An End-to-End Trainable Multi-Scale CRF Framework for Decision Fusion in Multi-Modal Remote Sensing Classification

Supplementary Material

In this supplementary paper, we provide the following analysis:

- **Loss convergence.** We present the training curves of our proposed model on the Houston HSI-LiDAR dataset in Fig. 7(a) and (b) and Augsburg dataset in Fig. 7(c) and (d) and justify the convergence of the proposed framework over the epochs.
- **Generated classification maps.** Fig. 8, 9, 10, and 11 provides the visualization of the false color composites of the two data modalities, their test labels, and their fused classification maps obtained from MFT, MSFMamba, KANfusion, and the proposed model for the Houston 2013 (MSI), Augsburg, Berlin, and Houston 2018 datasets, respectively. Table 7 shows the class legends of each dataset. The findings show that HSI–LiDAR fusion reliably outperforms both HSI–MSI and HSI–SAR fusion, highlighting its stronger capability to generate more accurate classification maps.
  Visual examination suggests that MFT tends to place greater emphasis on neighboring pixels when classifying a target pixel. As a result, colors often bleed across region boundaries, giving regions a more rounded appearance. In contrast, MSFMamba produces much sharper boundaries, likely because it assigns more weight to the central pixel within each patch. However, MSFMamba occasionally fails to detect extremely narrow boundaries—few pixels wide—between different land-cover classes. KANfusion yields comparatively better results, but still fails in segregating finer grained classes like commercial area from industrial area. Our proposed model appears to capture these fine details effectively.
- **Statistical analysis of model.** A calibrated neural network produces class probabilities that reflect the true likelihood of each class being correct. Calibration is especially important in safety-critical applications where accurate confidence estimates are essential for making reliable decisions. Fig. 5 show the reliability plots that depict the gap between accuracy and calibration, which results in miscalibration for the MSFMamba and KANfusion on the Augsburg dataset. The main paper has the miscalibration plots for our proposed model and MFT. We also report the expected calibration error (ECE) to quantify the miscalibration. A lower ECE score denotes a better calibrated model. Through the plots, we demonstrate that our model has better calibration than the existing competing models.

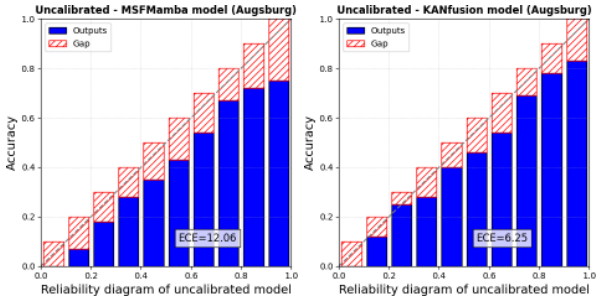
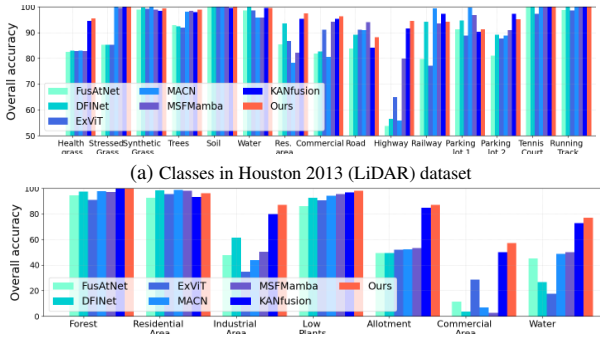


Figure 5. Calibration plot for MSFMamba and KANfusion on the Augsburg dataset.

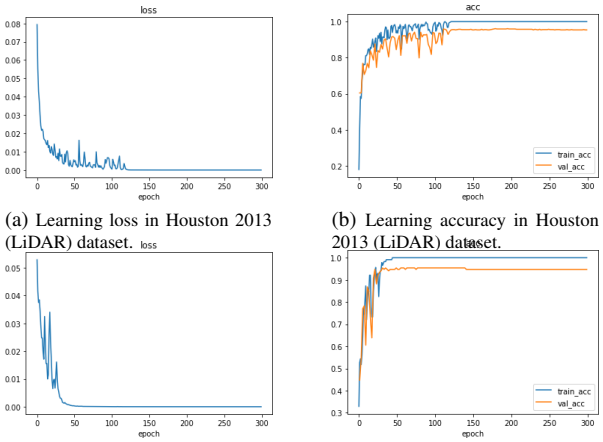


(a) Classes in Houston 2013 (LiDAR) dataset

(b) Classes in Augsburg dataset

Figure 6. Class-wise classification performance of different SOTA methods: FusAtNet, DFINet, ExViT, MACN, MSFMamba, KANfusion and ours. In some classes where SOTA fail miserably, our proposed model provides superior performance by a large margin.

- **Classwise precision.** In the Houston 2013 (LiDAR) dataset, Fig. 6(a) illustrates the class-wise results for all 15 categories and compares them with representative SOTA approaches. The highways class typically exhibits low performance in most SOTA methods; while MSFMamba and KANfusion raise its accuracy by about 15%, our model improves it by an additional 14%, reaching a precision of 94%. With the exception of `Parking lot 2`—which is often confused with `Parking lot 1`—and `Railway`, the proposed framework surpasses all baselines. The confusion between these classes is likely due to their highly similar spectral characteristics resulting from comparable material properties. Achieving more reliable separation may require additional training samples for these specific categories.



(a) Learning loss in Houston 2013 (LiDAR) dataset.

(b) Learning accuracy in Houston 2013 (LiDAR) dataset.

(c) Learning loss in Augsburg dataset.

(d) Learning accuracy in Augsburg dataset.

Figure 7. The training curve for our proposed model on the Houston HSI- LiDAR and Augsburg datasets.

Table 7. The land use/cover classes present in each of the datasets and their corresponding color legends used for generating the classification maps in Fig. 2-11.

Houston2013 HSI-LiDAR & HSI-MSI		**Houston2013** HSI-LiDAR & HSI-MSI	
Class name	**Color**	**Class name**	**Color**
Unlabelled/Background		Non-residential buildings	
Healthy grass		Roads	
Stressed grass		Highways	
Synthetic grass		Railways	
Trees		Parking lots 1	
Soil		Parking lots 2	
Water		Tennis court	
Residential buildings		Running track	
Augsburg/Berlin HSI-SAR/LiDAR		**Augsburg/Berlin** HSI-SAR/LiDAR	
Class name	**Color**	**Class name**	**Color**
Unlabelled/Background		Low plants	
Forest		Allotment	
Residential area		Commercial area	
Industrial area		Water	
Houston2018 HSI-LiDAR		**Houston2018** HSI-LiDAR	
Class name	**Color**	**Class name**	**Color**
Unlabelled/Background		Sidewalks	
Healthy grass		Crosswalks	
Stressed grass		Major thoroughfares	
Artificial turf		Highways	
Evergreen trees		Railways	
Deciduous trees		Paved parking lots	
Bare earth		Unpaved parking lots	
Water		Cars	
Residential buildings		Trains	
Non-residential buildings		Stadium seats	
Roads			

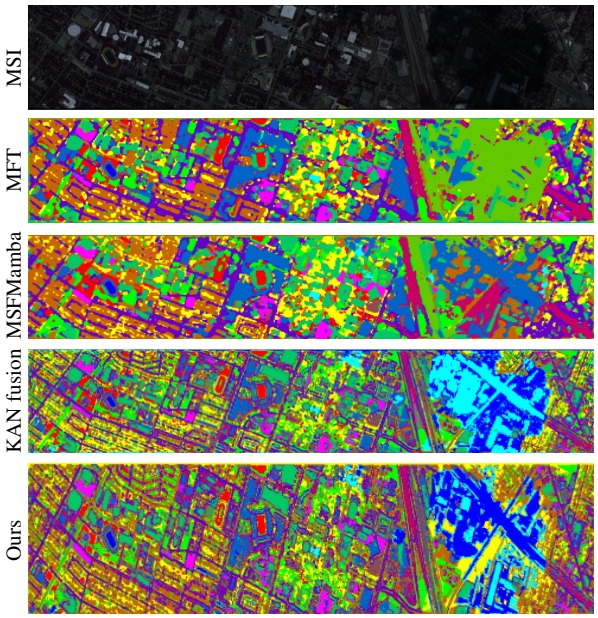


Figure 8. Classification map obtained on the Houston 2013 (MSI) dataset using MFT, MSFMamba,, KANfusion and the proposed model. The MSI FCC image was created using the bands (4, 3, 2) as RGB. The HSI and the label images are same as Fig. 2. The proposed model provides much more sharper region boundaries, as compared to SOTA.

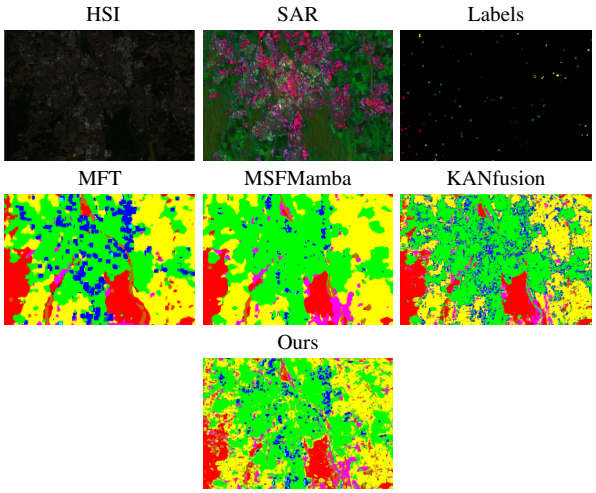


Figure 9. Classification map obtained on the Augsburg dataset using SOTA and the proposed model. The HSI and SAR FCC image was created using the bands (25, 15, 10) and features (3, 2, 1) as RGB, respectively. The proposed model provides much more sharper region boundaries.

For the Augsburg dataset, Fig. 6(b) reports class-wise performance across the 7 categories in comparison with SOTA methods. The `commercial areas` class generally shows weak performance in most existing models (precision below 30%). Although KANfusion attains a precision of 53%, our method surpasses the SOTA by an additional 3% improvement. Although the Houston 2018 dataset yields reasonably high overall accuracy (OA), a closer look at the per-class results shows that certain categories—such as `unpaved parking lots` and `crosswalks`—exhibit weak performance. Further

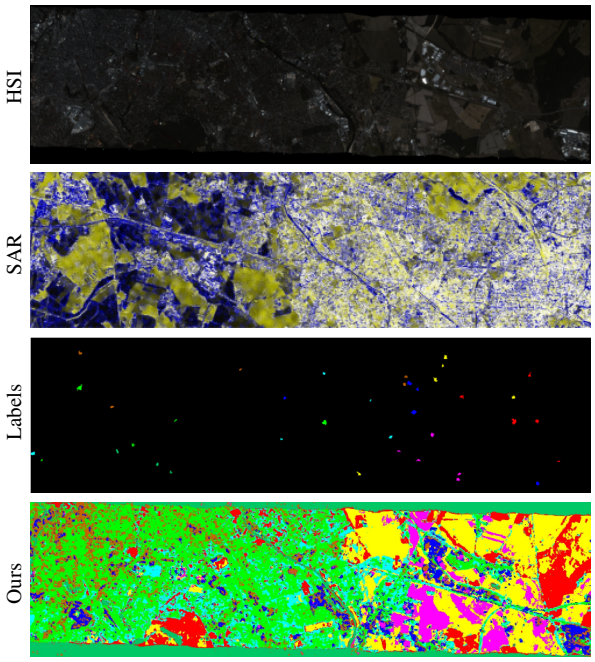


Figure 10. Classification map obtained on the Berlin dataset using the proposed model. The HSI FCC image was created using the bands (30, 20, 10) as RGB, respectively.

Table 8. Analysis of various network components and effects of different training protocols on the datasets A: Hosuton 2013 (LiDAR); B: Houston 2013 (MSI); C: Augsburg; D: Berlin & E: Houston 2018. We report the OA values.

	Task	A	B	C	D	E
Patch size	7×7	98.02	96.93	98.64	98.92	95.91
	9×9	98.27	97.39	98.92	99.24	96.17
	11×11	97.87	97.00	98.41	99.21	95.88
	13×13	97.45	97.23	98.32	98.00	95.90
	15×15	97.00	96.78	98.89	98.91	96.98
	17×17	96.97	96.22	98.45	96.33	96.95
Fusion	Modality-1 (HSI)	82.11	80.34	85.53	87.42	83.22
	Modality-2 (Others)	67.71	69.01	73.82	77.62	68.90
	Fused modalities	98.27	97.39	98.92	99.24	96.17

examination indicates that these categories are heavily underrepresented in the training set, which likely explains the model's difficulty in accurately classifying them.

- **Patch size.** Table 8 examines how different patch sizes influence the network's performance. Our earlier experiments used a 9×9 patch. For larger patches, we introduced an additional max-pool layer (kernel size 2, stride 2) before the self-attention module to keep the feature dimensions consistent. Patch sizes from 7×7 to 17×17 were evaluated. The first four datasets achieved their best results with a 9×9 patch, whereas the Houston 2018 dataset performed better with larger patches. Still, increasing patch size did not lead to a consistent improvement—particularly in overall accuracy (OA). This is likely because the training data contain sparse labels, causing the model to learn many unlabeled pixels (label 0) as if they were misclassified examples. For the Augsburg dataset, which has a ground sampling distance of 30 m, small variations in patch size produced minimal changes in OA, as shown in Table 8.
- **Individual modalities.** To better understand the specific contribution of data fusion and disentangle its influence from the rest of the network design, we performed an ablation study evaluating each modality independently. As shown in Table 8, each stan-

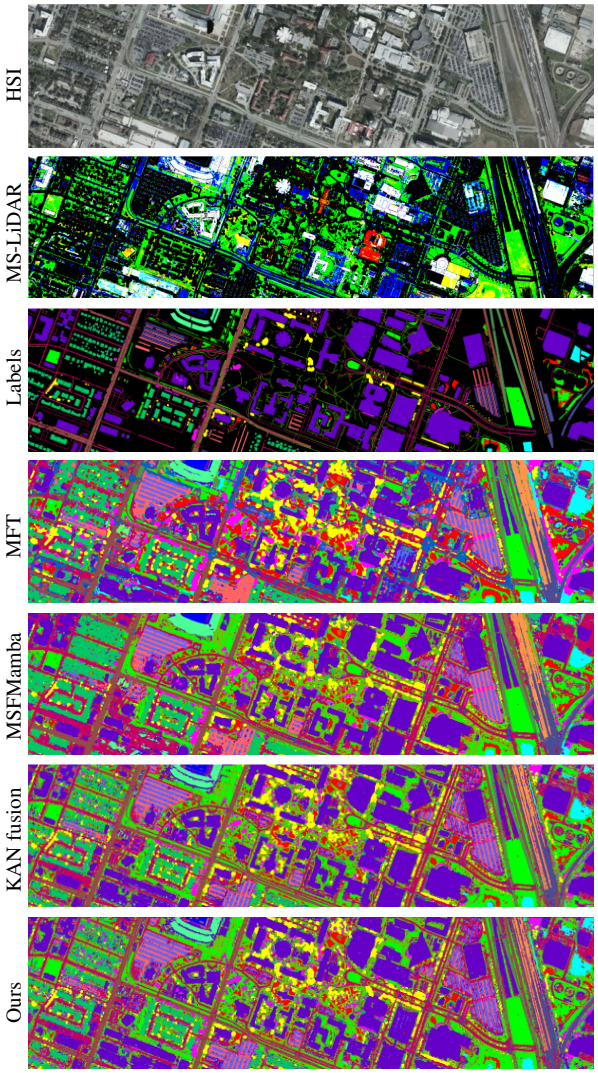


Figure 11. Classification map obtained on the Houston 2018 dataset using MFT [42], MSFMamba [20], and the proposed model. The multispectral-LiDAR image was created using the bands 1550 nm, 1064 nm, and 532 nm bands as RGB. The proposed model provides much more sharper region boundaries, as compared to MFT and MSFMamba.

Table 9. Sensitivity to hyper-parameters σ_α, σ_β, and σ_γ on the two Houston-2013 & Houston-2018 datasets.

Houston-13 HSI-LiDAR				**Houston-13** HSI-SAR				**Houston-18** HSI-LiDAR			
σ_α	σ_β	σ_γ	OA	σ_α	σ_β	σ_γ	OA	σ_α	σ_β	σ_γ	OA
0.4	0.5	0.4	97.08	0.4	0.5	0.4	97.55	0.4	0.5	0.4	96.31
4	5	4	98.27	4	5	4	97.39	4	5	4	96.17
40	50	40	98.18	40	50	40	96.95	40	50	40	95.83

dalone modality already achieves strong performance within the proposed framework. However, when the modalities are combined during inference, we observe a clear and consistent performance gain—typically in the range of 4–10% across all datasets. This highlights the complementary nature of the modalities and confirms that fusion provides substantial added value beyond what any single source can deliver.

- **Sensitivity to hyperparameters.** We performed a grid search over the most optimum values for σ_α, σ_β, and σ_γ for all the datasets and found the performance to be fluctuating less with the chosen values. Thus, the model is fairly generalizable and need not be tuned separately for each dataset for optimal performance. Table 6 and 9 report this in terms of OA.