MOS: Mitigating Optical-SAR Modality Gap for Cross-Modal Ship Re-Identification

Supplementary Material

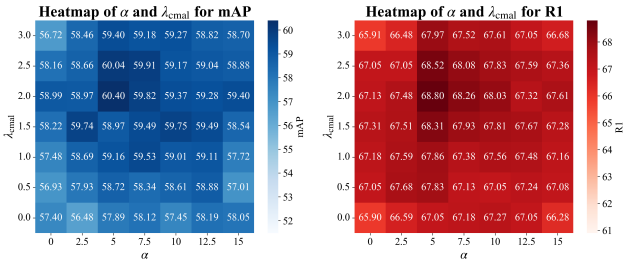

Figure 1. R1 heatmap based on α and λ_{cmal} in *ALL to ALL* protocol.

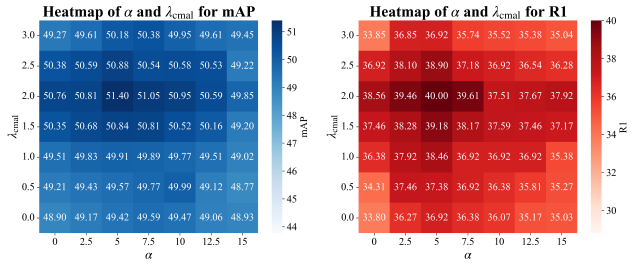

Figure 2. R1 heatmap based on α and λ_{cmal} in *Optical to SAR* protocol.

1. Wasserstein-2 distance

The Wasserstein-2 distance (W_2) is a fundamental metric from optimal transport that measures the minimal effort required to transform one probability distribution into another. For two distributions p and q defined on a metric space $\mathcal{X}$, the W_2 distance is given by:

$$W_2(p,q) = \left(\inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{(x,y)\sim\gamma} \left[\|x-y\|_2^2 \right] \right)^{1/2}, \quad (1)$$

where $\Gamma(p,q)$ denotes the set of couplings whose marginals are p and q. The infimum corresponds to the optimal transport plan that minimizes the quadratic transportation cost.

A key advantage of W_2 is that it admits a closed-form expression when both distributions are Gaussian. Let

$$p = \mathcal{N}(\mu_1, \Sigma_1), \qquad q = \mathcal{N}(\mu_2, \Sigma_2),$$

then the squared Wasserstein-2 distance is:

$$W_2^2(p,q) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2\,(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big). \quad (2)$$

This formulation shows that W_2 jointly aligns the means and covariances of the two distributions, providing a geometry-aware and numerically stable measurement of distributional discrepancy. Compared with KL divergence or simple ℓ_2 matching, W_2 remains finite for distributions with disjoint support and captures covariance differences naturally. These properties make it particularly effective for aligning Gaussian feature distributions in our method.

2. Heatmap in other evaluation protocols

Both heatmaps in Fig. 1 and Fig. 2 illustrate the sensitivity of our method with respect to the hyper-parameters α and λ_{cmal} under the *ALL to ALL* and *Optical to SAR* protocols. Despite the differences in evaluation settings, the two heatmaps exhibit highly consistent trends: the optimal performance is achieved when $\alpha = 5.0$ and $\lambda_{\text{cmal}} = 2.0$ across all protocols. This consistency indicates that our cross-modal alignment strategy is robust and not overly sensitive to the choice of α and λ_{cmal} across different protocols.