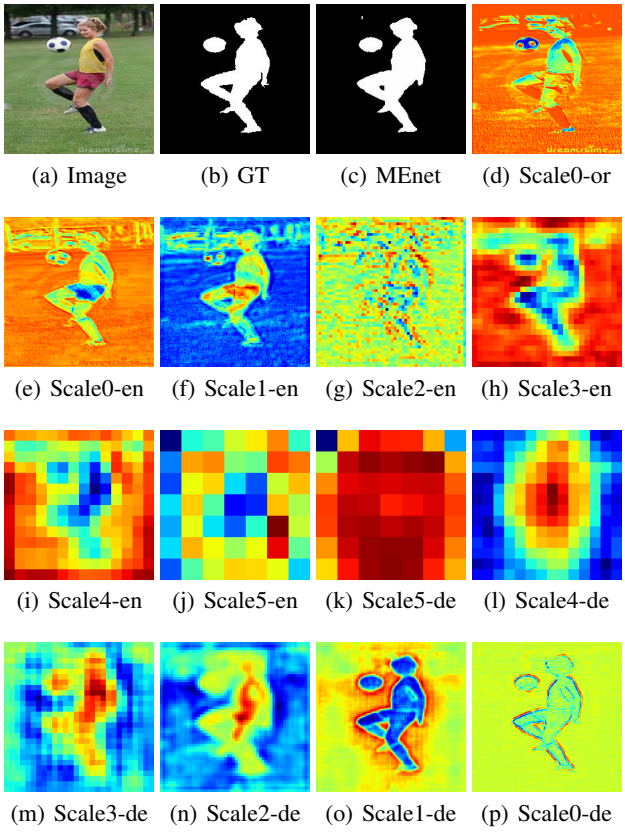
	Ours	SRM	MSRNet	NLDF	Amulet	UCF	DCL	DS	DHSNet	ELD	MC
s/img	0.086	0.091	4.678	0.071	0.061	0.111	0.53	0.104	0.019	0.78	1.8

Table 2: Running time of the compared methods.

bust to lighting conditions, deformation, and angle for human re-identification [Yi *et al.*, 2014], all of which can be regarded as "noise." Also, for vehicle re-identification, the distance similarity has been shown to provide vital information for robustly estimating the similarity among objects [Shen *et al.*, 2017]. In real-world scenes, images are easily impacted by noise and compression. Therefore, we consider the proposed work to be a more robust model.

	DUT-O		HKU-IS		ECSSD		MSRA1K		SOD	
	AWGN	JPEG	AWGN	JPEG	AWGN	JPEG	AWGN	JPEG	AWGN	JPEG
Ours	0.586 •	0.649 •	0.710 •	0.801 •	0.716 •	0.792 •	0.867 •	0.910 •	0.466 •	0.485 •
SRM	0.200	0.543	0.221	0.658	0.215	0.663	0.504	0.819	0.136	0.415
MSRNet	0.561 ◦	0.590 ◦	–	–	0.711 ◦	0.752	–	–	0.459 ◦	0.470 ◦
NLDF	0.402	0.561	0.531	0.700	0.565	0.693	–	–	0.352	0.433
Amulet	0.534	0.529	0.677 ◦	0.686	0.695	0.708	–	–	0.420	0.420
UCF	0.519	0.524	0.656	0.682	0.668	0.698	–	–	0.381	0.418
DCL	0.374	0.523	0.477	0.677	0.505	0.657	0.664	0.832	0.286	0.386
DS	0.368	0.497	0.477	0.611	0.532	0.649	0.619	0.771	0.313	0.405
DHSNet	–	–	0.605	0.735 ◦	0.622	0.753 ◦	–	–	0.394	0.461
ELD	0.454	0.548	0.531	0.686	0.603	0.730	0.737	0.841 ◦	0.376	0.444
MC	0.415	0.496	0.475	0.539	0.509	0.648	0.747 ◦	0.787	0.305	0.392

Table 3: Quantitative comparison with recent deep methods based on deep learning methods in difference distorted scenes via F-measure (lager is better). The top two results are indicated by • and ◦, respectively. DHSNet is trained on MSRA-B and DUT-O, MSRNet is trained on HKU-IS and MSRA-B, and UCF, Amulet and NLDF are all trained on MSRA-B dataset which contains MSRA1K. Therefore, we do not compare our model with these four models on these datasets. JPEG denotes JPEG Compression method.



(a) Image (b) GT (c) MEnet (d) Scale0-or

(e) Scale0-en (f) Scale1-en (g) Scale2-en (h) Scale3-en

(i) Scale4-en (j) Scale5-en (k) Scale5-de (l) Scale4-de

(m) Scale3-de (n) Scale2-de (o) Scale1-de (p) Scale0-de

Figure 6: Feature maps visualization, where (d)-(p) denote the different scale features as shown in Figure 1 (Feature Extraction Part).

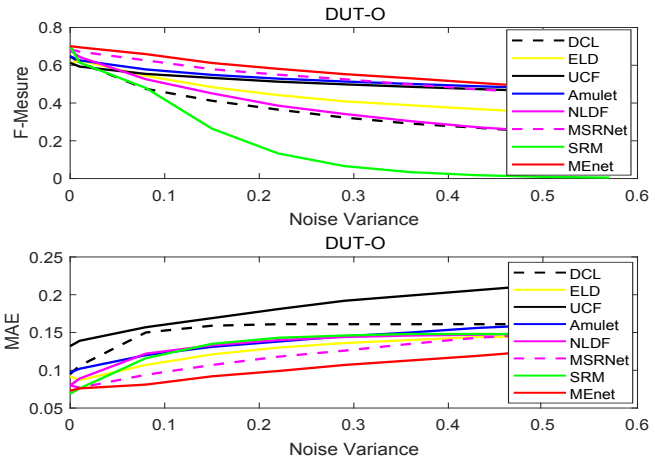


Figure 7: The curves of difference methods on DUT-O dataset under various noise variances.

Advantages of MEnet

In previous work, multi-scale features have been applied to produce saliency maps [Liu and Han, 2016; Zhang *et al.*, 2017a]. Although this is similar to our approach, there exist some differences in that these mentioned works predict saliency maps at each scale and so feature maps from the last layer of each scale may be similar. We propose to integrate multi-scale feature maps for classification and distance prediction. We only concatenate the feature maps to generate the final saliency maps.

To intuitively illustrate the advantage of MEnet, we select several feature maps for visualization. As we move to deeper layers, the receptive field of each neuron becomes larger. As shown in Figure 6, we observe that each convolutional layer contains different semantic information, and moving deeper allows the models to capture richer structures. Within the decoding part, Scale2-de, 3-de, 4-de are sensitive to the salient region, while Scale1-de has higher response against the background region. Other layers like Scale0-de can distinguish the boundary of salient objects.

Also, in [Liu and Han, 2016; Zhang *et al.*, 2017a], a convolutional layer with 1×1 kernels is used to fuse multi-scale features, which may lead to the receptive field being restricted. Instead of using 1×1 convolutions in the last layer, we instead use an $n \times n$ convolution in the last layer, containing more units to capture information from its neighborhood.

Data	Indexes	CE-plain	CE-only	MEnet
DUT-O	$F_\beta \uparrow$	0.631	0.678	**0.732**
	$MAE \downarrow$	0.098	0.084	**0.074**
HKU-IS	$F_\beta \uparrow$	0.803	0.872	**0.879**
	$MAE \downarrow$	0.064	0.056	**0.044**
ECSSD	$F_\beta \uparrow$	0.794	0.855	**0.880**
	$MAE \downarrow$	0.093	0.072	**0.060**
MSRA1K	$F_\beta \uparrow$	0.884	0.915	**0.928**
	$MAE \downarrow$	0.037	0.034	**0.028**
SOD	$F_\beta \uparrow$	0.525	0.555	**0.594**
	$MAE \downarrow$	0.156	0.159	**0.139**

Table 4: The performance of different strategies.

Data	Indexs	AML	Lu's	MEnet
ECSSD	$F_\beta \uparrow$	0.667	0.715	0.880
	$MAE \downarrow$	0.165	0.136	0.060
MSRA1K	$F_\beta \uparrow$	0.794	0.806	0.928
	$MAE \downarrow$	0.089	0.080	0.028

Table 5: Comparison with two traditional methods based on metric learning with F-measure and MAE scores.

To show the effectiveness of our proposed multi-scale feature extraction and loss function, we use different strategies for semantic saliency detection/segmentation as shown in Table 4. CE-only uses the cross entropy as its loss function, while CE-plain omits the feature extraction part and metric loss layer, and the loss layer is added directly to the decoder module in the framework. Therefore, the difference between CE-only and CE-plain is that CE-plain does not use multi-scale information which will lead to performance degradation. We also note that the performance of MEnet improves after introducing the metric loss. The multi-scale framework (encoder-decoder) and metric loss help make it feasible to distinguish saliency from background during training.

We compare MEnet with two other traditional metric learning methods for saliency segmentation, AML [Li *et al.*, 2015] and Lu [You *et al.*, 2016]. The results in Table 5 demonstrates the potential superiority of deep metric learning over traditional metric learning for semantic saliency segmentation.

4 Conclusion

In this paper, we present an end-to-end deep metric learning architecture called MEnet for salient object segmentation. We use multi-scale features extraction to obtain semantic information and combine with deep metric learning for mapping pixels into a "saliency space" where Euclidean distances can be used. The resulting mapping distinguishes salient image elements (pixels) from background efficiently. The proposed model is trained from scratch and does not require pre/post-processing. Experiments on benchmark datasets clearly demonstrate the effectiveness of our model, and robustness when handling distorted images.

Acknowledgments

This work was supported in part by grants from National Science Foundation of China (6151005, 61571382, 61103121, 81671766), the China Scholarship Council (201806155037), Guangdong Natural Science Foundation (2015A030313007, 2015A030313589), and the Science and Technology Research Program of Guangzhou, China (201804010429).

References

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[Borji and Itti, 2012] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 478–485. IEEE, 2012.

[Chen *et al.*, 2017] Zhuo Chen, Weisi Lin, Shiqi Wang, Long Xu, and Leida Li. Image quality assessment guided deep neural networks training. *arXiv preprint arXiv:1708.03880*, 2017.

[Cheng *et al.*, 2015] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[Du *et al.*, 2017] Bo Du, Wei Xiong, Jia Wu, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Stacked convolutional denoising auto-encoders for feature representation. *IEEE transactions on cybernetics*, 47(4):1017–1027, 2017.

[Fathi *et al.*, 2017] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017.

[Hariharan *et al.*, 2015] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[Harley *et al.*, 2017] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 7, 2017.

[Hu *et al.*, 2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings*

of the 22nd ACM international conference on Multimedia, pages 675–678. ACM, 2014.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5455–5463, 2015.

[Li and Yu, 2016] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

[Li *et al.*, 2015] Shuang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price. Adaptive metric learning for saliency detection. *IEEE Transactions on Image Processing*, 24(11):3321–3331, 2015.

[Li *et al.*, 2016] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930, 2016.

[Li *et al.*, 2017] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 247–256. IEEE, 2017.

[Liu and Han, 2016] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 678–686. IEEE, 2016.

[Liu *et al.*, 2011] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2011.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[Luo *et al.*, 2017] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *IEEE CVPR*, 2017.

[Martin *et al.*, 2001] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Shen *et al.*, 2017] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. *arXiv preprint arXiv:1708.03918*, 2017.

[Wang *et al.*, 2015] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3183–3192. IEEE, 2015.

[Wang *et al.*, 2017] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4019–4028, 2017.

[Yan *et al.*, 2013] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1155–1162. IEEE, 2013.

[Yang *et al.*, 2013] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3166–3173. IEEE, 2013.

[Yi *et al.*, 2014] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

[You *et al.*, 2016] Jia You, Lihe Zhang, Jinqing Qi, and Huchuan Lu. Salient object detection via point-to-set metric learning. *Pattern Recognition Letters*, 84:85–90, 2016.

[Zhang *et al.*, 2017a] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

[Zhang *et al.*, 2017b] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–221, 2017.

[Zhao *et al.*, 2015] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.