Multi-Type Self-Attention Guided Degraded Saliency Detection

Ziqi Zhou, Zheng Wang, Huchuan Lu, Song Wang, Meijun Sun

Abstract

Existing saliency detection techniques are sensitive to image quality and perform poorly on degraded images. In this paper, we systematically analyze the current status of the research on detecting salient objects from degraded images and then propose a new multi-type self-attention network, namely MSANet, for degraded saliency detection. The main contributions include: 1) Applying attention transfer learning to promote semantic detail perception and internal feature mining of the target network on degraded images; 2) Developing a multi-type self-attention mechanism to achieve the weight recalculation of multi-scale features. By computing global and local attention scores, we obtain the weighted features of different scales, effectively suppress the interference of noise and redundant information, and achieve a more complete boundary extraction. The proposed MSANet converts low-quality inputs to high-quality saliency maps directly in an end-to-end fashion. Experiments on seven widely-used datasets show that our approach produces good performance on both clear and degraded images.

Introduction

Saliency object detection (SOD) aims to distinguish the visually distinctive regions or objects and then segment the foreground targets out from the background. It plays an important role in content-based image/video understanding, such as visual tracking (Avytekin, Cricri, and Aksu 2018), and person re-ID (Zhao, Oyang, and Wang 2016).

Early SOD researches capture local details and global context based on hand-crafted features, e.g., color, texture, luminance (Harel, Koch, and Perona 2007). However, due to the lack of guidance at the semantic level, the developed methods perform poorly in many complex scenarios. Recently, thanks to the capability of extracting both low-level details and high-level semantic information simultaneously, convolutional neural networks (CNNs) have become the main force of SOD. However, many existing CNN-based SOD methods mainly focus on single image, and most of them are “picky-eaters”, which only accept clean and high-quality images. Nevertheless, due to the complexity of the natural environment and the diversity of acquisition methods, we may get many degraded images with lower quality in practice. Typical image degradations include e.g. low resolution, blur, noise, and atomization etc. On degraded images, the robustness of the SOD methods is inevitably influenced by losing information both inside and along the boundary of the objects of interest.

Similar to the discussion in (Guo et al. 2019), to address such a problem, a straightforward approach is to use the degraded image, denoted as \( x_d \), as adversarial examples, to form the training dataset. However, obvious gaps exist between \( x_o \) and \( x_d \) in terms of semantics and details and simple adversarial training cannot effectively improve the performance (see the performance of adv-Amulet in the later experiments). Another idea is to enhance the robustness by improving the quality of \( x_d \). However, given the diversity of degradation types and the randomicity of the degradation levels, general image restoration algorithms cannot handle all the situations well, not to mention that pre-processing of image restoration can be very time-consuming and storage demanding. Furthermore, such a pre-processing may not be combined to the end-to-end CNN training and this is usually not desirable in SOD by forwarding the errors between network modules.

In this paper, we propose a new MSANet for degraded salient object detection by incorporating multi-type self-
attentions, e.g. both global and local ones, which directly outputs the salient object segment from the degraded images in an end-to-end manner (Some visualization results are shown in figure.1). In particular, we introduce the “teacher-student” network to extract better features from degraded images. Under the guidance of the teacher network, our student net can implicitly capture the hidden mode of $x_d$ and extract more fully degraded features, thereby improving the accuracy of latter saliency inference. To sum up, the main contributions of this paper are:

- Systematically analyzing the current status of SOD on degraded images and developing a new MSANet, which greatly improves the SOD performance on degraded images without increasing the time cost.
- Proposing a multi-type self-attention network to select features from both global and local perspectives, thus to effectively suppress distracters while enhancing positive items.
- Proposing an attention transfer learning network to improve degraded SOD by narrowing the semantic gap between the target network and the source network.

Related work

Skip Connection Structure

Skip connection constructs a top-down path to transmit high-level semantic knowledge to shallower layers for enriched feature maps construction. For example, Zhang et al. (2017a) recursively embedded edge-ware low-level feature maps and the low-resolution predictions to promote boundary inference and semantic enhancement. Zhang et al. (2017b) fused high-level features to lower layers for acquiring adequate content information. For integration, they used a reformulated dropout to construct an uncertain ensemble of internal feature units. Zhang et al. (2018) designed symmetrical fully convolutional neural networks (FCNs) to learn complementary saliency features. For acceleration, Wu, Su, and Huang (2019) proposed a cascaded partial decoder framework, which discarded larger resolution features of shallower layers and directly utilized the generated saliency map to refine the features of backbone network.

However, such skip connection also passed through cluttered and noisy information. To solve this problem, some works introduced attention mechanism and recurrent structure. For example, Liu, Han, and Yang (2018) proposed a pixel-wise attention network. They generated attention maps for each pixel and selectively aggregated the contextual information to construct the attended contextual features. Liu et al. (2019) built two pooling-based modules, one module provided different layers the location information of potential salient objects, while the other merged features. Zhuge, Zeng, and Lu (2019) transformed prior information into an embedding space to select attentive features and to filter out outliers. Feng, Lu, and Ding (2019) proposed a boundary-enhanced loss for learning exquisite object boundaries. Deng et al. (2018) proposed a recurrent residual refinement network equipped with residual refinement blocks to detect salient regions. Wang et al. (2018b) constructed a global recurrent localization network and a local boundary refinement network. The former exploited contextual information of the weight map to avoid redundancy, while the latter learned the contextual message for each spatial position to recover object boundaries. Our approach also takes feature selection issues into account. The difference is that we are targeting degraded images. To this end, we first consider the completeness of features. And then, we propose a multi-type self-attention network to filter information from multiple perspectives.

Multi-Task Structure

Several studies employed multi-task strategies for rich supervision. e.g., Wang et al. (2018a) employed both saliency and segmentation labels for model training. Li et al. (2018) leveraged the contextual information provided by a well-trained contour extraction network to obtain fine boundary. Li et al. (2017) also trained using both contour and saliency labels. Qin et al. (2019) proposed a hybrid loss to learn from ground truth information in pixel-, patch- and map-levels. Moreover, Zhang et al. (2019) leveraged caption as auxiliary semantic knowledge to learn discriminative semantic features for salient objects. Different from the above models, MSANet refines the salient object boundary from the perspective of feature selection, and does not use any auxiliary supervision information.

Furthermore, all the models mentioned above are sensitive to image quality – they are vulnerable to degraded images. On the contrary, our MSANet learns more fully degraded features through attention transfer training so that it can adapt to various degraded situations and get more accurate degraded saliency maps.

MSANet

We choose VGGNet16 as the backbone network and discard the last pooling layer and dense layers to maintain the spatial structure. As shown in figure.2, the core ideas of MSANet are: 1) attention translation and 2) feature selection.

For the former, we propose an attention transfer network (ATN) for guidance learning, which helps to perceive the structural details and semantic content of degraded inputs. For the latter, we design a multi-type self-attention network (MSA), which implements pixel-level selection of features from multiple perspectives, so as to strengthen positive items while suppressing disturbances.

Attention Transfer Network (ATN)

When the input data is mixed with noise information (degraded), its distribution changes greatly compared with the clear one. The offset of data distribution will invalidate the feature extraction ability of the general network and cause the final detection to fail. Considering that the degraded image $x_d$ and the clear image $x_o$ largely share the same structural information, we design this attention transfer module (ATN) based on the “teacher-student” network to migrate the knowledge of the teacher network.

As shown in figure.3, ATN has a twin structure, which uses $\omega_t(\cdot)$ pre-trained on $x_o$ as the “teacher” net to guide the
Figure 2: The overall architecture of our MSANet, consisting of Attention Transfer Network (ATN), Recurrent Local Self-attention module (RLS), and Global Pixel Self-attention module (GPA). Details of RLS and GPA are shown in the right side, and the structure of ATN is shown in figure 3.

Figure 3: The architecture of attention transfer network.

learning process of the “student” net \( \omega_s(.) \). Since the higher features extracted from \( \omega_t(.) \) are semantic/task-related, it is reasonable to expect that the feature distribution of \( \omega_t(.) \) pre-trained on clean images should be similar to the feature distribution of \( \omega_s(.) \) fine-tuned using degraded images (Guo et al. 2019).

Specifically, let \{ \{x_d, x_o\}; y\} represents a pair of training samples, we use \{x_i; y\} to train \( \omega_t(.) \), and then, we use \{x_i; y\} to train \( \omega_s(.) \). In this phase, we freeze \( \omega_t(.) \) and use the encode features of \( \omega_t(x_o) \) as the guidance information to pilot the detail perception and semantical feature learning of \( \omega_s(.) \). Especially, we respectively calculate the spatial attention map \( M \) of \( \omega_t \) and \( \omega_s \), and minimize the attention transfer loss \( \mathcal{L}_T \). \( M \) decodes the attention of the current feature map, which is defined as \( M^i = \delta(\max_{j=1,2} B^j_i) \), \( B_i \) is the feature of \( i \)-layer, \( \delta \) is softmax normalization. The \( \mathcal{L}_T \) can be represented as:

\[
\mathcal{L}_T^t(\omega_t; x_d, x_o) = \|M_{\omega_s}^t - M_{\omega_t}^t\|^2_2 \\
\mathcal{L}_T^s(\omega_t; x_d, x_o) = \sum_{i \in I} \mathcal{L}_T^i(\omega_t; x_d, x_o)
\]

where \( M_{\omega_s} \) and \( M_{\omega_t} \) is the spatial attention map of \( \omega_s \), \( \omega_t \) respectively; \( I = \{2, 3, 4, 5\} \).

In order to minimize the objective function \( \mathcal{L}_T \), we constantly update the parameters of \( \omega_t \) and find a relatively better solution \( \omega_t^* \) to achieve the purpose of attention migration.

\[
\omega_t^* \leftarrow \arg\min_{\omega_t} \mathcal{L}_T^t(\omega_t; x_d, x_o)
\]

During training, the parameters of \( \omega_s \) is frozen and not be updated; during testing, we close \( \omega_s \) and only used the results of \( \omega_t^* \).

Multi-type Self-Attention (MSA)

We calculate the multi-type self-attention score to re-determine the pixel weights. Specifically, we respectively implement recurrent local self-attention (RLS) estimation and global pixel self-attention (GPA) inference.

For the four encoding blocks \( B_2 \sim 5 \), we perform RLS on \( B_3 \sim 5 \) and GPA on \( B_2 \). The main reasons are: i. Deploying RLS on lower-level features reduces computational speed and has little improvement in accuracy; ii. GPA calculates pixel-level relationship, while such correlation remains more complete in shallow layers.

Recurrent Local Self-Attention (RLS). In order to fully explore local context associations, we design an attention-based convolutional gated recurrent units, i.e. ACG. Similar to ConvLSTM, the proposed ACG also employ gates for information selection and integration (figure 2). Specifically, ACG consists of two convolutional gates (reset, update) and one hidden state \( (h) \). Unlike ConvLSTM, we weight the input information on each iteration by an attention layer, which enables more accurate local optimization of features while reducing the number of parameters.

As shown in figure 4, RLS consists of three layers: a convolution layer and two ACG layers. Convolution for channel formation selection and integration (figure 2). Specifically, ACG requires 4-D input data \( x_{acg}(t, w, h, c) \). Let the scaled feature is \( x_{org} \), \( x_{acg} \) is a set of copies of \( x_{org} \), and the number of copies is determined by \( t \), that means, at each time step \( t \), ACG accepts two sets of inputs, the old hidden state data \( h_{t-1} \), and the new input data \( x_t \).
Here, $x_t$ is constant, that is, it is always $x_{org}$, and the hidden state $h_t$ is continuously updated, which selectively fuses information from $x_t$ and $h_{t-1}$ through the gating structure, thus to locally optimize $x_{org}$. After a $t$ times optimization, the final $h_t$ is output as the optimized feature $x_{RLS}$.

Given the input feature $x_t$ ($x_{org}$) at time step $t$, the status update process is driven by the following formula:

$$
\bar{x}_t = x_t \ast (I + c \text{gap}(x_t)/\Sigma c \text{gap}(x_t))
$$

(4)

$$
r_t = \sigma(w_{xi} \ast \bar{x}_t + w_{hr} \ast h_{t-1} + b_t)
$$

(5)

$$
h_t = \tanh(r_t \ast w_{hh} \ast h_{t-1} + w_{xi} \ast \bar{x}_t + b_h)
$$

(6)

$$
z_t = \sigma(w_{xz} \ast \bar{x}_t + w_{hz} \ast h_{t-1} + b_z)
$$

(7)

$$
h_t = (1 - z_t) \ast \bar{h}_t + z_t \ast h_{t-1}
$$

(8)

where $I$ is a matrix with all values of 1, $\ast$ is the convolution operation, $\ast$ is pixel-wise multiplication, $X_{org}$ represents the weighted input. We discuss the setting of $t$ in table 1.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$F_p$</th>
<th>$\text{MAE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8925</td>
<td>0.0517</td>
</tr>
<tr>
<td>3</td>
<td>0.9008</td>
<td>0.0517</td>
</tr>
<tr>
<td>4</td>
<td>0.9015</td>
<td>0.0517</td>
</tr>
<tr>
<td>5</td>
<td>0.9010</td>
<td>0.0517</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.0578</td>
</tr>
</tbody>
</table>

Table 1: Performance evaluation results of ACG on ECSSD under different step settings. Considering the computational cost, we finally choose $t=3$ to achieve the best balance.

We define RLS as a local mode because there is no explicit global information-based attention calculating process in the feature optimization process. The maintaining or suppressing of each feature point is implicitly completed. The visualized features in figure 5 show that the deployment of the RLS effectively captures the boundary information and keeps the attention on the foreground area.

Figure 5: Visualization of B3-5 features of some channels before and after RLS implementation.

**Global Pixel Self-Attention (GPA).** GPA explicitly models channel attention (CA) and spatial attention (SA), the score for each position is obtained by comparison with all other positions, so GPA is a global mode.

For feature $B_i^{w \times h \times c}$, we firstly perform CA. Since each channel focus on different feature, it is necessary to highlight those channels that focus on the foreground objects. We calculate max and mean values simultaneously to obtain soft attention:

$$
c(B_i) = \delta^1(\text{AP}_c(B_i)) + \delta^2(\text{MP}_c(B_i))
$$

(9)

$\delta$ is a softmax normalization. When dealing with shallow features, each response can be thought as a detector for boundaries. Considering that $\delta^2$ returns only a single response, which focus on one discriminative part and ignore others, while $\delta^1$ encourages the detector to treat all locations on average, inevitably introducing noise. To this end, we calculate $c(B_i)$ to do a soft selection.

Then we calculate SA, which includes two items. The first item is similar to CA, which computes the spatial mean matrix and employs softmax for normalization. The second item $\text{LPPool}$ solves the local patch similarity. We scale the channels to 1 by dot-convolution and use average pooling ($2\times2$) to get the representative value of each patch, thus to ensure that the attention score of each pixel is calculated both locally and globally.

$$
s(B_i) = \delta(\text{AP}_s(B_i)) + \sigma(\text{LPPool}(B_i))
$$

(10)

Where $\sigma$ is a sigmoid function. It should be noted that we use softmax in spatial average pooling and use sigmoid in local-patch attention calculation, because the response of single position should be independent (sum to 1) while patch-wise response is related with others.

Finally, the global weighted feature $\bar{B}_i$ is calculated by:

$$
\bar{B}_i = B_i \ast (I + c(B_i)) \ast (I + s(B_i))
$$

(11)

where $c(B_i)$ and $s(B_i)$ is the channel and spatial attention scores respectively. $I$ is a matrix with all values of 1. We first employ CA and then employ SA.

**Joint Loss**

Our MSANet calculates multiple losses. Let $S_i$ represents the prediction, $G_i$ is the ground-truth. The total loss $L_{total}$ includes: i. Binary Cross-entropy loss $L_{bce}$. ii. Structural Similarity loss $L_{ssim}$, which helps to learn the structural information of $G_i$. iii. Attention Transfer loss $L_{AT}$, which promotes the model to obtain more powerful feature representations. $L_{total}$ is expressed as:

$$
L_{total} = \lambda_1 L_{bce} + \lambda_2 L_{ssim} + \lambda_3 L_{AT}
$$

(12)

where $L_{bce} = -\Sigma_i G_i \log S_i + (1 - G_i) \log(1 - S_i)$, $L_{ssim} = 1/N(1 - \text{SSIM}(G_i, S_i))$ and $\text{SSIM}()$ (Wang, Simoncelli, and Bovik 2003) is the standard structural similarity function. $\lambda_1$, $\lambda_2$ and $\lambda_3$ is set to 5, 10 and 2 respectively.

**Experiment**

**Implementation Details**

The whole architecture is built on the keras deep learning framework. We employ **DUTS-TR** (Wang et al. 2017) to train, which contains 10,553 images, and we perform image enhancement, i.e., reverse, mirroring to increase to 65k. All the experiments are run on NVIDIA Geforce GTX 1080Ti (11 GB memory) and i7-8700k cpu. The first five conv-blocks of MSANet are all initialized from VGGNet16, and the other parameters are all randomly assigned. All training and testing images are resized to 224×224. We use the Adam optimizer to train and the learning rate is initialized to 1e-5.
Datasets

Seven large benchmarks are used for evaluation, including DUT-OMRON (Yang et al. 2013), ECSSD (Shi et al. 2016), HKU-IS (Li and Yu 2016), PASCAL-S (Li et al. 2014), SOD (Movahedi and Eldar 2010), MSRA-B (Jiang et al. 2013), and DUTS-TE (Wang et al. 2017).

For realistic degraded simulation, we perform multi-type and multi-scale degradation on clear images in table 2. During training, the synthesis process guarantees two random-nesses: random type and random scale, and for test, we representatively reported the experimental results in one fixed degraded condition under each effect.

Table 2: Details of the synthetic dataset. LR: low resolution, $r$ represents the resize scale; GB: gaussian blur, $h$ is the fuzzy kernel size; MB: motion blur, $l$ is the motion length; GN: Gaussian white noise with a mean of $m=0$ and a variance of $v$; HZ: haze and $e$ is the degree of haze.

<table>
<thead>
<tr>
<th>Scale factor</th>
<th>LR</th>
<th>GB</th>
<th>MB</th>
<th>GN</th>
<th>HZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale range</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>0.05</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Evaluation Metrics

Five evaluation criteria, including P-R curves, F-measure, mean absolute error (Borji et al. 2015), S-measure (Fan et al. 2018), and E-measure (Fan et al. 2018) are used to reflect the model performance.

Analysis

We compare our model with other 13 state-of-the-art CNN-based models, including AFNet (Feng, Lu, and Ding 2019), BASNet (Qin et al. 2019), CPDNet (Wu, Su, and Huang 2019), MWSNet (Zeng et al. 2019), C2SNet (Li et al. 2018), PiCANet (Liu, Han, and Yang 2018), DGRL (Wang et al. 2018b), Amulet (Zhang et al. 2017a), UCFF (Zhang et al. 2017b), WSS (Wang et al. 2017), and MRSNet (Li et al. 2017). To be fair, we use their released code with default settings to calculate the degraded saliency maps.

To fully illustrate the effect of proposed ATN in dealing with degraded images, we choose Amulet (Zhang et al. 2017a) for adversarial training, that is, based on the already trained model, we further use the hybrid dataset (mixed by degraded and clear images) for finetuning, and get adv-Amulet. The training process is in full compliance with the description in their paper to ensure fairness.

Quantitative Evaluation. The evaluated results are illustrated in tables 3-7, the best result is highlighted in **bold**. Since LFRNet, RFCN, and UCF were trained on MSRA10K and MSRA-B is a subset of MSRA10K, so we did not report their results on MSRA-B. The SE scores and PR curves are illustrated in figures 6 and 8.

According to the results shown in tables 3-7, the performance of most existing algorithms drops dramatically when dealing with degraded images. In particular, CPDNet (Wu, Su, and Huang 2019) and LFRNet (Zhang et al. 2018), which performed well on clear images, show a significant decrease in performance on degraded images. Especially in low-resolution scenes, whose performance has dropped by 37.5%, 25.0%, respectively. What’s more, the overall best performing MRSNet (Li et al. 2017) and C2SNet (Li et al. 2018) can basically maintain a stable level when processing low-resolution images, but the performance decreases drastically when handling blurred and noisy images. This is because both two models leverage contour information.

To a certain extent. However, such adversarial training has a
fixed format depending on the type of training data. Exces-
In this paper, we systematically studied the degraded SOD and proposed a multi-type self-attention network, MSANet. Through attention transfer learning, feature extractor can learn more abundant semantic details of degraded images and perceive their hidden feature patterns. Meanwhile, MSANet performs global and local attention inference on multi-scale features for information filtering. Extensive experiments were conducted to verify the effectiveness of our model. In the future work, we plan to further extend this work to detect video saliency under motion blurs, as well as SOD for small objects.

## Conclusion

This work was supported by the Natural Science Foundation of China under Grant 61572351, 61876125, 61772360, 61672376 and U1803264.

## Acknowledgments


---

### Qualitative Evaluation

Figure 7 provides some visualization of MSANet and other 12 methods in some challenging scenarios, where our MSANet is superior to other methods. For example, in low-resolution scenes, most of the models lose the position of salient objects (2nd). In blurring, foggy, and noisy scenes, many methods cannot accurately detect salient objects due to the side effects of interference information, or incorrectly regard the background as a salient object (3rd-6th), yet our proposed model is able to capture complete salient regions in a variety of complex scenarios.

### Ablation Study

The results in rows 6-7 of table 8 show that the effect of introducing ATN is better than that of adversarial training, which indicates that ATN can effectively promote the understanding of the content of degraded images and help to obtain more abundant features. Meanwhile, the results of lines 5-6 also show that our proposed MSA structure is better than Amulet under the same adversarial training mode. It is worth noting how to calculate and where to teach has a great impact on performance. According to the results in lines 1-4, after the introduction of RLS, the performance has increased by 2.4%-4.3%, which indicates the superiority of RLS in suppressing the interference of non-target regions. And the introduction of GPA further improves performance slightly, which indicates that GPA is helpful for optimizing the details of salient objects.

### Table 8: Ablation study on HKU-IS under different architectures

<table>
<thead>
<tr>
<th>Method</th>
<th>$S_2$</th>
<th>$E_{ps}$</th>
<th>$F_2$</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline En-De-V</td>
<td>869</td>
<td>391</td>
<td>.826</td>
<td>.078</td>
</tr>
<tr>
<td>†En-De-V+RLS</td>
<td>893</td>
<td>929</td>
<td>.849</td>
<td>.044</td>
</tr>
<tr>
<td>†En-De-V+GPA</td>
<td>882</td>
<td>920</td>
<td>.845</td>
<td>.049</td>
</tr>
<tr>
<td>†En-De-V+RLS+GPA</td>
<td><strong>894</strong></td>
<td><strong>942</strong></td>
<td><strong>.896</strong></td>
<td><strong>.045</strong></td>
</tr>
<tr>
<td>adv-Amulet</td>
<td>.746</td>
<td>.851</td>
<td>.764</td>
<td>.090</td>
</tr>
<tr>
<td>†adv-En-De-V+RLS+GPA</td>
<td>826</td>
<td>880</td>
<td>.778</td>
<td>.081</td>
</tr>
<tr>
<td>†En-De-V+RLS+GPA+ATN</td>
<td><strong>.835</strong></td>
<td><strong>.891</strong></td>
<td><strong>.795</strong></td>
<td><strong>.074</strong></td>
</tr>
</tbody>
</table>

Table 8: Ablation study on HKU-IS under different architectures. †: clear image, ‡: degraded image with LR scale r=4.

Figure 8: The precision-recall curves of 15 methods. The red curve represents our proposed MSANet. The number in parentheses represents the degradation scale, as detailed in Table 2.


