A Sensor Agnostic Domain Generalization Framework for Leveraging Geospatial Foundation Models: Enhancing Semantic Segmentation via Synergistic Pseudo-Labeling and Generative Learning

Supplementary Material

A. Mathematical Insight

A.1. Notations and Definitions

Datasets Notations and Definitions: Let Y_i^τ represent the class (as a random variable) at the i^{th} pixel of an image $\mathbf{X}_\tau$ taken from domain τ, where $\tau \in \{S, T\}$ denotes the source or target domain, respectively. The random variable X_i^τ represents the channel vector at the i^{th} pixel in the image $\mathbf{X}_\tau$, i.e., $X_i^\tau \in \mathbb{R}^{C_\tau}$, where C_τ is the number of channels. An observation of of the random variable Y_i^τ is denoted as y_i^τ, and similarly, an observation of the random variable X_i^τ is denoted as x_i^τ.

Let the set $\mathbf{X}_S = \{x_i^S\}_{i=1}^{N_S=H_S\times W_S}$ represent the source domain image, where $\mathbf{X}_S \in \mathbb{R}^{C_S\times H_S\times W_S}$ is the set of channel vectors for all pixels in the source image, with H_S and W_S being the height and width of the image, respectively. Similarly, let the set $\mathbf{X}_T = \{x_i^T\}_{i=1}^{N_T=H_T\times W_T}$ represent the target domain image, where $\mathbf{X}_T \in \mathbb{R}^{C_T\times H_T\times W_T}$ is the set of channel vectors for all pixels in the target image.

The labels for the source domain are represented by the set $\mathbf{Y}_S = \{y_i^S\}_{i=1}^{N_S}$, where we assume the source domain is fully labeled. For the target domain, the labeled samples are represented by the set $\mathbf{Y}_T = \{y_i^T\}_{i=1}^{l}$, where l is the number of labeled pixels and $l \ll N_T$.

Models Defentions : Let f_θ: the feature extractor parametrized by θ

$h_{\theta_{\text{seg}}}$: the segmentation head parametrized by θ_{seg}

g_{θ_M}: MAE head parametrized by θ_M

Then we can define the segmentation model and the generative model we have in our proposed method as follows :

$$\mathbb{P}(y_i \mid \mathbf{X}; \theta, \theta_{\text{seg}}) = h_{\theta_{\text{seg}}}(f_\theta(\mathbf{X}), y_i) \tag{10}$$

The probability of class y at the i^{th} pixel given the entire image $\mathbf{X}$. By employing mean squared error (MSE) in the generative component, we inherently assume the following model :

$$\begin{aligned} X_i^T &= g_{\theta_M}\left(f_\theta(X_S)\right) + \epsilon, \\ &\text{where } \epsilon \sim \mathcal{N}(0, \Sigma), \\ &\text{and } \Sigma \in \mathbb{R}^{C_T\times C_T}. \end{aligned} \tag{11}$$

A.2. Task Learning as Joint Likelihood

The segmentation, generative, and domain adaptation (DA) tasks in the proposed method are governed by minimizing the following loss function:

$$\begin{aligned} \underset{\theta,\theta_{seg},\theta_M}{\operatorname{argmin}} \mathcal{L}_{Tot}(\lambda_{DA}, \lambda_{MAE}, \ldots) &= \mathcal{L}_{Seg} + \lambda_{DA}\mathcal{L}_{DA} \\ &+ \lambda_{MAE}\mathcal{L}_{MAE} \end{aligned} \tag{12}$$

Here, $\mathcal{L}_{Seg}$ denotes the multi-class cross-entropy loss, computed on the labeled source domain and a limited number of labeled target domain samples. The $\mathcal{L}_{DA}$ term facilitates domain alignment between the source and target domains, while $\mathcal{L}_{MAE}$ represents the mean squared error (MSE) used in MAE method.

For clarity in this mathematical insight, we focus solely on the impact of MAE learning on the segmentation task. Therefore, we set $\lambda_{DA} = 0$ and $\lambda_{MAE} = 1$ in Equation (12). Hence

$$\underset{\theta,\theta_{seg},\theta_M}{\operatorname{argmin}} \mathcal{L}_{Tot}(0, 1, \ldots) = \underset{\theta,\theta_{seg},\theta_M}{\operatorname{argmax}} J(\theta, \theta_{seg}, \theta_M) \tag{13}$$

Where $J(\theta, \theta_{seg}, \theta_M)$ is the joint probabilty of the *observed* source domain labels, $\mathbf{Y}_S$, the few *observed* labels from the target domain, $\mathbf{Y}_T$, and the *observed* target domain image $\mathbf{X}_T$, conditioned on the *observed* source domain image, $\mathbf{X}_S$, and parametrized by the model parameters :

$$J(\theta, \theta_{\text{seg}}, \theta_M) = \mathbb{P}(\mathbf{Y}_S, \mathbf{Y}_T, \mathbf{X}_T \mid \mathbf{X}_S; \theta, \theta_{\text{seg}}, \theta_M) \tag{14}$$

The joint distribution in Equation (14) can be factorized as follows:

$$\begin{aligned} \mathbb{P}(\mathbf{Y}_S, \mathbf{Y}_T, \mathbf{X}_T \mid \mathbf{X}_S; \theta, \theta_{\text{seg}}, \theta_M) &= \mathbb{P}(\mathbf{Y}_S \mid \mathbf{Y}_T, \mathbf{X}_T, \mathbf{X}_S) \\ &\times \mathbb{P}(\mathbf{Y}_T \mid \mathbf{X}_T, \mathbf{X}_S) \\ &\times \mathbb{P}(\mathbf{X}_T \mid \mathbf{X}_S), \end{aligned} \tag{15}$$

Assuming the conditional independence assumption holds for the output quantities given the inputs, we have $Y_i^\tau \perp \zeta \mid \mathbf{X}_\tau$ for the segmentation model, where ζ represents any variable other than Y_i^τ. Likewise, for the MAE model, the condition $X_i^T \perp \zeta \mid \mathbf{X}_S$ applies for all ζ not equal to X_i^T. Under these assumptions, the joint distribution in Equation (15) simplifies as follows:

$$\begin{aligned} \mathbb{P}(\mathbf{Y}_S, \mathbf{Y}_T, \mathbf{X}_T \mid \mathbf{X}_S; \theta, \theta_{\text{seg}}, \theta_M) &= \mathbb{P}(\mathbf{Y}_S \mid \mathbf{X}_S; \theta, \theta_{\text{seg}}) \\ &\times \mathbb{P}(\mathbf{Y}_T \mid \mathbf{X}_T; \theta, \theta_{\text{seg}}) \\ &\times \mathbb{P}(\mathbf{X}_T \mid \mathbf{X}_S; \theta, \theta_M) \end{aligned} \tag{16}$$

Utilizing the conditional independence assumption to write Equation (16) on the pixel level :

$$J = \prod_{n\in\{1,\dots,N_S\}} \mathbb{P}(y_n^s \mid \mathbf{X}_S;\theta,\theta_{\text{seg}}) \times \prod_{i\in\{1,\dots,l\}} \mathbb{P}(y_i^T \mid \mathbf{X}_T;\theta,\theta_{\text{seg}}) \times \prod_{k\in\{1,\dots,N_T\}} \mathbb{P}(x_k^T \mid \mathbf{X}_S;\theta,\theta_M) \tag{17}$$

A.2.1. Impact of Unlabeled Target Domain Sample Reconstruction on the Segmentation Task

Since our approach incorporates a few labeled samples from the target domain alongside labeled data from the source domain, the impact of these labeled target samples on the segmentation task can be inferred through their direct influence on the decision boundaries of the segmentation model. However, the use of the MAE loss has enabled the integration of unlabeled samples from the target domain. We are specifically focused on investigating the impact of these unlabeled samples on the segmentation task. To achieve this, we will reformulate Equation (17) to distinguish between the labeled and unlabeled pixels from the target domain. Before doing so, it is essential to first highlight the following product that appears in Equation (17).

$$\mathbb{P}(y_i^T \mid \mathbf{X}_T;\theta,\theta_{\text{seg}})\mathbb{P}(x_k^T \mid \mathbf{X}_S;\theta,\theta_M) \tag{18}$$

According to the basic rule of conditional probability and the conditional independence assumption, Expression (18) can be written as follows:

$$\begin{aligned}\mathbb{P}\left(y_i^T, x_k^T \mid \{x_r^T\}_{r\neq k}, \mathbf{X}_S; \bar{\theta}=[\theta,\theta_{\text{seg}},\theta_M]\right) &= \mathbb{P}\left(y_i^T \mid \mathbf{X}_T;\theta,\theta_{\text{seg}}\right) \\ &\times \mathbb{P}\left(x_k^T \mid \mathbf{X}_S;\theta,\theta_M\right)\end{aligned} \tag{19}$$

In words, Expression (18) represents the joint probability of Y_i^T and X_k^T, parameterized by $\bar{\theta}=[\theta,\theta_{\text{seg}},\theta_M]$, conditioned on $\mathbf{X}_S$ and all channel vectors of $\mathbf{X}_T$, excluding x_k^T, which we denote as $\{x_r^T\}_{r\neq k}$.

Substituting Equation (19) in Equation (17) and taking the *log* we got :

$$\begin{aligned}\log J = &\sum_{n\in\{1,\dots,N_S\}} \log\mathbb{P}\left(y_n^S \mid \mathbf{X}_S,\theta,\theta_{\text{seg}}\right) \\ &+ \sum_{i\in\{1,\dots,l\}} \log\mathbb{P}\left(y_i^T, x_i^T \mid \{x_r^T\}_{r\neq i}, \mathbf{X}_S, \bar{\theta}\right) \\ &+ \sum_{k\in\{1,\dots,N_T\}\setminus\{1,\dots,l\}} \log\mathbb{P}\left(x_k^T \mid \mathbf{X}_S,\theta,\theta_M\right).\end{aligned} \tag{20}$$

To evaluate how this learning paradigm influences the feature extractor during training, a key component of our proposed method for jointly learning the segmentation and MAE tasks, we will compute the partial derivative with respect to θ. This approach is motivated by gradient-based optimization methods, where the derivative is responsible for updating the model parameters, enabling us to observe how the feature extractor evolves and adapts throughout the training process:

$$\begin{aligned}\frac{\partial J}{\partial\theta} = &\sum_{n\in\{1,\dots,N_S\}} \frac{\partial}{\partial\theta}\log\mathbb{P}\left(y_n^S \mid \mathbf{X}_S,\theta,\theta_{\text{seg}}\right) \\ &+ \sum_{i\in\{1,\dots,l\}} \frac{\partial}{\partial\theta}\log\mathbb{P}\left(y_i^T, x_i^T \mid \{x_r^T\}_{r\neq i}, \mathbf{X}_S, \bar{\theta}\right) \\ &+ \sum_{k\in\{1,\dots,N_T\}\setminus\{1,\dots,l\}} \frac{\partial}{\partial\theta}\log\mathbb{P}\left(x_k^T \mid \mathbf{X}_S,\theta,\theta_M\right)\end{aligned} \tag{21}$$

By utilizing the following "trick " provided by [19]:

$$\begin{aligned}\forall x,y,\ \frac{\partial}{\partial\theta}\log\mathbb{P}(x|\theta) &= \frac{1}{\mathbb{P}(x|\theta)}\frac{\partial}{\partial\theta}\mathbb{P}(x|\theta) \\ &= \frac{1}{\mathbb{P}(x|\theta)}\frac{\partial}{\partial\theta}\left(\sum_y \mathbb{P}(x,y|\theta)\right) \\ &= \frac{1}{\mathbb{P}(x|\theta)}\sum_y \frac{\partial}{\partial\theta}\mathbb{P}(x,y|\theta) \\ &= \frac{1}{\mathbb{P}(x|\theta)}\sum_y \mathbb{P}(x,y|\theta)\left(\frac{1}{\mathbb{P}(x,y|\theta)}\frac{\partial}{\partial\theta}\mathbb{P}(x,y|\theta)\right) \\ &= \sum_y \frac{\mathbb{P}(x,y|\theta)}{\mathbb{P}(x|\theta)}\frac{\partial}{\partial\theta}\log\mathbb{P}(x,y|\theta) \\ &= \sum_y \mathbb{P}(y|x,\theta)\frac{\partial}{\partial\theta}\log\mathbb{P}(x,y|\theta).\end{aligned} \tag{22}$$

Applying the trick in Equation (22) for the last term in Equation (21), we got :

$$\begin{aligned}\frac{\partial J}{\partial\theta} = &\sum_{n\in\{1,\dots,N_S\}} \frac{\partial}{\partial\theta}\log\mathbb{P}\left(y_n^S \mid \mathbf{X}_S,\theta,\theta_{\text{seg}}\right) \\ &+ \sum_{i\in\{1,\dots,l\}} \frac{\partial}{\partial\theta}\log\mathbb{P}\left(y_i^T, x_i^T \mid \{x_r^T\}_{r\neq i}, \mathbf{X}_S, \bar{\theta}\right) \\ &+ \sum_{k\in\{1,\dots,N_T\}\setminus\{1,\dots,l\}} \sum_{y_k} \mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S;\bar{\theta}) \\ &\times \frac{\partial}{\partial\theta}\log\mathbb{P}(x_k^T, y_k \mid \mathbf{X}_S;\bar{\theta})\end{aligned} \tag{23}$$

The final term in Equation (23) holds particular importance, as it demonstrates how our learning paradigm dynamically adjusts the influence of unlabeled pixels on the classification process for each class. This adjustment is achieved

through the expression $\mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S; \bar{\theta})$, which functions as a dynamic weighting mechanism. In the following, we will further explain this dynamic weighting mechanism by relating it to the definitions of our models introduced earlier.

A.2.2. Dynamic Weighting in terms of the Segmentation Model

By applying the conditional probability rule and the conditional independence assumption, the joint distribution $\mathbb{P}\left(y_k, \{x_r^T\}_{r\neq k} \mid x_k^T; \bar{\theta}\right)$ can be written:

$$\begin{aligned}\mathbb{P}\left(y_k, \{x_r^T\}_{r\neq k} \mid x_k^T; \bar{\theta}\right) &= \mathbb{P}\left(y_k \mid \mathbf{X}_T, \mathbf{X}_S, \bar{\theta}\right) \\ &\times \mathbb{P}\left(\{x_r^T\}_{r\neq k} \mid \mathbf{X}_S, \bar{\theta}\right) \\ &= h(f(\mathbf{X}_T), y_k) \\ &\times \prod_{r\neq k} \mathcal{N}(x_r^T; g(f(\mathbf{X}_S, r)), \Sigma).\end{aligned} \tag{24}$$

$$\mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S, \bar{\theta}) \times \mathcal{N}(x_k^T; g(f(\mathbf{X}_S, k)), \Sigma) \tag{25}$$

Now we can direclty link the weight $\mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S; \bar{\theta})$ to our segmentation model $h(f(.))$ through marginalising the joint distribution in Equation (24)

$$\begin{aligned}\mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S, \bar{\theta}) = \int \cdots \int_{\{x_r^T\}_{r\neq k}} & h(f(\mathbf{X}_T), y_k) \\ &\times \mathbb{P}\left(\{x_r^T\}_{r\neq k} \mid \mathbf{X}_S, \bar{\theta}\right) \prod_{r\neq k} dx_r^T.\end{aligned} \tag{26}$$

In order to write Equation (26) in a compact way, we let $\mathbf{Z} = \{x_r^T\}_{r\neq k}$ such that Equation (26) can be written as follows:

$$\mathbb{P}(y_k \mid x_k^T, \mathbf{X}_S, \bar{\theta}) = \mathbb{E}_{\mathbf{Z}\sim\mathbb{P}(\mathbf{Z}\mid\mathbf{X}_S,\bar{\theta})}\left[h(f(\mathbf{Z}, x_k^T), y_k)\right] \tag{27}$$

Equation (27) demonstrates that the weighting mechanism is closely linked to the segmentation model's confidence regarding the unlabeled pixels in the target domain. If the model is incorrectly overconfident about these unlabeled pixels, it can adversely affect the learning of the segmentation task. Conversely, when the model is correctly confident about these unlabeled pixels, it enhances the model's generalizability over the target domain by effectively leveraging the unlabeled samples in conjunction with the limited labeled ones.

B. Implementation Details

The experiments were conducted using the PyTorch framework with the AdamW optimization algorithm [25]. A learning rate of 10^{-4} was used. Each experiment was repeated ten times for each datasets, with a unique random initialization of the network for each run. In each experiment, validation and testing were performed on datasets sampled from the target domain, and the best model was selected based on validation performance during training. Both hyperparameters, λ_{MAE} and λ_{DA}, were empirically set to a value of 1.

C. Inference Maps

C.1. C2Seg-AB

Fig. 4 displays segmentation inference maps for this dataset. Consistent with the quantitative results in Tab. 1, our proposed method performs comparably to PCS, with both showing superior performance relative to other methods. Notably, our method demonstrates improved segmentation for classes like Surface Water and Mine, Dump, and Construction Sites. It is worth mentioning that certain classes, such as Street, experience low precision across all methods, including ours.

C.2. FLAIR

Fig. 5 displays segmentation inference maps for this dataset, highlighting that our approach achieves the most accurate segmentation map compared to other methods. Notably, our method demonstrates improved precision and recall for classes such as Agricultural Land, Plowed Land, and Brushwood.

D. MAE-Based Generative Performance

To evaluate the generative performance of our proposed model across different data modalities (MSI and HSI), we train it exclusively on the generative task represented primarily by $\mathcal{L}_{MAE}$ until convergence.Specifically, this assessment aims to measure the model's ability to reconstruct the target domain image sequence from the concatenated source-target sequence, thereby encouraging the learning of generalizable domain-invariant features. For this evaluation, we employ a masking ratio of 50% for target domain input images while keeping source domain images fully unmasked to generate the source-target sequence.

D.1. HSI

We evaluated the trained model on the HSI modality using the C2Seg-AB dataset, varying the masking ratio of the target domain images across 50%, 75%, and 100%, while keeping the source domain images fully unmasked (0% masking) across all experiments. To evaluate the reconstruction performance both spatially and spectrally, Fig. 6 provides a visual assessment of spatial reconstruction quality, while Fig. 7 assess the reconstruction performance from a spectral perspective. The results from both spatial and

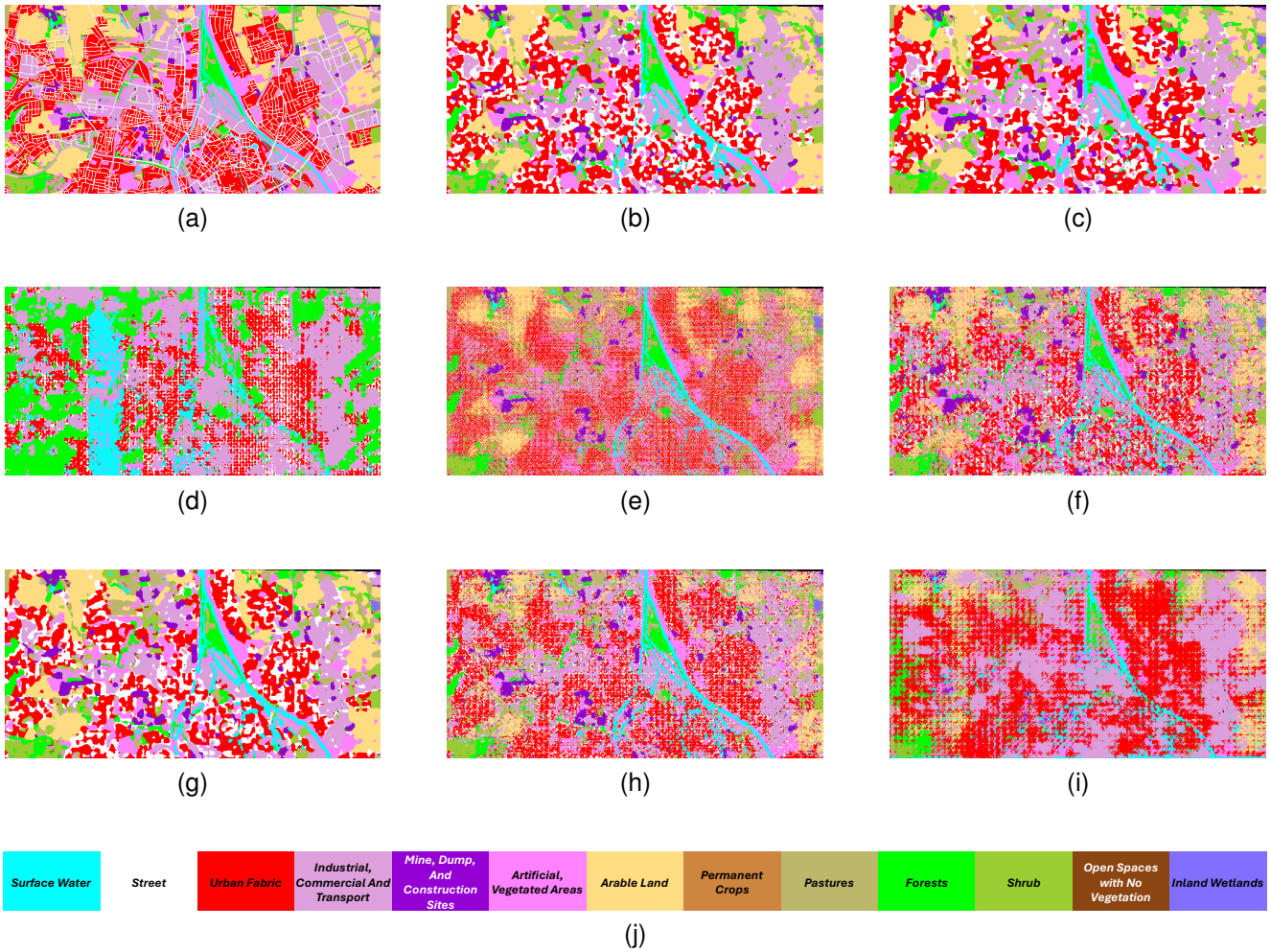


Figure 4. Comparative Segmentation Inference using the C2Seg-AB Dataset. Display includes (a) Ground Truth Mask, (b) Our Method, (c) PCS, (d) Zero Shot, (e) GDA, (f) CDS, (g) MIC, (h) CIA_UDA, (i) UDA_ME_BS, and (j) Colorbar.

spectral evaluations demonstrate strong reconstruction performance, highlighting the effectiveness of the proposed method. This outcome highlights the model's ability to capture informative features from one domain to aid reconstruction in another, promoting the learning of domain-invariant features. Importantly, our proposed framework extends MAE-based generative learning specifically to RS foundation models across multiple modalities, including HSI, which, as previously noted, has seen limited application with MAE-based self-learning approaches. For additional visual evaluation please refer to the supplementary material.

D.2. MSI

In a similar evaluation, we assessed the trained model on the MSI modality using the FLAIR dataset across various masking percentages for the target image, as shown in Fig. 8. The model consistently achieves near-perfect reconstruction quality under different masking ratios, demonstrating the flexibility and adaptability of our proposed method across multiple data modalities.

E. Additional Ablation Study Results

Similar to Tab. 3 in terms of the baseline represented by the second column, we observe that adding $\mathcal{L}_{MAE}$ alone degrades performance compared to the baseline, which can be attributed to the dynamic weighting mechanism discussed in the mathematical insight section. This mechanism depends on the model's predictions for unlabeled samples in the target domain, leading to potential issues when the model is mistakenly confident about these samples. However, combining $\mathcal{L}_{MAE}$ with $\mathcal{L}_{DA}$ yields the best performance, consistent with the previously described synergy between these two components.

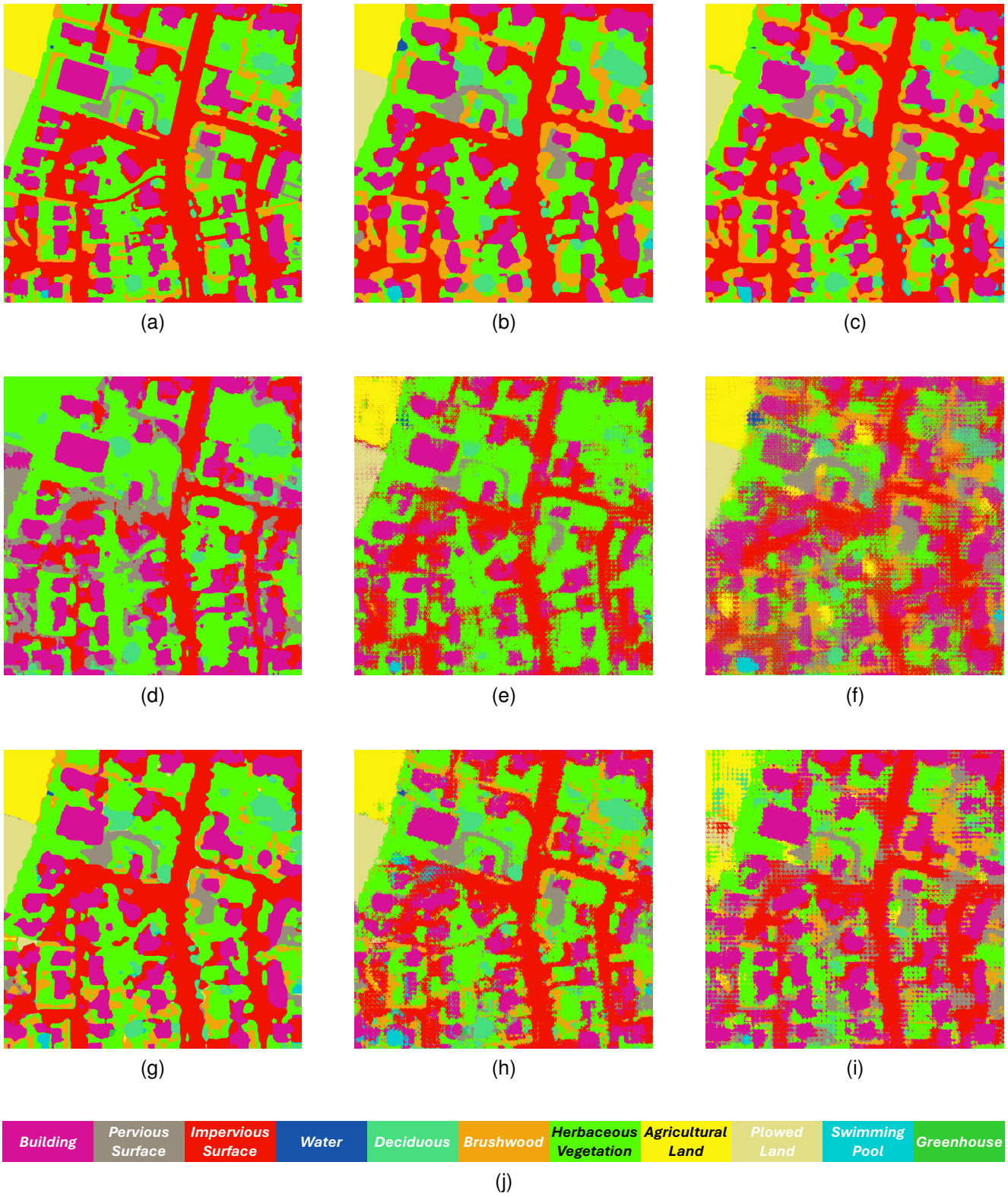


Figure 5. Comparative Segmentation Inference using FLAIR Dataset. Display includes (a) Ground Truth Mask, (b) Our Method, (c) PCS, (d) Zero Shot, (e) GDA, (f) CDS, (g) MIC, (h) CIA_UDA, (i) UDA_ME_BS, and (j) Colorbar.

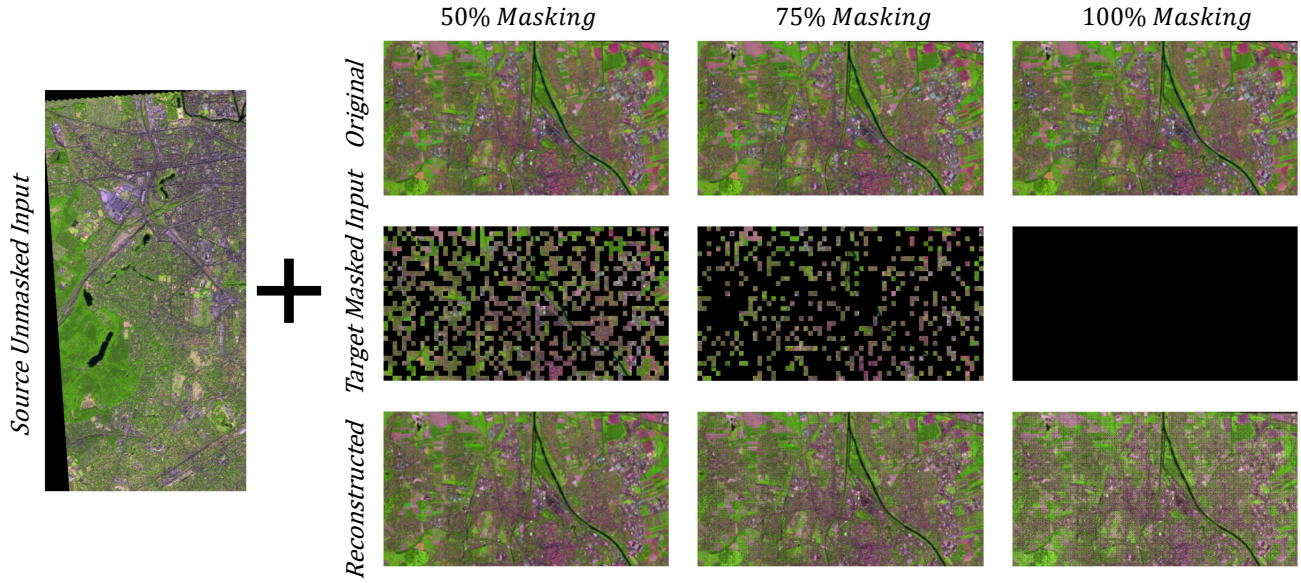
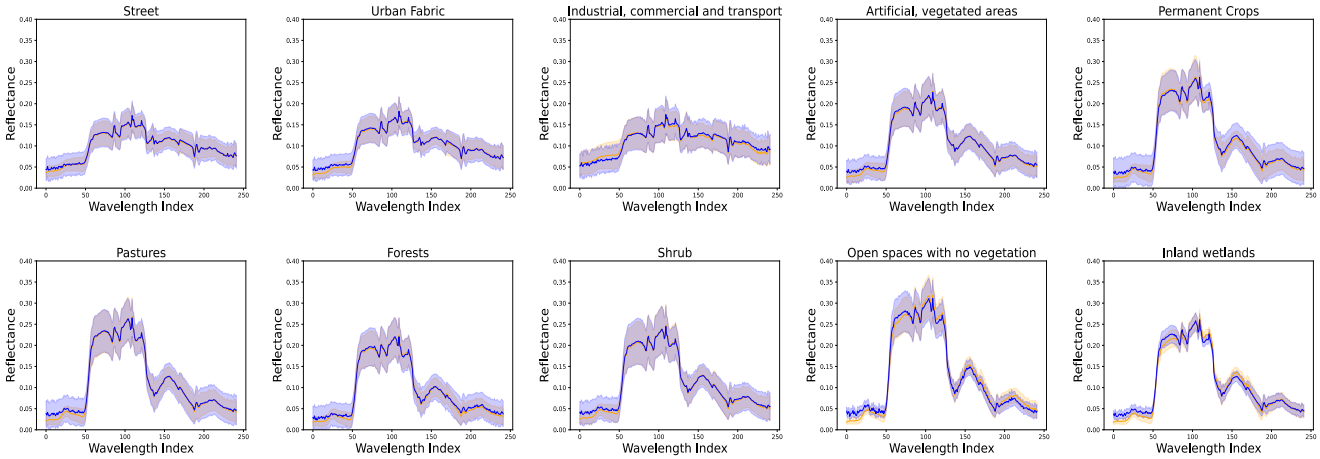


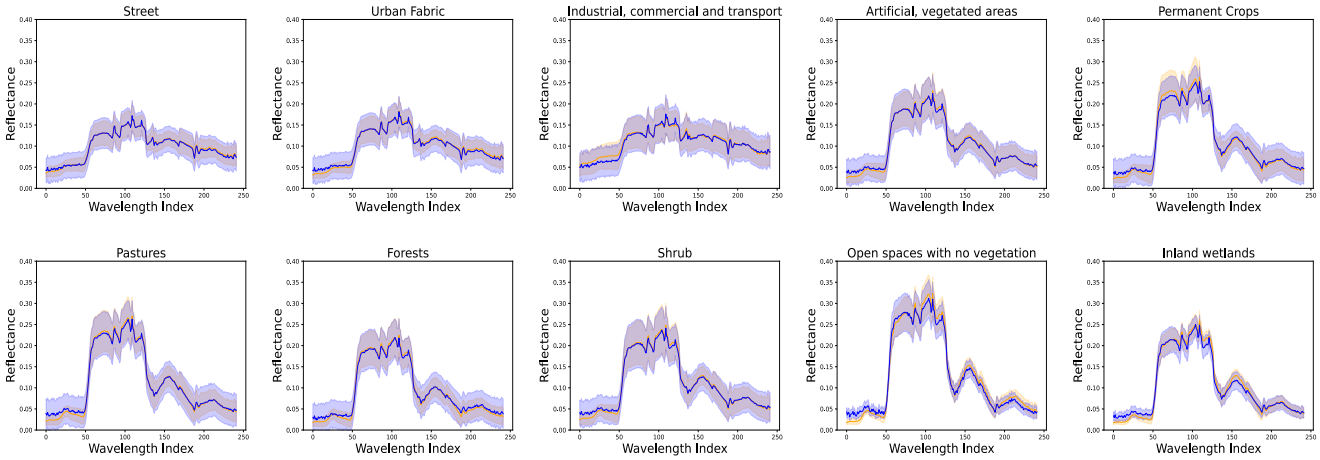
Figure 6. Generative task evaluation on HSI modality with C2Seg-AB dataset, showing original, masked, and reconstructed images across three target domain masking levels.

Table 4. Contribution of each component in our proposed frame- work to overall performance on the C2Seg-AB dataset.

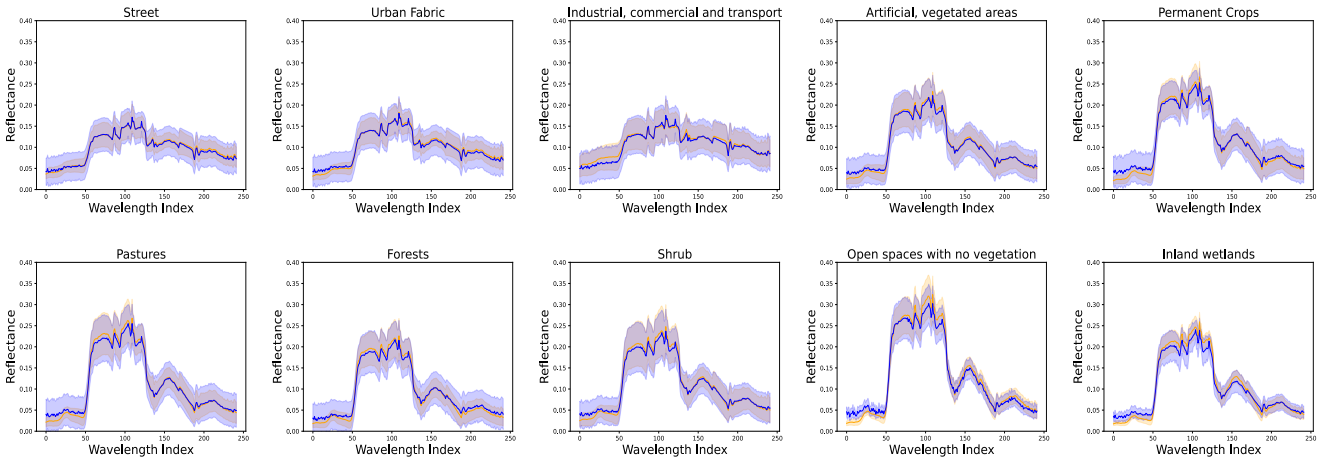
Classes	$\mathcal{L}_{Seg}$	$\mathcal{L}_{Seg} + \mathcal{L}_{DA}$	$\mathcal{L}_{Seg} + \mathcal{L}_{MAE}$	$\mathcal{L}_{Seg} + \mathcal{L}_{DA} + \mathcal{L}_{MAE}$ (Ours)
Surface water	0.4927	0.4971	0.4718	0.5138
Street	0.2400	0.3021	0.1360	0.3207
Urban Fabric	0.4555	0.6464	0.4950	0.6476
Industrial, commercial and transport	0.5101	0.7295	0.5256	0.7376
Mine, dump, and construction sites	0.5049	0.5611	0.3623	0.5949
Artificial, vegetated areas	0.4314	0.6504	0.4136	0.6615
Arable Land	0.4894	0.8324	0.4725	0.8382
Permanent Crops	0.2434	0.2569	0.1130	0.2358
Pastures	0.5354	0.6311	0.4016	0.6451
Forests	0.5686	0.6144	0.5022	0.6247
Shrub	0.3960	0.5305	0.2924	0.5404
Open spaces with no vegetation	0.0956	0.0140	0.0231	0.0217
Inland wetlands	0.3783	0.4423	0.2665	0.4503
MA(Avg)	0.4950	0.6252	0.3930	0.6381
MA(Std)	0.0621	0.0220	0.0654	0.0208
mIoU (Avg)	0.2704	0.3741	0.2245	0.3835
mIoU (Std)	0.0327	0.0174	0.0331	0.0161
mF1 (Avg)	0.4109	0.5160	0.3443	0.5255
mF1 (Std)	0.0420	0.0187	0.0482	0.0186



(a) 50% Target Domain Masking



(b) 75% Target Domain Masking



(c) 100% Target Domain Masking

Figure 7. Mean and standard deviation of the spectral distribution for the remaining classes not covered in the main text, evaluated for masked pixel classes in the target image at three different masking ratios using our MAE learning approach.

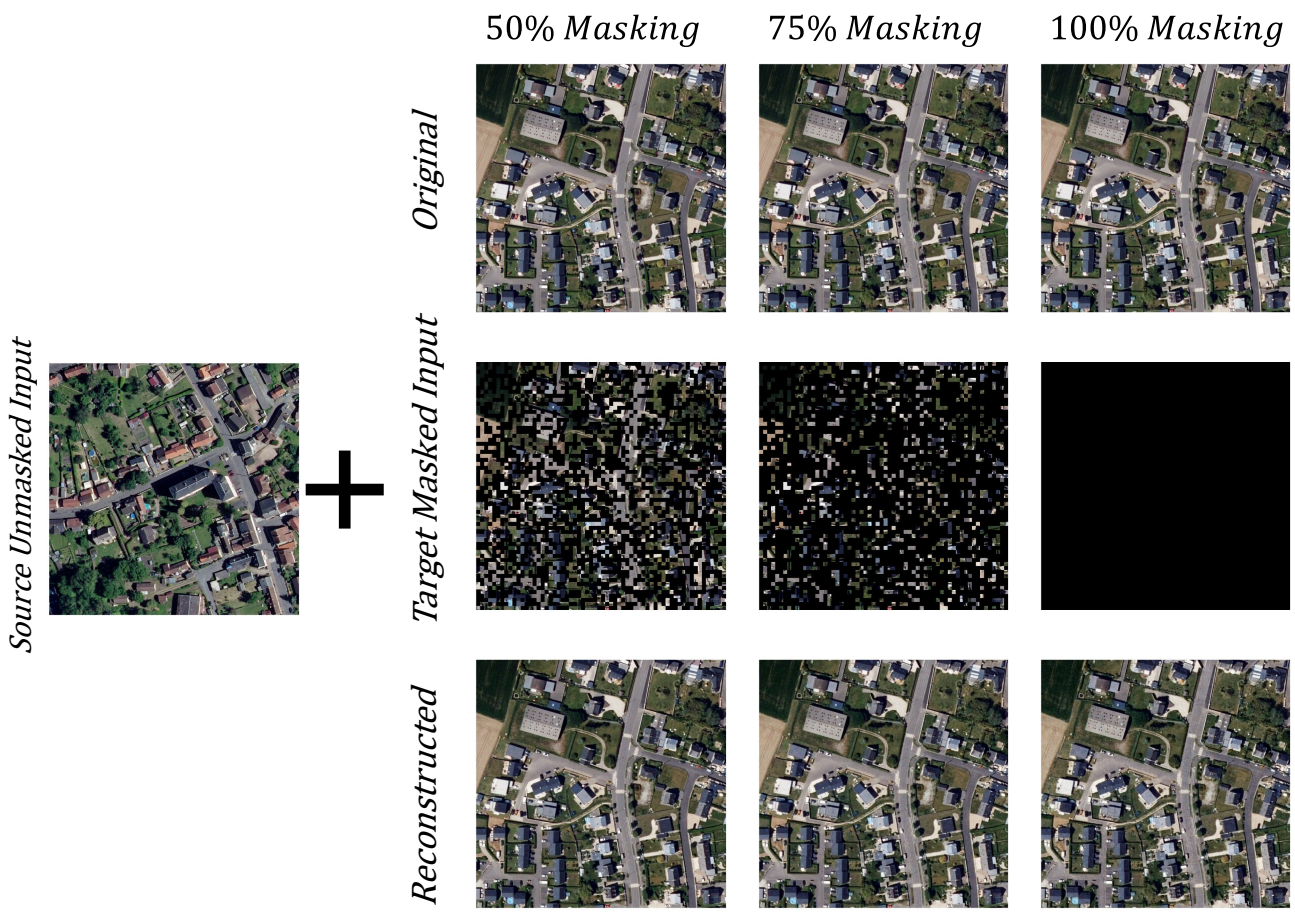


Figure 8. Generative task evaluation on MSI modality with FLAIR dataset, showing original, masked, and reconstructed images across three target domain masking levels.