

The paper suggests a methodology to classify ship types in fine-level by using casual attention and contrast learning. The idea will contribute to classifying the ship types, especially looks similar.

Although the paper involves very useful content, there is some need to modify the manuscript.

1) The table for Abbreviations (Terminology) and Symbols is required for readers to understand and not to be confused with the equations and expressions throughout this paper.

2) [Line 136, 160, 167, 169, 181, 190] Many operation symbols are broken.

image is input into the CNN backbone network to extract features to obtain the feature map $\mathbf{F} \in \mathbb{R}^{W \times H \times C}$, where W , H , and C are the feature map's width, height, and channel

multi-head attention mapping $\mathbf{A} \in \mathbb{R}^{W \times H \times M}$, where M is the number of attention channels. Considering the extreme aspect ratio of ship targets, the strip pooling module (SPM) [23]

To get the regional feature set $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M\} \in \mathbb{R}^{M \times C \times 1}$, $\mathbf{A}$ and $\mathbf{F}$ are multiplied (element by element), and the global average pooling is performed using the formula:

$$\mathbf{r}_m = \{\mathbf{r}_{mc} \mid \mathbf{r}_{mc} = \text{GAP}(\mathbf{A}_m \square \mathbf{F}_c), c \in \{1, 2, \dots, C\}, m \in \{1, 2, \dots, M\}\} \quad (1)$$

$$\begin{aligned} \text{GAP}(\mathbf{A}_{\text{sum}} \square \mathbf{F}) &= \text{GAP}\left(\sum_{m=1}^M \mathbf{A}_m \square \mathbf{F}\right) \\ &= \text{GAP}\left(\sum_{m=1}^M (\mathbf{A}_m \square \mathbf{F})\right) \\ &= \sum_{m=1}^M \text{GAP}(\mathbf{A}_m \square \mathbf{F}) \\ &= \sum_{m=1}^M \mathbf{r}_m \\ &= \mathbf{I} \end{aligned} \quad (4)$$

$$\hat{\mathbf{r}}_m = \{\hat{\mathbf{r}}_{mc} \mid \hat{\mathbf{r}}_{mc} = \text{GAP}(\hat{\mathbf{A}}_m \square \mathbf{F}_c), c \in \{1, 2, \dots, C\}, m \in \{1, 2, \dots, M\}\} \quad (6)$$

$$\hat{\mathbf{I}} = \sum_{m=1}^M \hat{\mathbf{r}}_m \quad (7)$$

3) [Line 143] Figure 2 represents the summary of the proposed model. Accordingly, many abbreviations are recorded in the figure, but there is no explanation in the caption of the figure. Thus, it is required to explain the abbreviations or make a table for abbreviations and symbols. For instance, SPM, GAP, FRM, and L_{agg} are recorded somewhere in this paper. But, it is difficult to find them in the context.

- 4) [Line 155] Figure 3 includes Chinese Mandarin in the title. And the full name of SPM and the explanation is required for understanding this figure. How can the $W \times 1 \times M$ be converted into $W \times H \times M$ without any operation? An operation symbol or an explanation is required. Likewise, the process from $1 \times H \times M$ to $W \times H \times M$ also requires an explanation.

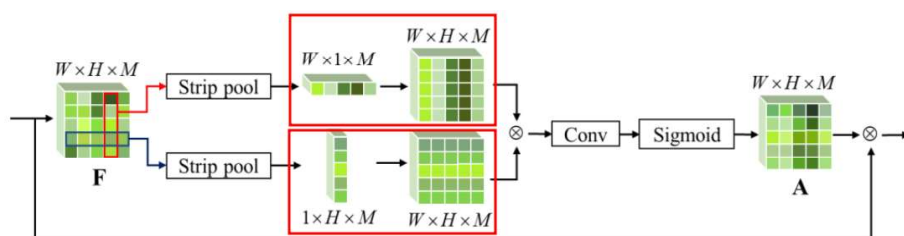


Figure 3. SPM 结构

155
156

- 5) [Line 170] What do you mean by the subscripts "mc" and "m"? The full name or explanation is required. "the c -th" is preferred than cth .

where r_{mc} is the value of the c th element in the region feature vector r_m ; GAP is global average pooling; $\otimes$ is element by element multiplication. The local features of the target

170
171

- 6) [Line 173-174] The equation shows lowercase "l" representing uppercase "L", but the first character next paragraph is uppercase "l" representing lowercase "i". Is it correct?
- 7) [Line 182, 298] "Fully connected layer is selected as the classifier". Then, what is the activation function for your model? And, why did you select the activation function?
- 8) [Line 184] Why the predicted output of multiple attention branches is recorded as Y_A instead of $\hat{Y}_A$? The hat symbol represents the prediction.
- 9) [Line 199] Are the multiple attention and predicted results expressed by Y_A and $\hat{Y}_A$?
- 10) [Line 235] "feature concatenation" is more general than "feature polymerization". If there is any intention of the authors, please describe it.

- 11) [Line 259, 262, 265] The sentences in Line 259 and 265 explain FAM. But Figure 3 shows FRM instead of FAM.
- 12) [Line 259, 262, 265] The sentences in Line 259 and 265 explain FAM. But Figure 3 shows FRM instead of FAM.
- 13) [Line 285] It may be better to express "Chen et al. (2020)" instead of "in reference 19".
- 14) [Line 287] What do you mean by the term "proxy vector"?
- 15) [Line 371-374] In equation 28, the formula represents "recall" of class c , not represents "accuracy". Thus, correct the " ACC_c " as " $Recall_c$ ". Accordingly, the average recall rate of all categories must be "Average Recall (AR)" or something, not to be "Average Accuracy (AA)".
- 16) [Line 373] Please, add the matrices "F1-score" and "Precision". If not possible, just explain. Then, it will be accepted.
- 17) [Line 439, 453] Figure 10 and Figure 10 have too small font sizes. Thus, it is difficult to identify the number and characters. Enlarge font size until we can identify.