Université Bretagne Sud

Multimodal data fusion for semantic segmentation

14.12.2022

SAAD AHMED JAMAL

JOSE GOMEZ

PROF. SEBASTIEN LEFEVRE

SURVEY PAPERS

- **Deep multimodal fusion for semantic image segmentation: A survey**
  Image and Vision Computing
  https://www.sciencedirect.com/science/article/pii/S0262885620301748?casa_token=USMUX0_tCXwAAAAA:sS71QjzaJsfO5hsPexi45ZSw02vdoZAK9qjAdnQw8wQJdFuqAaeYxLhiXUgaFybqyPfrr7u6s8E

- **Deep learning in multimodal remote sensing data fusion: A comprehensive review**
  International Journal of Applied Earth Observation and Geoinformation
  https://www.sciencedirect.com/science/article/pii/S1569843222001248

What is Semantic Segmentation?

- Semantic segmentation is a type of image analysis that involves assigning a label or category to each pixel in an image.e of image analysis that involves assigning a meaningful label or category to each pixel in an image.
- It provides a more precise understanding of the objects or features present in the scene.

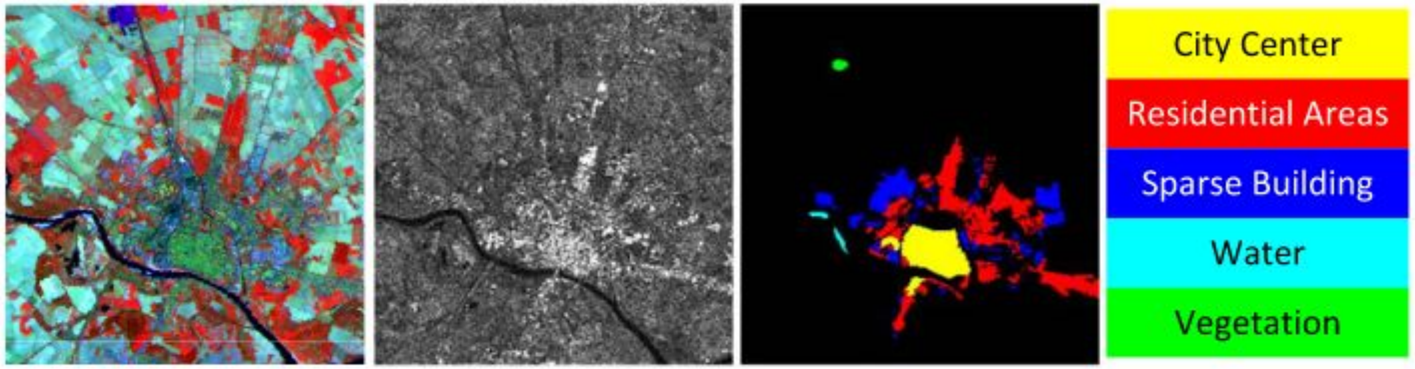


Image: Deep multimodal fusion for semantic image segmentation:A survey

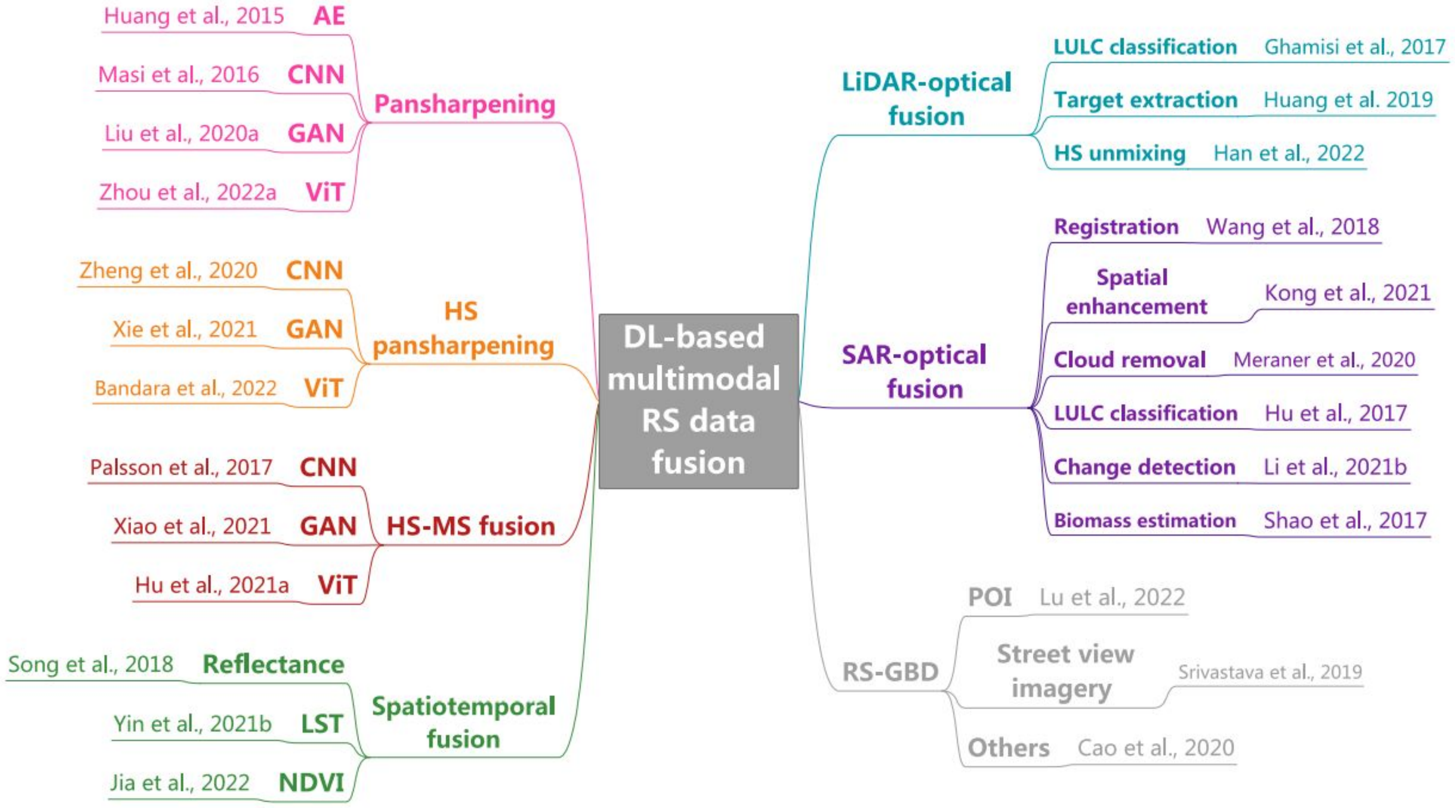
Data Fusion

Combining more than one data sets.

CASE UNDER STUDY:

In the following, we will be combining

- Hyperspectral
- LIDAR



DL-based multimodal RS data fusion

Pansharpening
Huang et al., 2015 AE
Masi et al., 2016 CNN
Liu et al., 2020a GAN
Zhou et al., 2022a ViT

HS pansharpening
Zheng et al., 2020 CNN
Xie et al., 2021 GAN
Bandara et al., 2022 ViT

HS-MS fusion
Palsson et al., 2017 CNN
Xiao et al., 2021 GAN
Hu et al., 2021a ViT

Spatiotemporal fusion
Song et al., 2018 Reflectance
Yin et al., 2021b LST
Jia et al., 2022 NDVI

LiDAR-optical fusion
LULC classification Ghamisi et al., 2017
Target extraction Huang et al. 2019
HS unmixing Han et al., 2022

SAR-optical fusion
Registration Wang et al., 2018
Spatial enhancement Kong et al., 2021
Cloud removal Meraner et al., 2020
LULC classification Hu et al., 2017
Change detection Li et al., 2021b
Biomass estimation Shao et al., 2017

RS-GBD
POI Lu et al., 2022
Street view imagery Srivastava et al., 2019
Others Cao et al., 2020

Deep learning in multimodal remote sensing data fusion:A comprehensive review

Methods

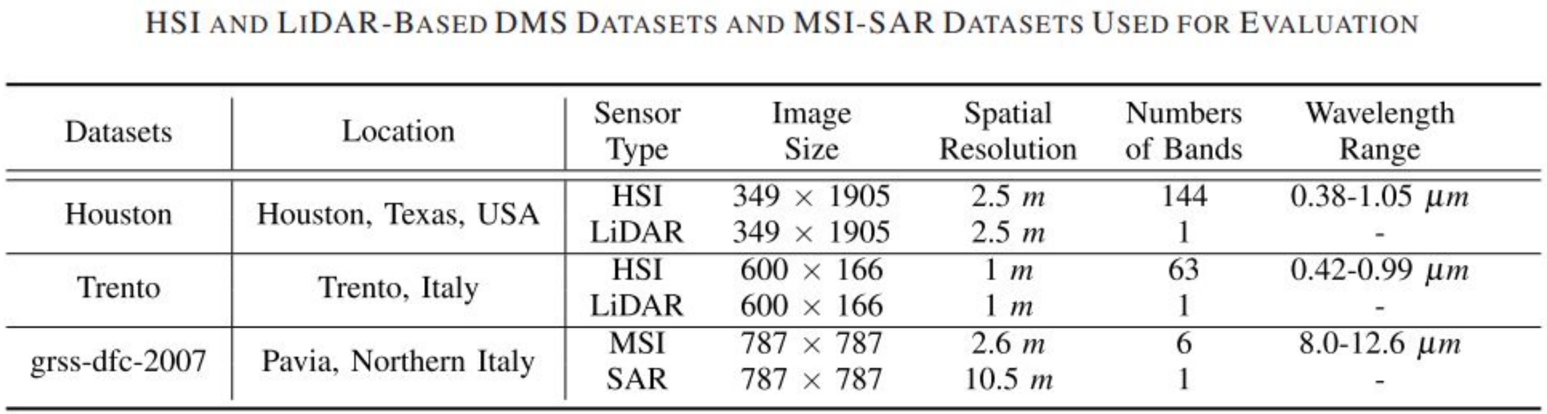
Adaptive Mutual-learning-based Multimodal Data Fusion Network (AM3Net) algorithm.

https://github.com/Cimy-wang/AM3Net_Multimodal_Data_Fusion

Deep Learning Methods (MMRS) provided by S. Fang.

https://github.com/likyoo/Multimodal-Remote-Sensing-Toolkit

BENCHMARK DATASETS



HSI AND LIDAR-BASED DMS DATASETS AND MSI-SAR DATASETS USED FOR EVALUATION

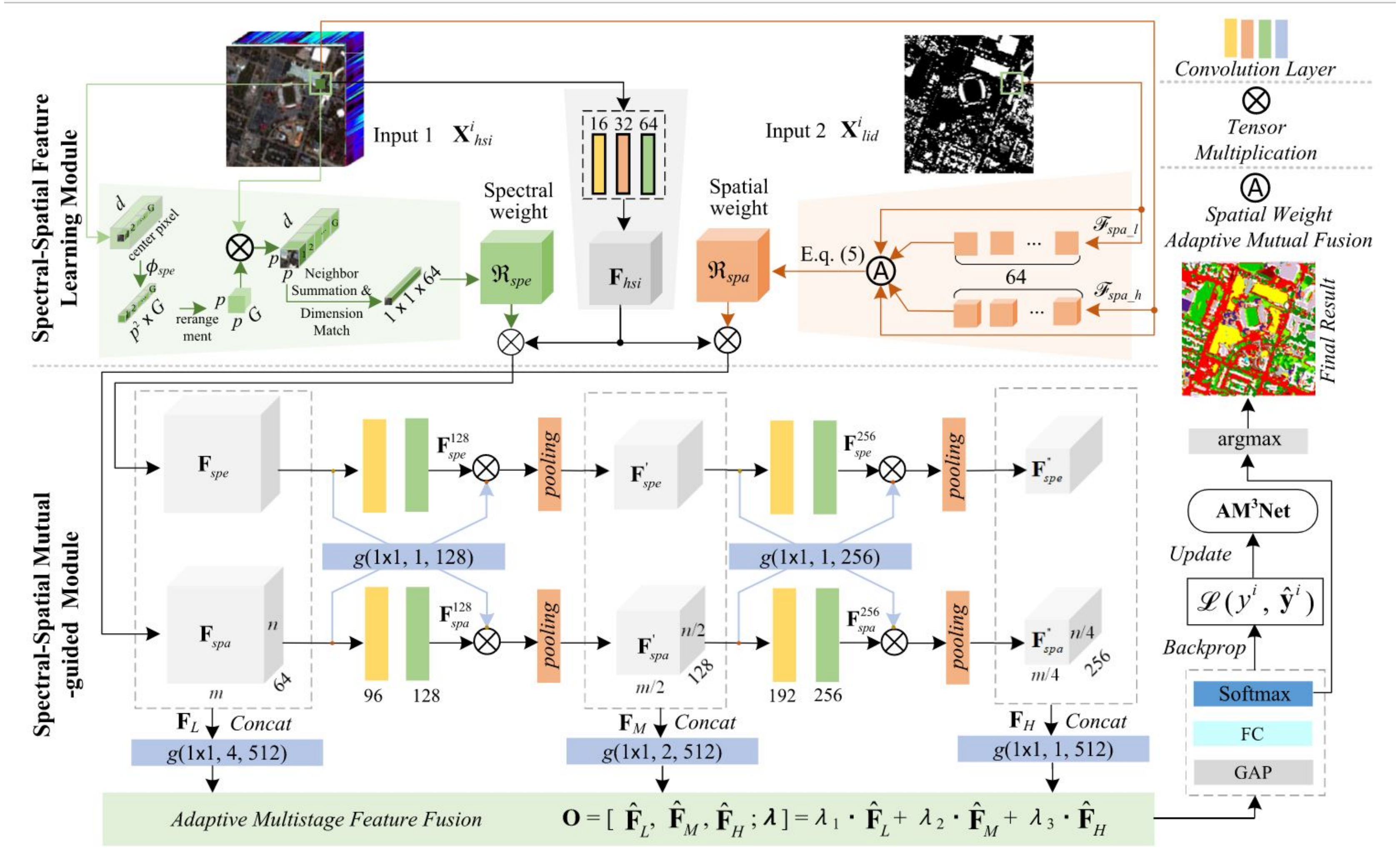
Datasets	Location	Sensor Type	Image Size	Spatial Resolution	Numbers of Bands	Wavelength Range
Houston	Houston, Texas, USA	HSI	349 × 1905	2.5 m	144	0.38-1.05 μm
		LiDAR	349 × 1905	2.5 m	1	-
Trento	Trento, Italy	HSI	600 × 166	1 m	63	0.42-0.99 μm
		LiDAR	600 × 166	1 m	1	-
grss-dfc-2007	Pavia, Northern Italy	MSI	787 × 787	2.6 m	6	8.0-12.6 μm
		SAR	787 × 787	10.5 m	1	-

Imported Trento dataset from AM3Net into MMRS.

Made it run by augmenting Trento files according to AM3Net DataLoader.

The MMRS takes input as 3 separate files for Hyperspectral (HSI), Light Detection and Ranging (LIDAR) and Ground Truth (GT) whereas in Trento Dataset from AM3Net all of these were present in the single file. To make it compatible, 3x copies of original Trento Data were made to make it compatible with MMRS input.

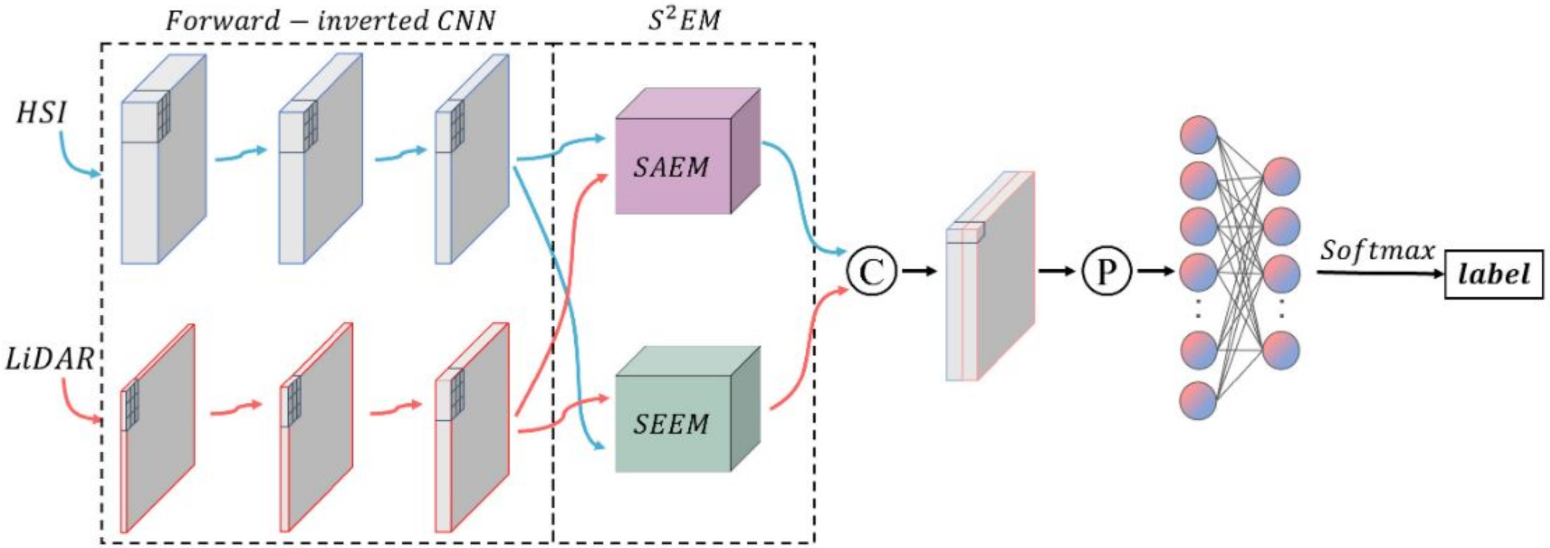
The method



Spectral-Spatial Feature Learning Module
Input 1 $\mathbf{X}^i_{hsi}$
Input 2 $\mathbf{X}^i_{lid}$
16 32 64
d
center pixel
ϕ_{spe}
$p^2 \times G$
rerangement
p
G
Neighbor Summation & Dimension Match
1 x 1 x 64
Spectral weight
$\Re_{spe}$
$\mathbf{F}_{hsi}$
Spatial weight
$\Re_{spa}$
E.q. (5)
$\mathscr{F}_{spa\_l}$
$\mathscr{F}_{spa\_h}$
64
Spectral-Spatial Mutual -guided Module
$\mathbf{F}_{spe}$
$\mathbf{F}_{spa}$
n
m
64
96
128
$\mathbf{F}^{128}_{spe}$
$\mathbf{F}^{128}_{spa}$
pooling
g(1x1, 1, 128)
$\mathbf{F}'_{spe}$
$\mathbf{F}'_{spa}$
n/2
m/2
128
192
256
$\mathbf{F}^{256}_{spe}$
$\mathbf{F}^{256}_{spa}$
g(1x1, 1, 256)
$\mathbf{F}''_{spe}$
$\mathbf{F}''_{spa}$
n/4
m/4
256
$\mathbf{F}_L$ Concat
g(1x1, 4, 512)
$\mathbf{F}_M$ Concat
g(1x1, 2, 512)
$\mathbf{F}_H$ Concat
g(1x1, 1, 512)
Adaptive Multistage Feature Fusion
$\mathbf{O} = [\hat{\mathbf{F}}_L, \hat{\mathbf{F}}_M, \hat{\mathbf{F}}_H; \boldsymbol{\lambda}] = \lambda_1 \cdot \hat{\mathbf{F}}_L + \lambda_2 \cdot \hat{\mathbf{F}}_M + \lambda_3 \cdot \hat{\mathbf{F}}_H$
Convolution Layer
Tensor Multiplication
Spatial Weight Adaptive Mutual Fusion
Final Result
argmax
AM³Net
Update
$\mathscr{L}(y^i, \hat{\mathbf{y}}^i)$
Backprop
Softmax
FC
GAP

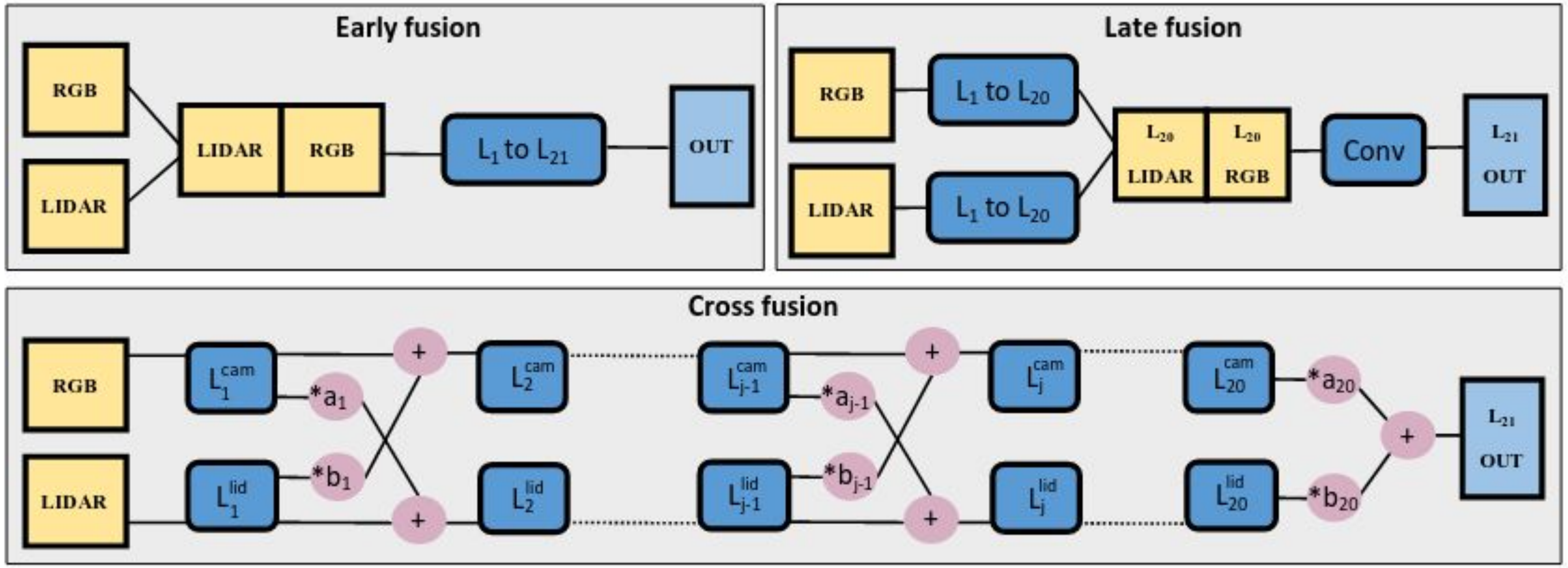
Benchmark methods

S2ENet: Spatial–Spectral Cross-Modal Enhancement Network for Classification of Hyperspectral and LiDAR Data



Benchmark methods

Cross Fusion CNN



LIDAR-Camera Fusion for Road Detection Using Fully Convolutional Neural Networks

Evaluation Metrics

- **S2ENet**

  Accuracy : **95.08%**

  Kappa: 0.9353

```
F1 scores :
      Unclassified: 0.0000
      Apple trees: 0.9986
      Buildings: 0.9054
      Ground: 0.9412
      Wood: 0.9601
      Vineyard: 0.9946
      Roads: 0.9615
```

```
Precisions :
      Unclassified: nan
      Apple trees: 0.9972
      Buildings: 0.8290
      Ground: 0.8985
      Wood: 0.9234
      Vineyard: 0.9903
      Roads: 0.9596
```

- **Cross_fusion_CNN**

  Accuracy : **99.50%**

  Kappa: 0.9934

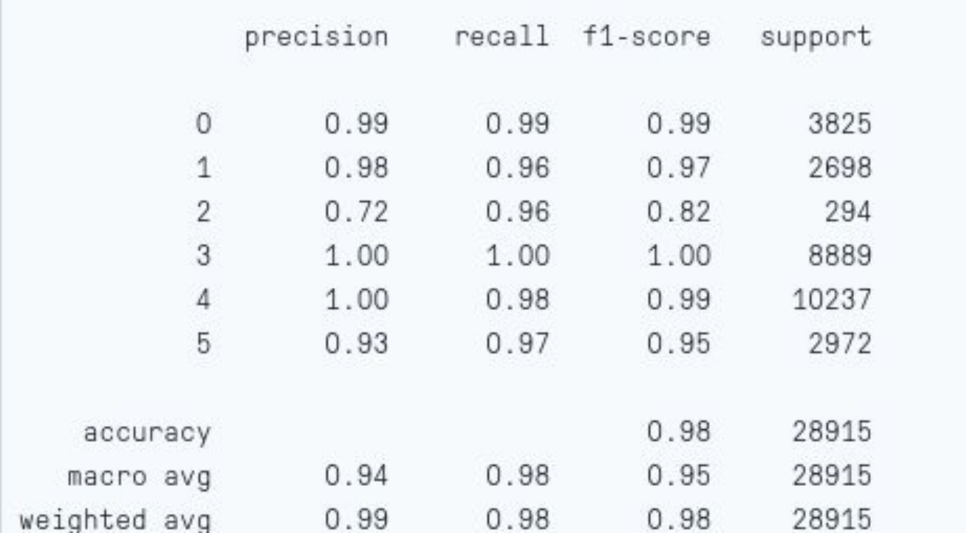
```
F1 scores :
      Unclassified: 0.0000
      Apple trees: 1.0000
      Buildings: 0.9789
      Ground: 1.0000
      Wood: 0.9979
      Vineyard: 1.0000
      Roads: 0.9925
```

```
Precisions :
      Unclassified: nan
      Apple trees: 1.0000
      Buildings: 0.9773
      Ground: 1.0000
      Wood: 0.9959
      Vineyard: 1.0000
      Roads: 0.9852
```

- **AM3NET**

  Accuracy : **98.4316%**

  Kappa: 0.96



```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      3825
           1       0.98      0.96      0.97      2698
           2       0.72      0.96      0.82       294
           3       1.00      1.00      1.00      8889
           4       1.00      0.98      0.99     10237
           5       0.93      0.97      0.95      2972

    accuracy                           0.98     28915
   macro avg       0.94      0.98      0.95     28915
weighted avg       0.99      0.98      0.98     28915
```

Evaluation Metrics (Houston)

- S2ENet

  Accuracy : 93.6542%

  Kappa: 0.9311

```
F1 scores :
      Unclassified: 0.0000
      Healthy grass: 0.9106
      Stressed grass: 0.8712
      Synthetic grass: 0.8914
      Trees: 0.9835
      Soil: 0.9976
      Water: 0.9481
      Residential: 0.9646
      Commercial: 0.8608
      Road: 0.7731
      Highway: 0.7076
      Railway: 0.9066
      Parking Lot 1: 0.7896
      Parking Lot 2: 0.9038
      Tennis Court: 1.0000
      Running Track: 0.9793
```

- Cross_fusion_CNN

  Accuracy : 88.4316%

  Kappa: 0.8744

```
F1 scores :
      Unclassified: 0.0000
      Healthy grass: 0.9078
      Stressed grass: 0.9313
      Synthetic grass: 0.9740
      Trees: 0.9850
      Soil: 0.9995
      Water: 0.8476
      Residential: 0.9268
      Commercial: 0.9341
      Road: 0.8754
      Highway: 0.9060
      Railway: 0.9413
      Parking Lot 1: 0.9079
      Parking Lot 2: 0.9308
      Tennis Court: 1.0000
      Running Track: 1.0000
```

Reported metrics

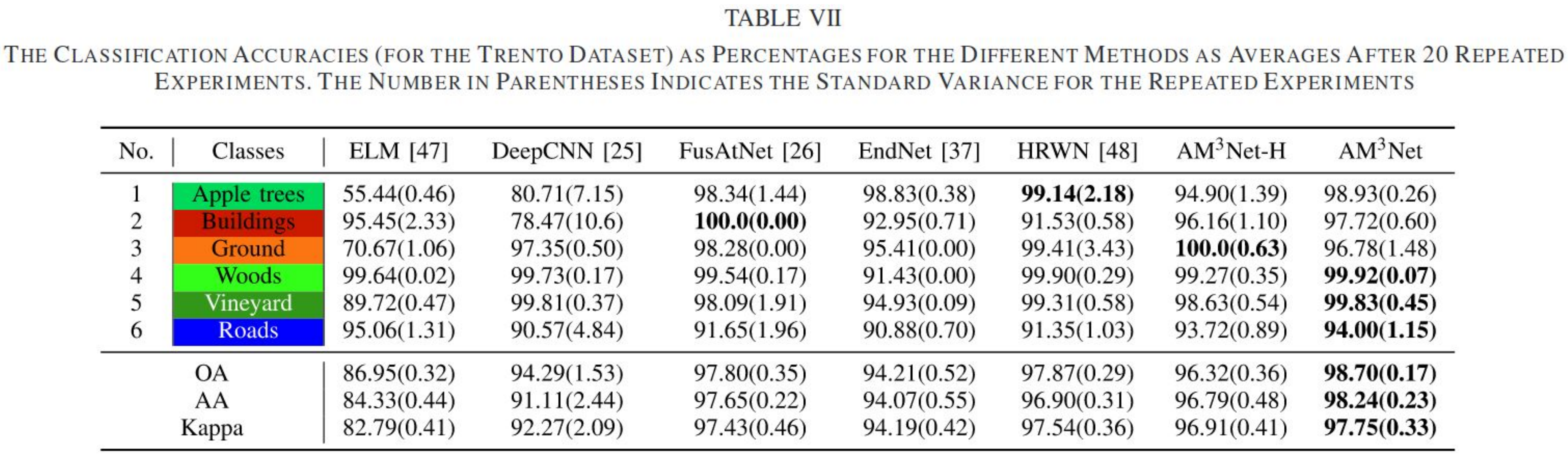


TABLE VII

THE CLASSIFICATION ACCURACIES (FOR THE TRENTO DATASET) AS PERCENTAGES FOR THE DIFFERENT METHODS AS AVERAGES AFTER 20 REPEATED EXPERIMENTS. THE NUMBER IN PARENTHESES INDICATES THE STANDARD VARIANCE FOR THE REPEATED EXPERIMENTS

No.	Classes	ELM [47]	DeepCNN [25]	FusAtNet [26]	EndNet [37]	HRWN [48]	AM^{3}Net-H	AM^{3}Net
1	Apple trees	55.44(0.46)	80.71(7.15)	98.34(1.44)	98.83(0.38)	**99.14(2.18)**	94.90(1.39)	98.93(0.26)
2	Buildings	95.45(2.33)	78.47(10.6)	**100.0(0.00)**	92.95(0.71)	91.53(0.58)	96.16(1.10)	97.72(0.60)
3	Ground	70.67(1.06)	97.35(0.50)	98.28(0.00)	95.41(0.00)	99.41(3.43)	**100.0(0.63)**	96.78(1.48)
4	Woods	99.64(0.02)	99.73(0.17)	99.54(0.17)	91.43(0.00)	99.90(0.29)	99.27(0.35)	**99.92(0.07)**
5	Vineyard	89.72(0.47)	99.81(0.37)	98.09(1.91)	94.93(0.09)	99.31(0.58)	98.63(0.54)	**99.83(0.45)**
6	Roads	95.06(1.31)	90.57(4.84)	91.65(1.96)	90.88(0.70)	91.35(1.03)	93.72(0.89)	**94.00(1.15)**
	OA	86.95(0.32)	94.29(1.53)	97.80(0.35)	94.21(0.52)	97.87(0.29)	96.32(0.36)	**98.70(0.17)**
	AA	84.33(0.44)	91.11(2.44)	97.65(0.22)	94.07(0.55)	96.90(0.31)	96.79(0.48)	**98.24(0.23)**
	Kappa	82.79(0.41)	92.27(2.09)	97.43(0.46)	94.19(0.42)	97.54(0.36)	96.91(0.41)	**97.75(0.33)**

Ablation Study

MMRS: Trento Dataset

- Training samples

  3000 → Accuracy: 99.55%

  27000 → Accuracy: 99.50% (Significantly slower)
- Batch Size

  64 → Accuracy: 98.91%

  256 → Accuracy: 83.04%
- Epochs

  100 → Accuracy: 98.91%

  500 → Accuracy: 99.23%

AM3NET: Trento Dataset

- Training samples

  3000 → Accuracy: 96.45%

  27000 → Accuracy: 96,9% (Significantly slower)
- Batch Size

  64 → Accuracy: 96,5%

  256 → Accuracy: 96.4%
- Epochs

  100 → Accuracy: 96.6%

  500 → Accuracy: 98.5%

Ablation Study

MMRS:Trento Dataset

- Different Models

  S2ENET → Accuracy: 98.91%

  Late_fusion_CNN → Accuracy: 99.15%

  Cross_fusion_CNN → Accuracy: 99.55%

- Network Architecture

  S2ENET original → Accuracy: 93.65%

  S2ENET augmentation of fully connected layers → Accuracy: 42.74%

MMRS: Houston Dataset

- Different Models

  S2ENET → Accuracy: 93.65%

  Late_fusion_CNN → Accuracy: 92.89%

  Cross_fusion_CNN → 88.43%

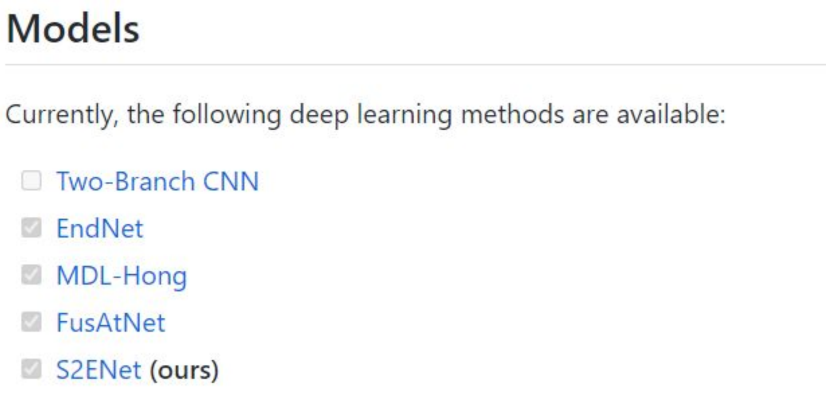
- Network Architecture

  S2ENET original → Accuracy: 93.65%

  S2ENET augmentation of fully connected layers → Accuracy: 42.74%

DIFFICULTIES FACED

- Several models were suppose to run according to the documentation but only a few of them worked.
- Tried to do the vice versa by importing houston dataset from MMRS to AM3Net algorithm but the attempt was not successful.



Models

Currently, the following deep learning methods are available:

- □ Two-Branch CNN
- ☑ EndNet
- ☑ MDL-Hong
- ☑ FusAtNet
- ☑ S2ENet **(ours)**

Papers and Github Repositories
References:

- AMM-FuseNet: Attention-Based Multi-Modal Image Fusion Network for Land Cover Mapping
- Deep multimodal fusion for semantic image segmentation: A survey
  https://www.sciencedirect.com/science/article/pii/S0262885620301748?casa_token=USMUX0_tCXwAAAAA:sS71QjzaJsf
  5hsPexi45ZSw02vdoZAK9qjAdnQw8wQJdFuqAaeYxLhiXUgaFybqyPfrr7u6s8E
- Deep learning in multimodal remote sensing data fusion: A comprehensive review
  https://arxiv.org/abs/2205.01380
- Gated Fully Fusion for Semantic Segmentation
  (https://ojs.aaai.org/index.php/AAAI/article/view/6805)
- Adaptive Mutual-learning-based Multimodal Data Fusion Network (AM3Net) algorithm.
  https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9698196

Github Repositories:
https://github.com/likyoo/Multimodal-Remote-Sensing-Toolkit
https://github.com/Cimy-wang/AM3Net_Multimodal_Data_Fusion