SUPPLEMENTARY MATERIAL

1. EXPERIMENT

1.1. Visual Results

Fig. 1 shows the visualization results of different methods for some images on the NUAA [1] and IRSTD1k [2] datasets. It can be seen that our method is still able to detect the target effectively even in the case of low contrast, complex background, and more noise interference. All other methods have different degrees of missed and false detection. Compared with our method, which not only has fewer missed and false detection but also can segment the targets better and the segmentation results are more fine compared with other methods.

1.2. Model Design

In this section, we investigate CLFT modules in the encoder part and UCDC modules in the decoder part. We have conducted experiments on the NUAA [1] dataset using the standard ABC model as a baseline.

1.2.1. Encoder

As shown in Table 1, we replaced the convolution module in the first layer of the encoder section with a CLFT module, which showed a significant degradation in performance. We believe that the infrared image has no clear semantic information and has a low signal-to-noise ratio. Since the CLFT module is dimensionally compressed when computing the attention matrix. If the complex original infrared image is directly used as the input, the CLFT module will be disturbed more when computing the attention matrix, and thus the wrong global information will be passed out. Therefore, we let the infrared image pass through a convolution module first, and the infrared image is coarsely extracted by the convolution module to filter out some background clutter so that the feature map input to the CLFT module does not have much interference. We find that the CLFT module has better results in processing the feature maps than the original infrared images directly.

1.2.2. Decoder

As shown in Table 2, we replaced the UCDC module in the first layer of the decoder part with a convolution module, and its performance dropped slightly. We believe that when the feature map enters the decoder part through the transition layer, the resolution of the feature map is small. And after being processed by the UCDC module of the transition layer, the feature map is fine enough. The UCDC module has a larger receptive field than the convolution module, and it is more effective to process small-resolution feature maps than the convolution module. The two UCDC modules process feature maps at different scales, which can effectively filter out the noise interference around the target and make the target outline clearer.

Table 1: $IoU(\%)$, $nIoU(\%)$, $F_1(10^{-2})$ of whether use the ConvModule in the encoder.

ConvModule	IoU ↑	nIoU ↑	F_1 ↑
N	79.29	77.72	88.45
Y	**81.01**	**79.00**	**89.51**

Table 2: $IoU(\%)$, $nIoU(\%)$, $F_1(10^{-2})$ of whether use the UCDC in the decoder.

UCDC	IoU ↑	nIoU ↑	F_1 ↑
N	79.82	77.19	88.78
Y	**81.01**	**79.00**	**89.51**

1.3. Ablation Study

In this section, we use the standard ABC model as a baseline and conduct ablation experiments on some hyperparameters in the model on the NUAA [1] dataset.

1.3.1. Impact of Input Dimension

We study the impact of different input dimensions C on the performance of the model, which is set to 64 by default in the paper. As shown in Table 3, when C is set to 16, 32, and 64 respectively, the performance is also enhanced. But we found that when C is set to 96, the performance drops slightly. We believe that since the infrared image has no clear semantic information, the inductive bias of the model will be affected when the input dimension is too large so that the model cannot effectively extract the target features.

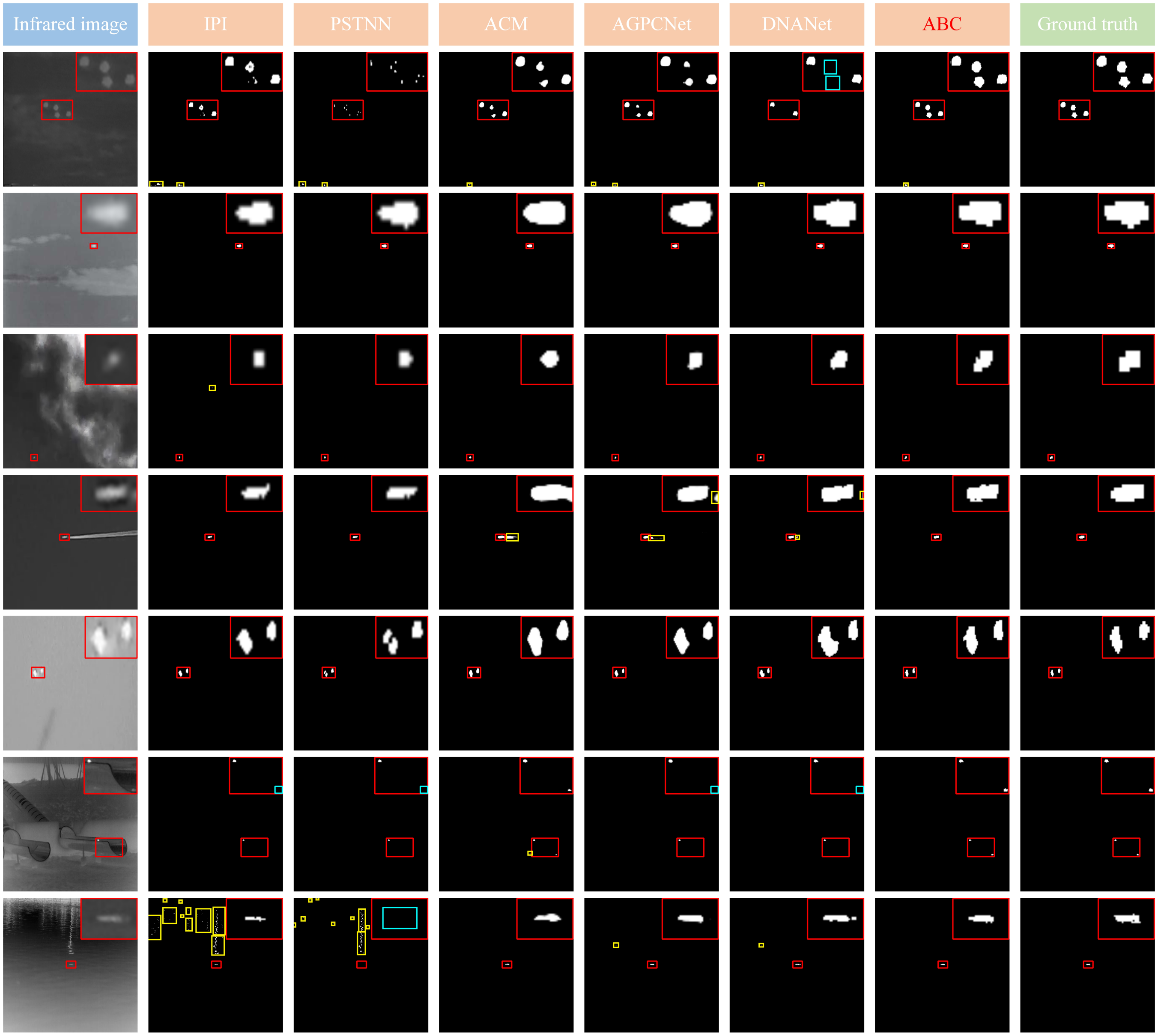

Fig. 1: Partial image visualization results of different methods on NUAA and IRSTD1k datasets. The red box, the yellow box, and the cyan box represent the correct detection box, the false detection box, and the missed detection box, respectively.

Table 3: Ablation study of the input dimension in $IoU(\%)$, $nIoU(\%)$, $F_1(10^{-2})$.

C	IoU ↑	nIoU ↑	F_1 ↑
16	79.86	78.04	88.80
32	80.21	78.15	89.02
64	**81.01**	**79.00**	**89.51**
96	79.92	77.87	88.84

1.3.2. Impact of Dilation Rate

We study the effect of setting different dilation rates on the performance of the model in the three dilated convolutional

Table 4: Ablation study of the dilated rate in $IoU(\%)$, $nIoU(\%)$, $F_1(10^{-2})$.

Dilation Rate	IoU ↑	nIoU ↑	F_1 ↑
1, 1, 1	80.12	77.55	88.96
1, 2, 4	79.80	77.92	88.76
2, 2, 2	80.91	78.40	89.45
2, 4, 6	80.04	78.04	88.91
2, 4, 2	**81.01**	**79.00**	**89.51**

Table 5: Hyperparameter settings

Dataset	Epochs	Lr	Batch
NUAA [1]	1500	0.0003	4
IRSTD1k [2]	500	0.0001	4
SIRSTAUG [3]	500	0.0001	16
NUDT [4]	1500	0.0001	16

layers in the CLFT module. As shown in Table 4, when we set the dilation rates of the three layers of dilated convolutional layers to the common 2, 4, and 6 respectively, the performance decreased slightly. The advantage of dilated convolution is that it has a larger receptive field, so when the dilation rate is set larger on conventional semantic segmentation tasks, it can bring more performance improvements. However, due to the small size of the infrared small target, using a larger dilation rate will introduce additional noise, resulting in blurred target features after feature fusion with the convolution branch. When the dilated rates are set too small, the receptive field will be relatively smaller, which will lead to the inability to effectively perceive the information around the target, resulting in performance degradation. Therefore, setting the dilation rates to 2, 4, and 2 respectively can obtain long-distance information without introducing additional noise due to a large dilation rate.

1.4. Experiment Details

We trained four different models on four datasets. Due to differences in dataset distribution and resolution, we set different hyperparameters for each dataset. Due to space limitations, we did not provide detailed information in the main text. However, we created a configuration file for each dataset in our code, which readers can easily access and reproduce the experiments. Additional experimental details are provided in Table 5, in which the number of GPUs is 4, the scheduler uses the poly strategy, and none of the pretrained model is used.

2. REFERENCES

[1] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard, "Asymmetric contextual modulation for infrared small target detection," in *WACV*, 2021, pp. 950–959.

[2] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo, "Isnet: Shape matters for infrared small target detection," in *CVPR*, 2022, pp. 877–886.

[3] Tianfang Zhang, Siying Cao, Tian Pu, and Zhenming Peng, "Agpcnet: Attention-guided pyramid context networks for infrared small target detection," *arXiv preprint arXiv:2111.03580*, 2021.

[4] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo, "Dense nested attention network for infrared small target detection," *IEEE T IMAGE PROCESS*, 2022.