Response to Reviewer 2 Comments

Dear reviewer,

We gratefully thank you for giving us the opportunity to revise the manuscript entitled " **MsIFT : Multi-source Image Fusion Transformer"**. We really appreciate all insightful comments and valuable suggestions, which are definitely helpful for improving the quality of our manuscript. We have studied all of your comments to revise the manuscript, and the responses to your comments are as follows (the replies are highlighted in red).

Point 1: The vision transformer should be introduced in considerable detail because it is the heart of the proposed method. The authors are highly suggested to describe how the transformer models long-range dependence through the attention mechanism. If possible, please utilize mathematical expressions and visual illustrations for an intuitive description.

Response 1: Thank you for the meaningful suggestion. To better understand the advantage of the attention mechanism in MsIFT, we have added the qualitative analysis of the attention mechanism in section 3.2.2. Please refer to Response 3. *(line 196 to 203, 208 to 214)*

Besides, Eq. 7~11 and Eq.17~21 are the mathematical expression of the self-attention mechanism and the cross-attention mechanism, respectively. The visual illustration is shown in Fig. 5. We display the attention map of the selected feature points. The brighter area shows the attention area of the selected feature points. It can be seen that each selected feature point has a global receptive field on the whole image, i.e., long-range dependence.

Point 2: What does the minus sign on the right side the MCA block in Figure 2 mean? Its description should be added to the caption of Figure 2.

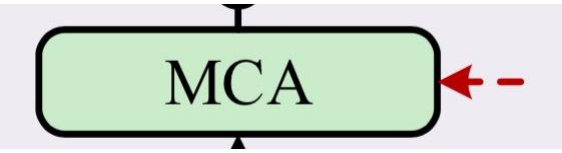


Fig.2 MCA block

Response 2: Thank you for the meaningful comment. The sign beside the MCA block is not a minus sign but a dashed line. It means that input the features of another source into the MCA. We have added a description beside the dashed line per your suggestion.

Point 3: The description of the feature fusion transformer is quite limited, causing difficulties to access the paper ' s novelties. Thus, it is advisable to provide more details about the proposed transformer, notably the MCA and the MSA.

Response 3: Thank you for the meaningful suggestion. We added the qualitative analysis of the feature fusion transformer module in section 3.2.2. *(line 196 to 203, 208 to 214)*

Feature fusion transformer consists of two components, i.e., fusion transformer encoder (FTE) and fusion transformer decoder (FTD). The former is used to extract the single-source feature, and the latter is used to fusion the multi-source feature.

In FTE, assume that F is the CNN feature map of a single-source image. F is fed into the FTE. For each feature point $F^{(i,j)}$ on the spatial position (i,j) of F, the multihead self-attention (MSA) module in the FTE performs a global search on the F to find a feature point set $\{F^k\}$ with similar semantic information to $F^{(i,j)}$. Specifically, in the search process, as shown in Eq. (10), the cosine distance is used to calculate the semantic similarity matrix between $F^{(i,j)}$ and the feature point on each spatial position of F. The similarity matrix is normalized by the softmax operation. The feature points on the F are weighted and summed using this normalized matrix (attention map). Finally, the obtained features are fused with $F^{(i,j)}$. Since MSA makes the features have a larger receptive field, it enhances the discriminative of the features.

In FTD, assume that F_A is the feature map of source A, F_B is the feature map of source B. To avoid the problem of semantic bias, feature point $F_A^{(i,j)}$ at the spatial position (i,j) of F_A is not directly fused with $F_B^{(i,j)}$. Instead, the multihead cross-attention (MCA) module in the FTD performs a global search on the F_B to find a feature point set $\{F_B^k\}$ with similar semantic information to $F_A^{(i,j)}$. Specifically, in the search process, as shown in Eq. (20), the cosine distance is used to calculate the semantic similarity matrix between $F_A^{(i,j)}$ and the feature point on each spatial position of F_B. The similarity matrix is normalized by the softmax operation. The feature points on the F_B are weighted and summed using this normalized matrix. Finally, the obtained features are fused with $F_A^{(i,j)}$. **Therefore, even if there is the semantic bias between the two sources, FTD will fuse the features of source A and those of source B with similar semantic information according to the semantic similarity matrix.**

Point 4: It is suggested to double-check and revise (if necessary) equations and notations in the paper for correctness and consistency. For example, the reviewer could not find definitions of Datt in Equation (10) and Laux on line 206. Also, as the number of patches N is an integer, its calculation should be protected by a rounding, flooring, or ceiling operator.

Response 4: Thanks for the reviewer's careful suggestions. We have revised all the equations and notations. The supplementary description has been given for symbols that lack definitions. *(line 193 to 194, line 204 to 205, line 243 to 247)*

Point 5: According to the results of SF-SRDA [30], it appears that the performance comparison in Table 4 was conducted using daytime images. If so, please revise the description in the main text for clarity.

Response 5: Point taken and thanks. We have added the description in the experiment section. *(line 296)*