Infrared Small Target Detection Based on Multi-Scale Local Contrast Network

Yuanteng Liu and Yuehuan Wang*

1National Key Laboratory of Science and Technology on Multispectral Information Processing, Huazhong University of Science and Technology, Wuhan, China
*Corresponding author’s e-mail: yuehwang@hust.edu.cn

Abstract. To mitigate the challenge of the lack of the intrinsic features of the small target which possesses only few pixels, we designed a multi-scale local contrast network which combines features extracted from different scales and layers for single-frame infrared small target detection in this paper. We extract the multi-scale contrast features in the same layer by multi-scale local contrast module and fuse the cross-layer contrast features by local contrast fusion module. Moreover, we embed these modules at each stage as a depth-wise nonlinear layer into an end-to-end convolutional network to predict the small target. We conduct extensive experiments and comparison with other SOTA methods to verify the performance of our method. The results show that our method outperforms many of its competitors, and when taking computation cost into consideration, it will be more outstanding.

1. Introduction
Infrared small target detection has always been prevalent not only in computer vision field, but also in the real-world application such as warning or surveillance system for decades. The good performance of infrared imaging can alleviate the effect of illumination deficiency and obstruction. However, it still remains to be a great challenge because of the lack of the texture or shape characteristic of the object which usually possesses only few pixels in the image.

To address the problem of the lack of intrinsic features, many works consider to utilize the temporal information to detect the target under the assumption that the target trace between two adjacent frames in the image sequence can be regarded to be continuous to a great extent while the background between two adjacent frames remains static. Considering that the assumption could be strict because the background is fast-changing and the target trace is not continuous enough between adjacent frames under the real-world sensor platforms and that the performance of the multi-frame detection is greatly depends on the precision of the single-frame, single-frame detection draws more and more attention. In addition, the need of the effectiveness of these detection tasks makes researches on single-frame detection which cost less time than multi-frame detection more important. Therefore, our work aims to detect small target in single frame.

Conventional approaches to detect infrared small targets can be classified as follows: model-driven approach detects the targets in a point-wise way, they regard the small target as a few correlated pixels different from the relatively smooth background, these methods need very strong prior assumption and are sensitive to the precision of the model; image processing approaches always utilize the grayscale or gradient information in the spatial domain, sometimes transferring into frequency domain, to distinguish the target from the background, and these methods lack semantic features so that they can only detect salient target [1]; machine learning ways show good performance recently but are also...
faced challenges, the major is that the size of the targets limits the depth of the network which can learn the features better when deeper.

In our work, we combine the advantages of these different methods to handle the insufficient intrinsic features issue. Particularly, we designed a multi-scale local contrast network for single frame detection. In this structure, the high-level features of the input image are firstly extracted through backbone networks, which avoid utilizing primary features directly like image processing; then, multi-scale contrasts are calculated respectively in each stage to fit the various size of the targets; finally, we combine both primary contrast features and high-level features as a module in to the network and the final feature is used to predict the target. In this way, our method retains the learning ability of machine learning, utilizes the multi-scale contrast features, combines multi-layer features to detect the target in an end-to-end network. The experiments following verify the effectiveness and efficiency of our method.

2. Related Works

Human visual attention methods depend greatly on the contrast measurement, which can be seen as the difference of the target and the background in small object detection tasks. In these models, different contrasts are proposed to measure the saliency of the target out of the background and suppress the clutters in the meantime. However, many of these models have strong assumptions, such as the target size is well match the patch size [2], and many of these models utilize the hand-crafted features [3], such as the grayscales, gradients, or their entropy. Our method has no assumption of the size or the local contrast of the target, and we embed the local multi-scale contrasts into the convolutional network for target detection.

Cross-layer feature fusion is universally known as an effective measure for the various scales of the targets. Table 1 shows some combination ways, where \( X \) and \( Y \) represent the feature maps in shallow and deeper layer respectively, \( G \) and \( L \) represent the global contrast and local contrast measurements respectively. Considering the main purpose is to preserve the small target feature in deep layers, we employ the local contrast fusion way to realize the cross-layer fusion.

| Manner                  | Formulation  | Reference |
|-------------------------|--------------|-----------|
| Addition                | \( X+Y \)    | [4]       |
| Bottom-up Modulation    | \( X+L(X)\times Y \) | [5]       |
| Top-down Modulation     | \( G(Y)\times X+Y \) | [6]       |
| Local Contrast Modulation | \( L(X)\times Y \) | ours |

3. Proposed method

The overall architecture of our method is illustrate as figure 1 which mainly contains multi-scale contrast (MLC) module and local contrast fusion (LCF) module. For an input image, firstly, we utilize the modified ResNet-20 [7] which can preserve the feature of the small target well as the backbone to extract the feature maps. Then we subsample twice at stage 2 and stage 3 with a stride of 2 to get the multi-scale features of the small targets. For each stage, we calculate its multi-scale contrast respectively to get three contrast feature maps in different scales and fuse them in the local contrast fusion module to get the final feature map which can be regarded as a saliency map. And finally it is used to predict the target.
3.1. multi-scale contrast (MLC) module
Unlike traditional local measure methods which calculate contrast in a patch-wise way. In these methods, the size of the patch is fixed so that the predict results are depended greatly on whether the size of the target match the size of the patch or not. Our work utilizes the multi-scale contrast (MLC) module to measure the contrast which can be illustrate in Figure 2(a).

Particularly, given a feature map (F) with the size of \(C \times H \times W\) and a contrast rate \(d\), the local contrast of position \((c, i, j)\) can be calculated as:

\[
D(c,i,j) = (F(c,i,j) - F(c,i-x,j-y)) \times (F(c,i,j) - F(c,i+x,j+y))
\]

\[C(c,i,j) = \min_{(x,y)}(D(c,i,j))\text{ (2)}\]

where \(C\) denotes the local contrast in one scale of \(d\), and \((x, y)\) values in \{(d, d), (d, 0), (d, -d), (0, -d)\}.

To accelerate the computation of the \(C\), we use eight preset kernels to shift the feature map circularly in a stride of \(d\) and then the local contrast of the whole feature map can be calculated as:

\[
D = (F - F(x,y)) \times (F - F(-x,-y))
\]

where \(F(x, y)\) and \(F(-x, -y)\) denote the shifted feature map at the direction of \((x, y)\) and \((-x, -y)\) respectively.

After calculating the local contrast in one scale, we repeat these steps to get the different local contrasts called LC1, LC2... LCn and concatenate them into a feature map of size of \(n \times C \times H \times W\). Lastly we perform the max-pooling and squeeze the feature map to get the MLC (FM) whose size is same as the original input feature map, which brings much convenience to further computation.

3.2. local contrast fusion (LCF) module
In order to utilize features from different layers, the cross-layer feature fusion shows to be an important step. The local contrast fusion (LCF) module in our network can be illustrate in Figure 2(b). Where X and Y denotes feature maps from different layers. And specifically X is the low-level and Y is the deeper. The final fusion feature can be calculated as:

\[
LCF = L(X) \times Y
\]

where \(L(X)\) denotes the local contrast of feature map X. When X and Y denotes the MLC(X) and MLC(Y), the LCF will be the final feature map.

The motivation of the LCF module is to embed low-level feature into the deep-level feature to utilize the multi-layer features. Because the local contrast often works well than the global contrast, we choose the local contrast feature as the weight matrix. On the other hand, considering that the low-
level feature contains more details information while the deep-level feature contains better semantic information, low-layer feature is more appropriate for the contrast computation, the following experiments also validate this. Moreover, in order to reduce the quantity of the parameters of the convolution kernels and further reduce the computational complexity, our method utilize the bottleneck structure as illustrated in Figure 2(b).

![Figure 2. The main modules of the multi-scale local contrast network.](image)

(a) MLC, (b) LCF.

4. Experiments
To verify the effectiveness and efficiency of the proposed multi-scale local contrast network, we conduct extensive experiments on the public dataset SIRST provided by [5] which is the largest open dataset for single-frame infrared small target detection according to the author. It contains 427 representative images from different scenarios form hundreds of real-world videos. And the dataset is split into train set, validation set, test set, whose portion are roughly 50%, 20% and 30%.

The evaluation metrics we take are IOU (intersection over union) to evaluate the precision and nIOU (normalized IOU) which is designed specifically for the dataset SIRST according to [5] as a more balanced metric of different methods. It is defined as:

\[ nIOU = \frac{1}{N} \sum_{i} \frac{TP[i]}{T[i] + P[i] - TP[i]} \]  \hspace{1cm} (5)

where TP, T and P denote the true positive, true and positive, respectively.

4.1. Detection results
Figure 4 presents some detection results of representative images from the SIRST dataset. The original images contain different sceneries, where the size and the brightness of the targets are various and the smoothness and complexity of the background are different. Our multi-scale local contrast network can detect the target accurately, which illustrate the effectiveness and robustness of our method, the quantity results are showed in the following part.
Figure 3. The visualization of detection results of representative infrared images from SIRST. The first row shows the original images and the second row shows the detection results.

4.2. Comparison with SOTA approaches

We compare our method with the various forms of the FPN [4] and the ALCNet [5], and the results are showed in Table 2. Where the parameter b denotes the block number or the parameters number in each stage. And other main parameters are as follows: total epochs of 400, weight decay of $10^{-4}$, learning rate of 0.1.

Taking cross-layer fusion manner into consideration alone, the compare results between FPN and Max-FPN shows that the multi-scale contrast can perform better than the single-scale; The compare results between Max-FPN and BGA-FPN shows that the local contrast can perform better than the global contrast. The whole results show that cross-layer feature fusion is necessary for the task while the manner should be further research, but when considering only the dataset SIRST, the all manners which consist of local contrast can be accepted.

| Architecture | Contrast | Scale | Formulation | IOU(b=1) | IOU(b=3) | nIOU(b=1) | nIOU(b=3) |
|--------------|----------|-------|-------------|----------|----------|-----------|-----------|
| FPN          | none     | single| X+Y         | 0.674    | 0.729    | 0.669     | 0.702     |
| Max-FPN      | none     | multi | max(X,Y)    | 0.665    | 0.722    | 0.674     | 0.706     |
| BGA-FPN      | global   | multi | X+G(X)×Y    | 0.676    | 0.731    | 0.679     | 0.704     |
| TLA-FPN      | local    | multi | L(X)×X+Y    | **0.688**| 0.750    | 0.688     | 0.722     |
| ALC Net      | local    | multi | X+L(X)×Y    | 0.677    | 0.753    | 0.686     | **0.724** |
| ours         | local    | multi | L(X)×Y      | 0.680    | **0.756**| **0.689** | 0.719     |

We also compare our method with the SMSL [8], NIPPS [9] and RIPT [10], and the results are showed in Table 3. Their main parameters are as follows: patch size of 50×50, stride of 10, threshold factor of 10. The results show the learning ability of our method and when it takes efficiency into consideration, our method can cost less compute time to get the similarly performance on the IOU and nIOU metric.

| Metric | SMSL | NIPPS | RIPT | FPN | ALC Net | Ours |
|--------|------|-------|------|-----|---------|------|
| IOU    | 0.081| 0.473 | 0.146| 0.720| 0.757   | 0.755|
5. Conclusion
In this paper, we designed a multi-scale local contrast network for single frame detection. We extract the multi-scale contrasts features in the same layer by MLC module and fuse the cross-layer contrast features by LCF module. Moreover, we embed these modules as a layer into an end-to-end convolutional network to predict the small target. The extensive experiments and comparison with other SOTA methods verify the effectiveness and efficiency of our method and shows that cross-layer fusion manners still hold the potential for better performance.

References
[1] Y. Zhao, H. Pan, C. Du, Y. Peng, and Y. Zheng. (2014) Bilateral two-dimensional least mean square filter for infrared small target detection, Infrared Physics & Technology, vol. 65, pp.17–23.
[2] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang. (2014) A local contrast method for small infrared target detection, IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pp. 574–581.
[3] T. X. Bai and Y. Bi. (2018) Derivative Entropy-Based Contrast Measure for Infrared Small-Target Detection, IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2452–2466.
[4] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, (2017) Feature pyramid networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 936–944.
[5] Y. Dai, Y. Wu, F. Zhou and K. B. (2020) Attentional Local Contrast Networks for Infrared Small Target Detection. IEEE TGRS, arXiv: 2012.08573.
[6] H. Li, P. Xiong, J. An, and L. Wang. (2018) Pyramid attention network for semantic segmentation, British Machine Vision Conference (BMVC), Newcastle, UK pp. 1–13.
[7] K. He, X. Zhang, S. Ren, and J. Sun. (2016) Identity mappings in deep residual networks, 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, pp. 630–645.
[8] X. Wang, Z. Peng, D. Kong, and Y. He. (2017) Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 10, pp. 5481–5493.
[9] Y. Dai, Y. Wu, Y. Song, and J. Guo. (2017) Non-negative infrared patch-image model: Robust target-background separation via partial summinimization of singular values, Infrared Physics & Technology, vol. 81, pp. 182–194.
[10] Y. Dai and Y. Wu,. (2017) Reweighted infrared patch-tensor model with both nonlocal and local priors for singleframe small target detection, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 8, pp. 3752–3767.