Supplementary Material for:

UniDiff: Parameter-Efficient Adaptation of Diffusion Models for Multi-Modal Remote Sensing Land-Cover Classification with Sparse Annotations

Yuzhen Hu
University of Houston
Houston, TX, USA
yhu34@uh.edu

Saurabh Prasad
University of Houston
Houston, TX, USA
saurabh.prasad@ieee.org

A. Additional Implementation Details

Pretrained Backbone. We use the ADM U-Net model pretrained on 64×64 ImageNet from Dhariwal and Nichol [1]. The ADM backbone consists of 12 decoder layers with encoder–decoder skip connections. During adaptation, we update only the FiLM layers while keeping the backbone weights frozen. Domain adaptation uses the Adam optimizer (lr=0.003) with a cosine noise schedule, 2000 training steps, and MSE loss.

Classifier Training. Pixel-level features are extracted separately for each modality, concatenated, and passed into an MLP classifier. For Augsburg we use a 2-layer MLP, and for Berlin a 3-layer MLP to match the larger label set. Adam optimizer is used (lr=0.001, weight decay 5×10^{-4}) with early stopping applied after a minimum of 10 epochs if validation loss increases for 3 consecutive epochs.

Input Data Normalization.

- *Pseudo-RGB*: Selected HSI bands (Berlin: 40, 30, 15; Augsburg: 21, 11, 6) with 2nd–98th percentile normalization.
- *PCA-HSI*: Top 3 principal components with 2nd–98th percentile normalization.
- *SAR*: Pauli-decomposed channels with 2nd–98th percentile normalization.

Feature Extraction Settings. We use layer 10 for Augsburg and layer 11 for Berlin, consistent with GeoDiffNet-F [2]. Optimal timesteps were identified through systematic ablation: $t = 0$ for Augsburg and $t = 300$ for Berlin.

Training Details.

- Batch sizes: 32 (domain adaptation), 64 (classifier training)
- Training duration: 2000 steps (adaptation), 10 epochs (classifier)
- Hardware: NVIDIA GeForce RTX 4080 GPUs
- Software: PyTorch 2.2, CUDA 12.1
- Random seeds fixed for reproducibility

B: Performance Across Timesteps and Modalities

We provide additional analyses to complement the main results. First, we examine UniDiff's stability across denoising timesteps. Second, we compare modality configurations (HSI-only vs HSI+SAR) to demonstrate scalability. We present results separately for the Berlin and Augsburg datasets.

B.1 Berlin Dataset

Timestep Stability. Tab. 1 reports mean F1 scores as well as OA, AA, Kappa, and mIoU across selected denoising timesteps. Adapted UniDiff variants consistently outperform their non-adapted counterparts, showing stable improvements across timesteps 100–700, with peak performance observed at timestep 300 for both HSI and HSI+SAR inputs.

Modality Comparison. At optimal timesteps, UniDiff achieves strong performance with HSI-only inputs and further improves when SAR is added. Tab. 2 summarizes this comparison.

Table 1. Adaptation stability on the Berlin dataset across denoising timesteps. HSI refers to pseudo-RGB and PCA-reduced HSI input, while HSI+SAR corresponds to fusion of pseudo-RGB, PCA-reduced HSI, and SAR. Results are reported in Overall Accuracy (OA) and mean F1 score (mF1), both in %. Bold entries denote the best adapted performance at timestep 300.

Timestep	Pretrained HSI		Adapted HSI		Pretrained HSI+SAR		Adapted HSI+SAR	
	OA (%)	mF1 (%)	OA (%)	mF1 (%)	OA (%)	mF1 (%)	OA (%)	mF1 (%)
T100	71.1	60.6	72.6	59.5	75.9	63.6	77.0	64.6
T200	70.3	61.2	72.6	61.5	74.0	64.5	77.6	65.3
T300	71.0	60.9	**74.9**	**61.4**	75.2	63.7	**80.9**	**66.3**
T400	68.1	59.0	73.8	61.3	74.6	64.0	77.9	65.2
T500	70.7	59.9	72.9	60.6	73.4	62.4	74.4	63.0
T600	70.9	58.5	73.8	59.5	73.8	60.4	75.4	64.3
T700	73.5	59.3	73.1	56.7	72.1	59.7	75.1	63.1

Table 2. Berlin dataset: UniDiff performance with HSI-only vs HSI+SAR at the optimal timestep (300). All metrics are reported in %.

Metric	HSI-only	HSI+SAR
OA (%)	74.96	**80.96**
AA (%)	66.08	**68.08**
Kappa (%)	62.81	**70.05**
mF1 (%)	61.44	**66.25**
mIoU (%)	47.78	**53.34**

B.2 Augsburg Dataset

Timestep Stability. Tab. 3 reports mean F1, OA, AA, Kappa, and mIoU across timesteps. Optimal performance occurs at timestep 0 for both HSI-only and HSI+SAR.

Table 3. Adaptation stability on the Augsburg dataset across denoising timesteps. HSI refers to pseudo-RGB and PCA-reduced HSI input, while HSI+SAR corresponds to fusion of pseudo-RGB, PCA-reduced HSI, and SAR. Results are reported in Overall Accuracy (OA, %) and mean F1 score (mF1, %). Bold entries denote the best adapted performance at timestep 0.

Timestep	Pretrained HSI		Adapted HSI		Pretrained HSI+SAR		Adapted HSI+SAR	
	OA (%)	mF1 (%)	OA (%)	mF1 (%)	OA (%)	mF1 (%)	OA (%)	mF1 (%)
T0	91.87	66.94	**91.97**	**69.62**	92.92	69.12	**93.08**	**72.10**
T100	90.87	65.11	90.09	63.23	92.65	66.37	92.75	68.27
T200	89.45	62.92	89.01	62.06	92.34	66.28	93.10	65.76
T300	86.14	59.52	88.85	59.85	92.61	64.29	92.76	64.13
T400	84.78	57.33	88.36	59.75	91.79	62.30	92.79	63.49
T500	83.75	57.28	86.37	59.19	91.19	60.56	92.61	61.76

Modality Comparison. Tab. 4 compares UniDiff under HSI-only and HSI+SAR at optimal timesteps. The model shows strong accuracy with HSI-only inputs, while SAR fusion provides additional gains.

Table 4. Augsburg dataset: UniDiff performance with HSI-only vs HSI+SAR at the optimal timestep (0). Bold values indicate the better performance for each metric.

Metric	HSI-only	HSI+SAR
OA (%)	92.09	**93.17**
AA (%)	69.11	**69.99**
Kappa (%)	88.58	**90.13**
mF1 (%)	67.61	**70.11**
mIoU (%)	58.13	**60.81**

C. Detailed Component and Per-Class Results

We provide comprehensive results for different modality combinations on both datasets, including overall metrics and per-class F1 scores. These results expand upon the component analysis in Sec. 4.4, illustrating how individual modalities and their combinations contribute to joint adaptation. Berlin results are reported in Tab. 5, while Augsburg results are given in Tab. 6.

Table 5. Per-class F1 scores and overall performance on the **Berlin** multimodal dataset using features extracted from the adapted model (jointly trained for 3000 steps) at layer 11 and diffusion **timestep 300**. Best scores per row are shown in **bold**.

Class / Metric	pRGB	PCA	SAR	pRGB+PCA	pRGB+SAR	PCA+SAR	pRGB+PCA+SAR
Forest	79.72	72.77	66.37	80.26	84.91	71.46	**82.91**
Residential Area	84.28	86.68	81.96	83.79	84.97	86.80	**89.99**
Industrial Area	45.96	44.16	34.27	43.82	45.57	48.86	**52.63**
Low Plants	71.60	69.05	46.25	75.82	81.41	74.40	**82.30**
Soil	76.36	80.44	51.77	79.18	76.32	76.99	**81.96**
Allotment	14.71	32.17	24.81	26.68	23.86	**35.39**	**31.90**
Commercial Area	31.61	31.92	18.70	33.16	32.11	32.46	**33.26**
Water	40.60	69.10	61.61	70.48	65.66	74.40	**75.06**
OA (%)	73.31	74.52	64.58	72.97	75.03	75.03	**80.96**
Kappa (%)	60.05	61.77	47.94	60.74	63.09	62.63	**70.05**
AA (%)	58.89	64.43	53.00	67.81	66.15	67.39	**68.08**
mF1 (%)	55.60	60.79	48.22	61.65	61.85	62.59	**66.25**
mIoU(%)	42.33	46.62	34.22	47.93	48.63	48.31	**53.34**

Table 6. Per-class F1 scores and overall performance on the **Augsburg** multimodal dataset using features extracted from the adapted model (jointly trained for 1500 steps) at diffusion **timestep 0**, with three modalities (pRGB, PCA-compressed hyperspectral, and SAR) under joint conditional adaptation. Best scores per row are in **bold**.

Class / Metric	pRGB	PCA	SAR	pRGB+PCA	pRGB+SAR	PCA+SAR	pRGB+PCA+SAR
Forest	94.69	94.39	93.67	95.33	96.96	97.01	**97.44**
Residential Area	94.27	93.73	89.19	94.93	94.41	93.54	**94.79**
Industrial Area	64.25	57.43	28.58	**65.13**	61.00	57.20	63.39
Low Plants	96.07	94.67	94.34	96.42	97.17	96.92	**97.93**
Allotment	63.57	50.41	26.44	61.53	61.23	61.28	**69.41**
Commercial Area	18.65	11.08	15.24	19.86	21.96	15.54	**19.56**
Water	21.48	32.77	17.93	54.17	36.32	31.53	**62.18**
Overall Accuracy (%)	91.32	90.16	86.56	91.97	91.99	91.67	**93.08**
AA (%)	64.11	63.97	53.68	71.02	66.62	63.34	**71.68**
Kappa (%)	87.38	85.69	80.42	88.40	88.40	87.83	**89.97**
mF1 (%)	64.71	62.07	52.20	69.62	67.01	64.72	**72.10**
mIoU (%)	55.39	52.41	43.98	59.34	57.22	55.35	**62.36**

D. Cross-Modal Feature Organization

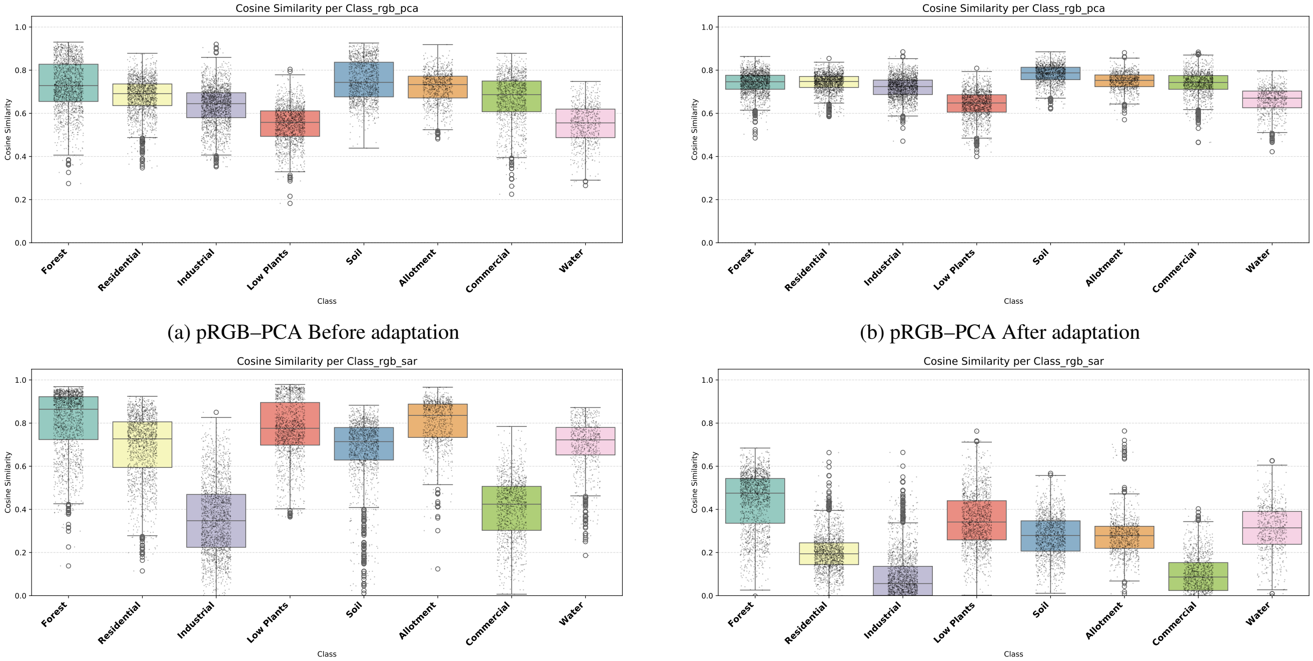
We analyze cross-modal feature organization by measuring cosine similarity between pRGB-PCA and pRGB-SAR feature representations before and after joint adaptation. This analysis provides insight into how adaptation reshapes representational relationships across modalities.

D.1 Berlin Dataset

Fig. 1 shows cosine similarity distributions across semantic classes for the Berlin dataset. Joint adaptation produces consistent improvements in pRGB-PCA organization, with similarities becoming both higher and more tightly distributed across classes. Commercial Area and Industrial Area show particularly notable improvements in consistency. For pRGB-SAR relationships, adaptation creates more structured patterns, with most classes showing reduced similarity and tighter distributions, indicating clearer differentiation between optical and structural modalities.

D.2 Augsburg Dataset

Fig. 2 presents corresponding results for Augsburg. Similar to Berlin, pRGB-PCA similarities show consistent improvement with reduced variance across semantic classes. However, pRGB-SAR patterns differ from Berlin, with some classes (e.g.,



(a) pRGB–PCA Before adaptation

(b) pRGB–PCA After adaptation

(c) pRGB–SAR Before adaptation

(d) pRGB–SAR After adaptation

Figure 1. Cosine similarity distributions across semantic classes on the **Berlin** dataset, comparing pseudo-RGB (pRGB)–PCA and pRGB–SAR features before and after joint adaptation.

Forest) showing increased organization while others maintain separation. This dataset-dependent behavior suggests that joint adaptation learns context-appropriate cross-modal relationships rather than applying uniform transformations.

D.3 Discussion

The most consistent finding across both datasets is the reduction in within-class variance for cross-modal similarities. This indicates that joint adaptation creates more reliable and predictable feature relationships, facilitating systematic multimodal fusion rather than relying on erratic cross-modal patterns. The dataset-dependent pRGB-SAR patterns demonstrate the framework's ability to learn appropriate modality relationships for different geographic and semantic contexts.

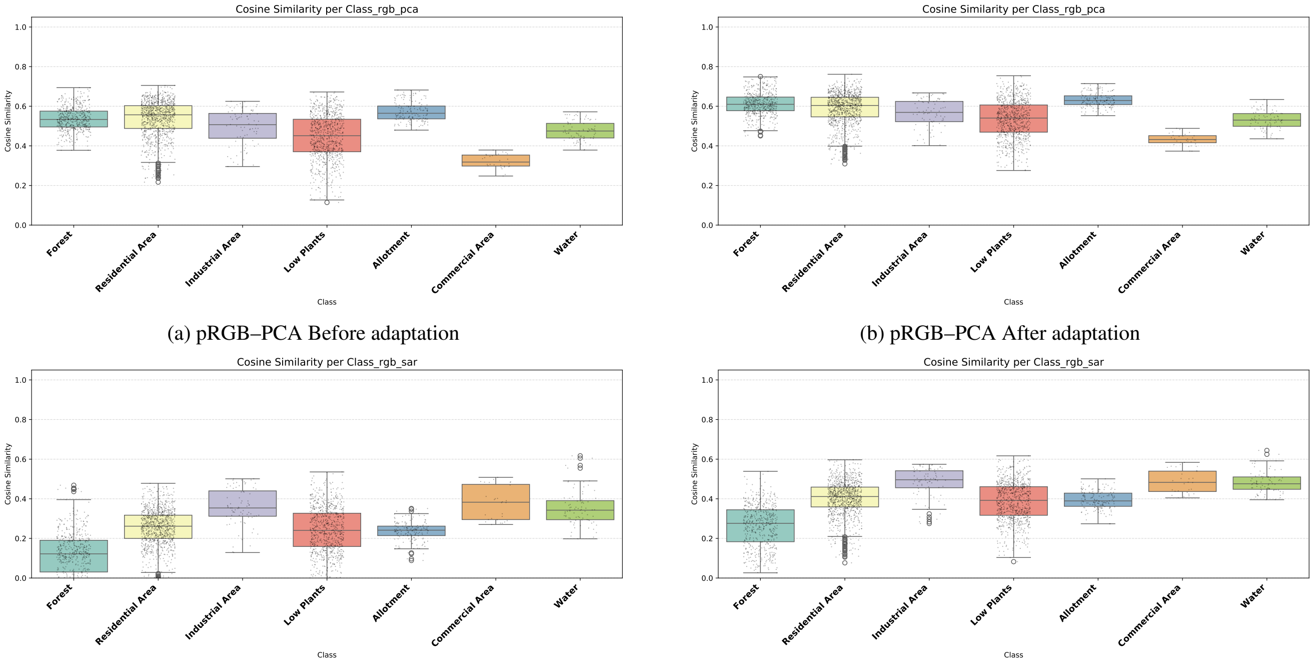
E. Code and Reproducibility

We provide an anonymized repository with training and evaluation scripts, data preprocessing routines, and pretrained checkpoints for Berlin and Augsburg. Instructions for reproducing all reported results are also included. https://github.com/hutuhehe/UniDiff-code

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[2] Yuzhen Hu, Biplab Banerjee, and Saurabh Prasad. Label-efficient hyperspectral image classification via spectral film modulation of low-level pretrained diffusion features. In *TerraBytes-ICML 2025 workshop*, 2025.



(a) pRGB–PCA Before adaptation

(b) pRGB–PCA After adaptation

(c) pRGB–SAR Before adaptation

(d) pRGB–SAR After adaptation

Figure 2. Cosine similarity distributions across semantic classes on the Augsburg dataset, comparing pseudo-RGB (pRGB)–PCA and pRGB–SAR features before and after joint adaptation.