Supplemental Information for FuseForm: Multimodal Transformer for Semantic Segmentation

Justin McMillen
University of South Florida
Tampa, Florida, USA
jmcmillen@usf.edu

Yasin Yilmaz
University of South Florida
Tampa, Florida, USA
yasiny@usf.edu

1. Whu-Opt-SAR:

We show a more detailed comparison between FuseForm and other methods on the Whu-Opt-SAR [6] dataset in Table 1. FuseForm demonstrates superior overall performance with the highest mIoU, indicating robust segmentation quality across all classes. FuseForm excels in classifying city regions but shows some variability in other classes. The strong performance in high-percentage classes such as farmland and water (35% and 38%, respectively) combined with competitive scores in all other classes, underscores its effectiveness. While there is room for improvement in specific classes such as road and others, FuseForm's mIoU metric makes it the best-performing method in this comparison.

2. Final Expansion Layer:

The final layer of the FuseForm decoder is a linear projection layer when expands the segmentation map from $[H/4, W/4, C_1]$ to $[H/2, W/2, C_1 \times 4]$, which allows the decoder's output to attend to more pixels per output token. In this section, we explore the effects this has on decoder performance. On the MCubeS dataset with all modalities, we follow the training parameters outlined in the Experiments section. The results are shown in Table 2.

Altering the decoder output size has a notable impact on the performance on FuseForm. The half resolution $(H/2, W/2)$ configuration provides the best mIoU, suggesting that some degree of upsampling enhances performance. The model still performs competitively when the embedding layer is enabled but does not upsample the output. Upsampling by a factor of 4 to the full input resolution reduces the performance by 1.77%, showing the limits of how much a single layer can perform meaningful upsampling. Removing the embedding layer altogether decreases the mIoU by 0.96%, indicating the final expansion layer has value regardless of the output size.

3. Dataset Visualizations:

Some qualitative examples from the MCubeS [7] dataset are shown in Figure 3. Segmentation images are taken with our FuseForm model using all modalities (RGB-A-D-N).

Some qualitative examples on the Whu-Opt-SAR [6] dataset are shown in Figure 1. The first image is a composition of 9 tile images, while the other 5 are single images, enlarged to show the detail in the segmentation maps.

Figure 2 shows two samples from the Next Day Wildfire Spread dataset [5]. The first 12 columns are input to the model and the rightmost column is the ground truth for that sample.

More examples from our other two tested datasets, DeLiVER [8] and MFNet [3], are shown in Figures 4 and 5.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. 2

[3] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 1

[4] Wei Hu, Xinhui Wang, Feng Zhan, Lu Cao, Yong Liu, Weili Yang, Mingjiang Ji, Ling Meng, Pengyu Guo, Zhi Yang, and Yuhang Liu. Opt-sar-msnet: A multi-source multi-scale siamese network for land object classification using remote sensing images. *Remote Sensing*, 16:1850, 05 2024. 2

[5] Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 1, 3

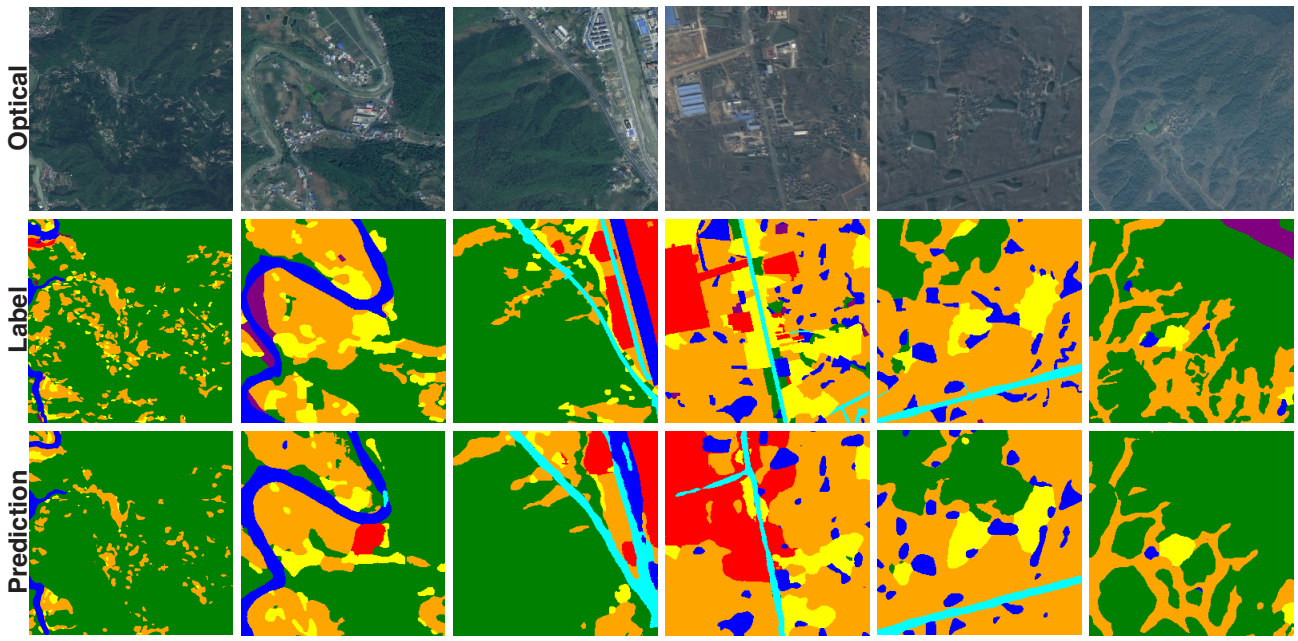

Figure 1. Example results from the Whu-Opt-SAR dataset.

Table 1. Land classification mean Intersection over Union (mIoU), as well as accuracies for each class for the Whu-Opt-SAR dataset [6]. Results are shown when each method is using all modalities.

			Class Accuracy (%)						
Method	**mIoU**	**OA**	**Farmland**	**City**	**Village**	**Water**	**Forest**	**Road**	**Others**
SegNet [1]	37.4	75.7	76.5	42.8	45.1	68.4	**96.9**	42.8	14.0
DeeplabV3+ [2]	41.2	80.9	79.5	65.8	39.3	75.2	94.2	79.0	12.7
MCANet [6]	42.9	81.7	79.7	58.8	49.7	78.6	95.8	35.2	27.2
OPTSARMSNet [4]	45.2	**84.3**	72.3	53.7	**75.9**	**79.6**	92.2	**86.8**	**28.5**
FuseForm	**48.4**	83.7	**80.4**	**73.6**	62.9	77.9	92.1	60.9	24.6

Table 2. mIoU results for the MCubeS dataset [7] when using all available modalities and altering the decoder output size through the final expansion layer.

Method	**mIoU**
H, W	52.93
H/2, W/2	**54.70**
H/4, W/4	54.60
Embedding Layer Removed	53.74

[6] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022. 1, 2

[7] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19800–19808, 6 2022. 1, 2

[8] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation, 2023. 1

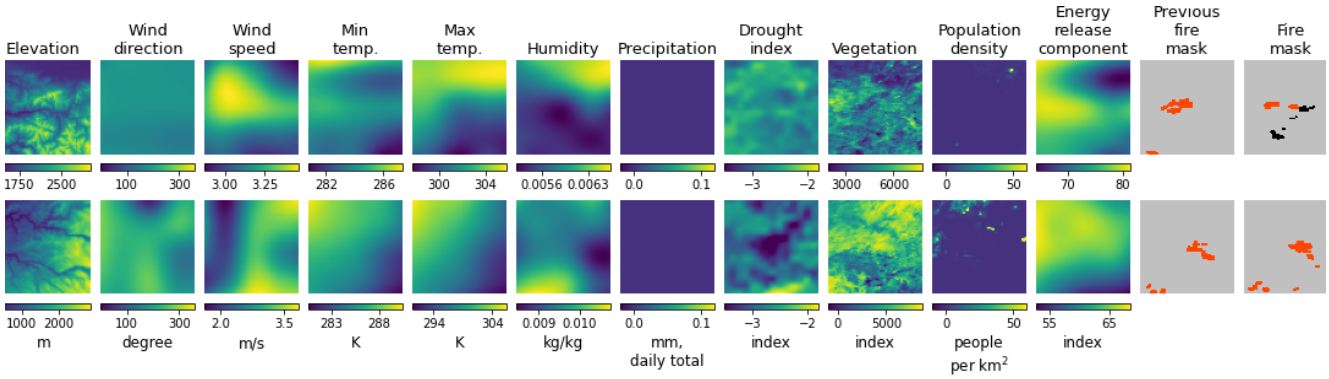

Figure 2. Two samples from the Next Day Wildfire Spread dataset from [5].

Figure 3. Example outputs from the MCubeS test set.

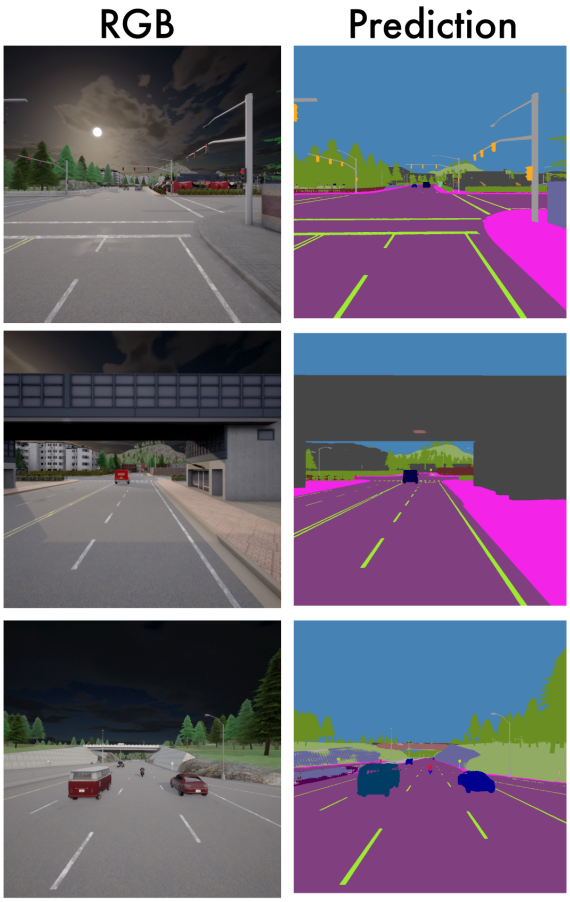

Figure 4. Example outputs from the DeLiVER validation set.

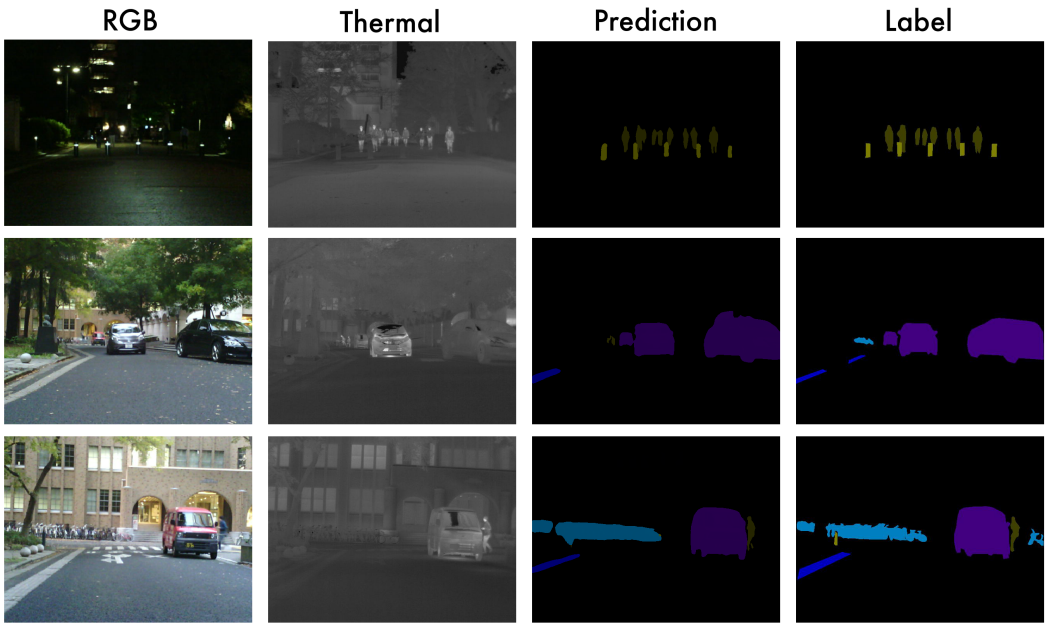

Figure 5. Example outputs from the MFNet test set.