Multimodal aerial view object classification with disjoint unimodal feature extraction and fully-connected-layer fusion

Saurav Singh[a], Manish Sharma[a], Jamison Heard[a], Jesse D. Lew[c], Eli Saber[a], and Panos P. Markopoulos[b]

[a]Rochester Institute of Technology, Rochester, NY, USA
[b]The University of Texas at San Antonio, San Antonio, TX, USA
[c]New York University, New York, NY, USA

ABSTRACT

Fusion of multimodal data can offer enhanced machine learning. One of the most common fusion approaches in deep learning is end-to-end training of a neural network on all available modalities. However, paired multimodal data from all the modalities is required to train such a network. Collecting paired data from multiple modalities can be challenging and expensive due to the requirement of specialized equipment, atmospheric conditions, limitation of individual modalities to probe a scene, data integration from modalities with different spatial and spectral resolutions, and annotation challenges for obtaining ground truth. A two-phase multi-stream fusion approach is presented in this work to counteract this issue. First, we train the unimodal streams in parallel with their own decision layers, loss, and hyper-parameters. Then, we discard the individual decision layers, concatenate the last feature map of all unimodal streams, and jointly train a common multimodal decision layer. We tested the proposed approach on the NTIRE-21 dataset. Our experiments corroborate that in multiple cases, the proposed method can outperform the alternatives.

Keywords: Aerial image classification, limited multimodal data fusion, remote sensing.

1. INTRODUCTION

Aerial imagery has become a popular data source for various remote sensing applications such as classification, detection, urban planning, disaster management, and surveillance.[1–5] Object classification is an essential task in this domain, where the goal is to identify and categorize objects of interest in aerial images automatically.[2,6] With the increasing availability of aerial imagery data, there has been a growing interest in developing automated methods for object classification.[7] Recent advancements in machine learning techniques, particularly in deep learning (DL) and computer vision, have significantly improved the accuracy and efficiency of object classification in aerial imagery.[7,8] However, object classification in aerial imagery is a challenging problem due to the complexity of the scene, occlusion issues, and the high variability in object appearance.[9,10] Thus, relying on a single sensing modality is insufficient to achieve high classification accuracy.

There has been a trend towards incorporating multiple modalities in aerial object classification to improve performance and robustness.[11–14] This includes different types of imagery, such as RGB, infrared, hyperspectral, multispectral, synthetic aperture radar (SAR), and light detection and ranging, as well as other data sources such as terrain maps and building footprints. Multimodal approaches have shown promising results, as they leverage complementary information from different sources to provide a more complete and accurate representation of the scene.[11,15] However, collecting data from multiple modalities can be challenging and expensive due to the

Further author information: (Send correspondence to Saurav Singh or Manish Sharma)
Saurav Singh: E-mail: ss3337@rit.edu
Manish Sharma: E-mail: ms8515@rit.edu
Jamison Heard: E-mail: jrheee@rit.edu
Jesse D. Lew: E-mail: jl8429@nyu.edu
Eli Saber: E-mail: esseee@rit.edu
Panos P. Markopoulos: E-mail: panagiotis.markopoulos@utsa.edu
*Saurav Singh and Manish Sharma contributed equally to this work.

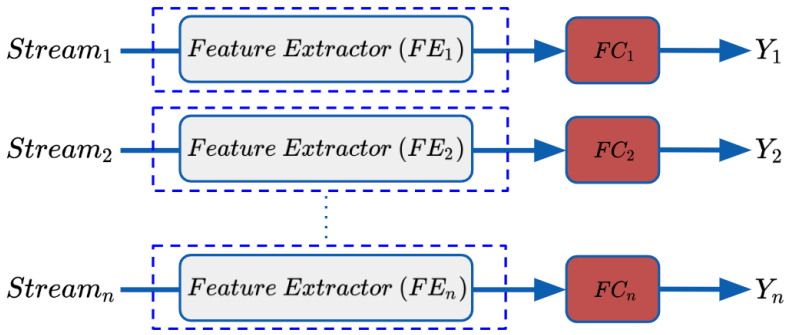

Figure 1: Unimodal networks for n unimodal streams with their corresponding feature extractors in blue dashed outlined boxes.

requirement of specialized equipment, atmospheric conditions, limitation of individual modalities to probe a scene, data integration from modalities with different spatial and spectral resolutions, and annotation challenges for obtaining ground truth.[16,17] Researchers often encounter the issue of limited *paired* multimodal data for training an end-to-end multimodal fusion network,[16] where paired multimodal data samples simultaneously agree with the following conditions: 1) spatial and temporal correspondence and 2) synchronous occurrence/availability.

Based upon these underlying conditions, the following research questions arise: 1) is it feasible and computationally advantageous to fuse legacy unimodal networks pre-trained on unpaired multimodal data samples and 2) what are the optimal approaches to train a fusion network if all paired multimodal data is available? Example approaches for training fusion networks are: i) training a multimodal network from scratch on all the paired multimodal data[11,18–20] and ii) training unimodal networks in phase one and using the corresponding feature extractors along with a fraction of paired multimodal data to fuse them in phase two. This leads to needing to understand what training paradigms perform the best (e.g., freezing or not freezing the feature extractor weights, data split between phase one and phase two) in terms of accuracy and robustness against different noise profiles.

This work focuses on multimodal limited data fusion; a way to fuse multiple modalities with a limited number of paired multimodal data using disjoint pre-trained unimodal feature extractors for object classification in aerial imagery. Our approach is similar to transfer learning in the unimodal case.[21,22] A two-phase multi-stream fusion approach is presented that fuses legacy unimodal networks with limited paired multimodal data. Our approach avoids retraining multimodal networks from scratch by using pre-trained unimodal feature extractors and fusing them using limited paired multimodal data via the standard concatenation fusion method before the final decision layer. To the best of our knowledge, our approach of using pre-trained unimodal networks trained on unpaired data has not been explored previously for feature-level multimodal fusion. The proposed approach requires significantly less computation power and training time. Using this approach, we strive to explore previously raised research questions.

Section 2 of this paper describes the unimodal legacy networks, most common fusion approach used for multimodal fusion with availability of new multimodal data, and finally introduces the proposed limited multimodal fusion method. In Section 3, we introduce the multimodal aerial dataset, baseline DL network, and experimental hyperparameters used during experimentation phase. This section also describes different multimodal fusion network training configurations for comparative studies, their results followed by discussion. Lastly, in Section 4, we summarize our contributions and important results.

2. METHODOLOGY

The limited paired multimodal data issue acts as a bottleneck towards training an end-to-end multimodal DL network. In this section, we present a two-phase multi-stream fusion approach to counteract this problem.

Generally, for a modality steam i, an individual disjoint unimodal network (shown in Figure 1), consists of a trainable feature extractor $FE_i(\theta_i)$, composed of several convolutional and pooling layers. These layers learn to extract features as an intermediate output followed by a non-trainable reshaping/vectorization operation

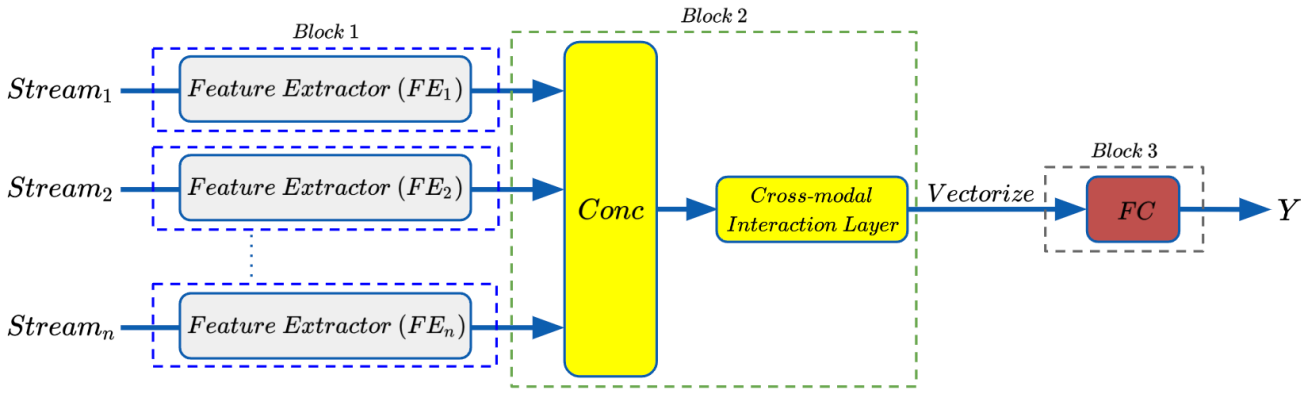

Figure 2: Multimodal fusion network.

ultimately leading to a trainable fully-connected layer (FCL) $FC_i(\phi_i)$ that classifies the overall data. In this paper, we call such modality-specific unimodal networks legacy networks. These pre-trained legacy networks are often neglected when new multimodal data becomes available (e.g., introduction of new sensors).[11,18–20] Such approaches not only discard the prior knowledge in the form of those legacy networks, but also consume significant computational resources and training time.

We utilize the feature-level fusion method, fusing n distinct unimodal data modality streams as shown in Figure 2. For the first phase, we take the n trained feature extractors $\{FE_i(\theta_i)\}_{i=1}^{n}$ denoted by *Block 1* and discard their corresponding decision layers. In the second phase, these n feature embeddings are concatenated and passed through a trainable layer to learn cross-modal interactions $CMI(\theta)$ represented by *Block 2*. The fused modality streams are followed by a non-trainable reshaping/vectorization operation ultimately leading to a standard trainable FCL $FC(\phi)$ denoted by *Block 3* to classify the data.

Multimodal paired data are required only to train the cross-modal interaction layer and a standard FCL, resulting in a small fraction of trainable parameters. Moreover, our approach avoids retraining the whole multimodal joint-stream network from scratch, saving computational resources and training time.

3. EXPERIMENTAL STUDIES

In this section, we present experimental studies for the proposed fusion approach along with comparative studies for unimodal networks and joint stream end-to-end trainable multimodal network.

3.1 Dataset & Network

The NTIRE-21 dataset[23] is an aerial imagery dataset that consists of two modalities, Electro-optical (EO) and Synthetic Aperture Radar (SAR). Aerial and satellite images are usually captured using EO sensors which depict an aerial view as the human eye perceives it. However, EO is susceptible to occlusion (e.g., clouds) or poor lighting (e.g., night time); thus, additional modalities such as Synthetic Aperture Radar (SAR) or Infrared (IR)[4,5] are required to support object detection algorithms.[23] This presents an interesting challenge, since EO is the dominant modality with more information in normal weather conditions. The NTIRE-21 dataset presents a 10-class classification problem comprised of Sedan, SUV, Pickup Truck, Van, Box Truck, Motorcycle, Flatbed Truck, Bus, Pickup Truck with Trailer, and Flatbed Truck with Trailer.

The EO and SAR unimodal legacy networks used for the NTIRE-21 dataset consist of a feature extractor with ResNet18[24] as the backbone. The extracted features are flattened and fed to the decision layer which consists of FCLs for class prediction. The EO-SAR multimodal network used for the NTIRE-21 dataset consists of two branches (one for each modality) with ResNet18 as the backbone of the feature extractor as shown in Figure 2, Block 1. The extracted features are flattened and concatenated to fuse the features from each modality (In-Network data fusion) as depicted in Figure 2, Block 2. The fused data is fed to the decision layer which consists of FCLs for class prediction, Figure 2 Block 3.

The networks were trained using Adam optimizer with 0.001 as the learning rate over 250 epochs. Since the dataset is imbalanced, where class Sedan has the most number of samples, and class Bus has the least number of samples, 624 samples (number of samples in the class with the least number of samples) were used from each class. 524 samples from each class were used as the training set, and 100 samples from each class were used as the test set.

3.2 Experiment Configurations

The proposed approach was validated by following the two-phase process described in Section 2 with the NTIRE-21 dataset. The first phase consists of training the EO and SAR unimodal networks with a fraction of the training set. The unimodal networks were trained with 20%, 40%, 60%, and 80% of the samples for the first phase. The second phase consists of training the EO-SAR multimodal dataset with the remaining 80%, 60%, 40%, and 20%, respectively. The EO-SAR multimodal network's feature extractors for each modality (Block 1 in Figure 2) are initialized with the EO and SAR unimodal network weights, respectively. This is done to utilize pre-trained legacy unimodal networks and fuse them with limited paired multimodal data in the second phase.

Two different configurations in phase two were used for the multimodal network: Frozen and Non-Frozen. The frozen configuration consists of frozen or non-trainable branched feature extractors (Block 1 in Figure 2). On the other hand, the entire network is trainable in the non-frozen configuration. As a baseline, EO unimodal network, SAR unimodal network, and EO-SAR joint stream multimodal network are trained with 100% of the training data. Table 1 summarises the different training configurations.

Table 1: Limited fusion training configurations for a multimodal network, Figure 2

Configuration	Block 1 Initialization	Block 1 Frozen	Block 2 Initialization	Block 2 Frozen	Block 3 Initialization	Block 3 Frozen
Unimodal	Random	No	N/A	N/A	Random	No
Joint Stream	Random	No	Random	No	Random	No
Frozen	Unimodal weights	Yes	Random	No	Random	No
Non-Frozen	Unimodal weights	No	Random	No	Random	No

Effects of noise on fusion networks were further studied by introducing zero-mean Gaussian noise with incrementally higher variance during inference to either one of the modalities or both modalities. Minimal Gaussian noise was also introduced while training the network (zero-mean Gaussian noise and variance 0.02). The effects of the noise were studied by observing the network classification accuracy on the testing set.

In summary, the independent variables for this experimental study consist of noise type, training data split, and fusion training configuration. The dependent variables are network accuracy and training time.

3.3 Results

Gaussian noise with incremental noise was added to either one or both modalities during inference to study the effects of noise on the fusion network performance. The results in Figure 3 show that unimodal EO and unimodal SAR performance declines drastically when noise is added to those modalities, which was expected. In the absence of noise, EO performs better than the SAR modality. EO-SAR multimodal fusion network with concatenation fusion block performance as good as a unimodal EO network in the absence of noise. However, when noise is introduced in the dominant EO modality, the performance of the unimodal EO network dips significantly while the fusion network still performs well. When noise is added to the non-dominant SAR modality, the multimodal network's performance stays unaffected. When noise was introduced in both modalities, the multimodal network performance drops significantly with the performance curve lying between unimodal EO and unimodal SAR performance curves.

The proposed limited fusion training method was validated by training EO and SAR unimodal networks on a fraction of training data and the knowledge was transferred to a multimodal fusion network by training the network on the remaining fraction of training data. Figure 4 shows the accuracies of the networks trained using the proposed method with training data splits of 20%-80%, 40%-60%, 60%-40%, and 80%-20% with different noise profiles during training. Gaussian noise with 0.0 mean and 0.06 variance was added to the noisy modality during inference for the experiment.

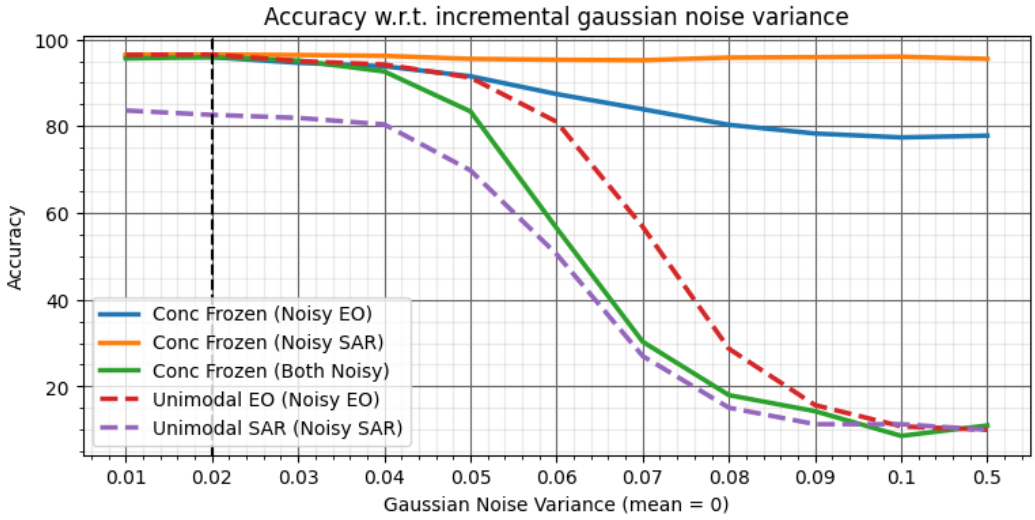

Figure 3: Effects of introducing noise in either one or both modalities on the fusion network during inference. The black vertical line represents the amount of noise introduced in the training data.

The network accuracy of EO-SAR Frozen pre-trained Nets increases as the training data split goes from a left-sided split to a right-sided split. Frozen pre-trained Nets feature extractors are trained only during the first phase with the first split of the training data, followed by training only the decision block during the second phase with the second split of the training data. Hence, this behavior is observed due to more data being available to train the feature extractor during the first phase on the right-sided training data split, 80%-20% generally performing the best. The network accuracy of EO-SAR Non-Frozen pre-trained Nets decreases as the training data split goes from a left-sided split to a right-sided split. Non-Frozen pre-trained Nets feature extractors are trained during the first phase with the first split of the training data, followed by further training of the feature extractors and decision block during the second phase with the second split of the training data. Hence, this behavior is observed due to training both - feature extractors and the decision block - on more data during the second phase on the left-sided training data split, 20%-80% generally performing the best. The less amount of data available for fine-tuning the feature extractor while training the decision-making blocks also adds the risk of overfitting the network over a small amount of data.

Since the EO modality is the dominant modality, the performance of unimodal EO is generally the best performance, except in the case of noisy EO modality. The unimodal EO network performance is closely followed by the joint stream (EO-SAR multimodal network trained end-to-end on the entire training dataset), frozen, and non-frozen pre-trained networks. The proposed approach outperforms the unimodal and joint stream end-to-end training methods in the case of noisy dominant (EO) modality. These trends; however, break down in case of the presence of noise in both modalities.

There is a clear benefit of limited fusion of pre-trained nets, since the training time of the multimodal network with the proposed approach is significantly less than training the entire fusion network from scratch, shown in Figure 5. Pre-training (light green) time is the average time to train the unimodal EO and SAR networks during the first phase. The pre-training time increases as the size of the per-training dataset increase, i.e., the training data split goes from a left-sided split to a right-sided split. Frozen and Non-Frozen training time decrease as the size of the fusion dataset decreases. The overall training time for pre-trained networks (pre-training time + fusion time) is significantly lower than that of the EO-SAR joint stream multimodal network. The total training time with Frozen increases as the splits goes from left to right while the total training time with Non-Frozen decreases as the splits go from left to right. This is due to the small number of trainable parameters with the Frozen network when compared to the Non-Frozen network.

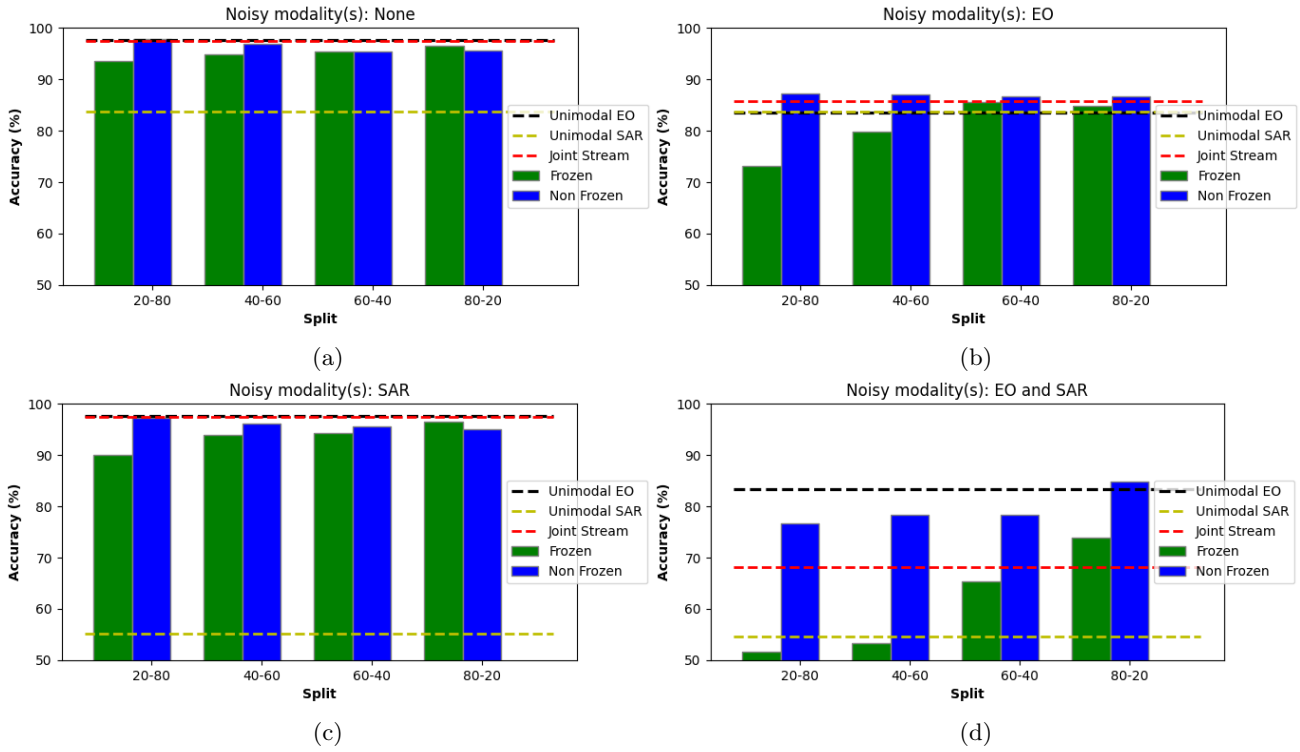

Figure 4: Network accuracy of the Unimodal EO, Unimodal SAR, Joint Stream training, Frozen pre-trained network and Non-Frozen pre-trained network in different data splits and presence of noise in different modalities during inference.

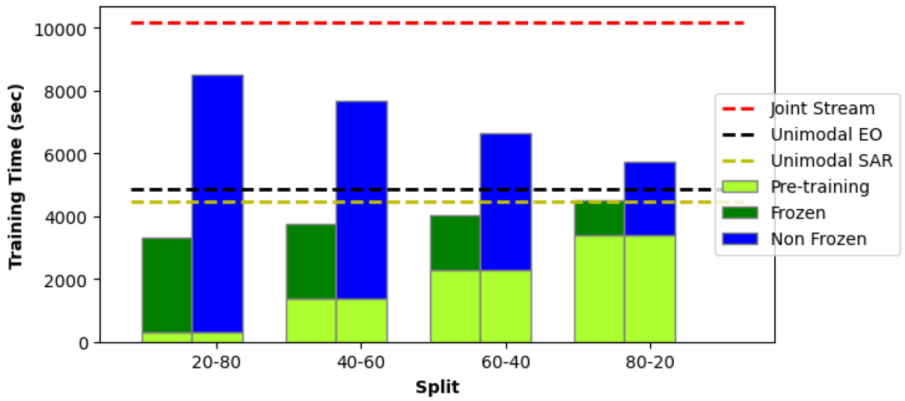

Figure 5: Total time in seconds to train the Unimodal EO, Unimodal SAR, Joint Stream training, Frozen pre-trained network and Non-Frozen pre-trained network. Pre-training (light green) time is the average time to train the unimodal EO and SAR networks for that training data split.

3.4 Discussion

Data collected from remote sensors can be corrupted or noisy due to inherent sensor noise and errors introduced during data transmission. Some image processing techniques can also introduce a loss of information and introduce noise to the data. Results in Figure 3 show that if there is noise present in one or the other modality, fusion networks can utilize the information from the other modality and still perform better than unimodal networks. This shows the merit of using multimodal networks over unimodal networks if the data is noisy.

The presented limited fusion training method presents a way to fuse legacy unimodal networks trained on

unpaired data into a multimodal network using limited paired multimodal data. Figures 4 and 5 show that the training time of the multimodal network with the proposed method is significantly less than joint stream end-to-end training of the multimodal network; however, there is a small yet acceptable drop in the performance accuracy. This also enhances the usability of the legacy unimodal networks while transitioning to the multimodal sensing paradigm. If all the paired multimodal data is available for training, a multimodal network can be trained from scratch on all the paired or joint stream multimodal data, or it can be trained in two phases using the proposed method. The networks trained using the two methods perform similarly to each other; however, the total time required to train the network using the proposed method is much less than the traditional end-to-end training method. The reduction in the training time of the multimodal network using the proposed method is more significant than it appears in Figure 5. Since the approach aims to use legacy unimodal networks, the training process will only include the second phase of the proposed method (shown in blue and dark green in Figure 5). The total training time for the proposed method also depends on how the training data is split for training the unimodal networks in phase one and the multimodal fusion network in phase two.

The second phase of the proposed training method can be carried out in two configurations: 1) Frozen, where each modality's feature extractor parameters in the multimodal network are frozen and only the decision layer is trained, or 2) Non-Frozen, where the entire multimodal network is trained including the feature extractor and the decision layer. The data split is an important hyperparameter with the proposed training method since more data is used to train each modality's feature extractors with a right-sided split (60%-40%, and 80%-20% splits) while more data is used to train the decision layers (and fine-tune the feature extractor in case of Non-Frozen network) with a left-sided split (20%-80%, and 40%-60% splits), see Figure 5.

It is observed in Figure 4 that a right-side split is better for Frozen training configuration since it uses more data to train the feature extractors (more parameters to train in the first phase) and uses limited data to train only the decision layer (fewer parameters to train in the second phase). A left-sided split is better for a Non-Frozen training configuration since it uses more data to train the decision layer and fine-tune the feature extractors (more parameters to train in the second phase). A right-sided split for Non-Frozen training configuration conditions the network to overfit the entire network on limited data during the second phase.

Even though the observations made in this paper are based on one dataset (NTIRE-21), the dataset is representative of various aerial imagery datasets. Such datasets generally consist of an Electro-Optical (EO) or RGB camera sensor modality as the dominant modality which depicts an aerial view as the human eye perceives it. They also consist of other support modalities like Near-Infrared (NIR), Synthetic Aperture Radar (SAR), or hyperspectral imaging sensors which are more useful in cloudy or nighttime conditions when the dominant modalities may be more susceptible to occlusion or poor lighting.

4. CONCLUSIONS

In conclusion, this paper presented a two-phased training process for multimodal networks. The proposed approach was validated using the NTIRE-21 dataset, a 10-class aerial imagery classification problem. The network accuracy and training time was compared with end-to-end trained multimodal network and unimodal networks as baselines. The training time of the multimodal network with the proposed method is significantly less than joint stream end-to-end training of the multimodal network; however, there is a small yet acceptable drop in the performance accuracy. Effects of the presence of noise during inference and different data split on network performance were also studied. The proposed two-phase multimodal network training method provides a way to fuse legacy unimodal networks trained on unpaired data from different modalities into a multimodal network. This also enhances the usability of the legacy unimodal networks while transitioning to the multimodal sensing paradigm and would benefit industries such as satellite surveillance, and autonomous vehicles which use multimodal data.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the Department of Defense under award HM04761912014 and the Air Force Office of Scientific Research (AFOSR) under award FA9550-20-1-0039.

REFERENCES

[1] Dhanaraj, M., Sharma, M., Sarkar, T., Karnam, S., Chachlakis, D. G., Ptucha, R., Markopoulos, P. P., and Saber, E., "Vehicle detection from multi-modal aerial imagery using YOLOv3 with mid-level fusion," in [*Big data II: learning, analytics, and applications*], **11395**, 22–32, SPIE (2020).

[2] Chen, C., Yan, J., Wang, L., Liang, D., and Zhang, W., "Classification of urban functional areas from remote sensing images and time-series user behavior data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 1207–1221 (2020).

[3] Fatima, S. A., Kumar, A., Pratap, A., and Raoof, S. S., "Object recognition and detection in remote sensing images: a comparative study," in [*2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*], 1–5, IEEE (2020).

[4] Sharma, M., Dhanaraj, M., Karnam, S., Chachlakis, D. G., Ptucha, R., Markopoulos, P. P., and Saber, E., "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 1497–1508 (2021).

[5] Sharma, M., Markopoulos, P. P., and Saber, E., "YOLOrs-lite: A lightweight cnn for real-time object detection in remote-sensing," in [*2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*], 2604–2607 (2021).

[6] Ma, L., Li, M., Ma, X., Cheng, L., Du, P., and Liu, Y., "A review of supervised object-based land-cover image classification," *ISPRS Journal of Photogrammetry and Remote Sensing* **130**, 277–293 (2017).

[7] Ahmad, M., Shabbir, S., Roy, S. K., Hong, D., Wu, X., Yao, J., Khan, A. M., Mazzara, M., Distefano, S., and Chanussot, J., "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE journal of selected topics in applied earth observations and remote sensing* **15**, 968–999 (2021).

[8] Li, Y., Zhang, H., Xue, X., Jiang, Y., and Shen, Q., "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(6), e1264 (2018).

[9] Datta, D., Mallick, P. K., Bhoi, A. K., Ijaz, M. F., Shafi, J., and Choi, J., "Hyperspectral image classification: Potentials, challenges, and future directions," *Computational Intelligence and Neuroscience* **2022** (2022).

[10] Yu, J., Chang, H., Lu, K., Zhang, L., and Du, S., "Scene clustering based pseudo-labeling strategy for multi-modal aerial view object classification," *arXiv preprint arXiv:2205.01920* (2022).

[11] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., and Chanussot, J., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation* **112**, 102926 (2022).

[12] Liu, J., Inkawhich, N., Nina, O., and Timofte, R., "NTIRE 2021 multi-modal aerial view object classification challenge," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 588–595 (2021).

[13] Low, S., Nina, O., Sappa, A. D., Blasch, E., and Inkawhich, N., "Multi-modal aerial view object classification challenge results-pbvs 2022," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 350–358 (2022).

[14] Jahan, C. S., Savakis, A., and Blasch, E., "Cross-modal knowledge distillation in deep networks for SAR image classification," in [*Geospatial Informatics XII*], **12099**, 20–27, SPIE (2022).

[15] Blasch, E. and Savakis, A., "Methods of fused EO/SAR deep learning," in [*Automatic Target Recognition XXXII*], PC1209604, SPIE (2022).

[16] Lahat, D., Adali, T., and Jutten, C., "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE* **103**(9), 1449–1477 (2015).

[17] Zhu, B., Zhou, L., Pu, S., Fan, J., and Ye, Y., "Advances and challenges in multimodal remote sensing image registration," *IEEE Journal on Miniaturization for Air and Space Systems* (2023).

[18] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N., "MDETR-modulated detection for end-to-end multi-modal understanding," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 1780–1790 (2021).

[19] Chai, W. and Wang, G., "Deep vision multimodal learning: Methodology, benchmark, and trend," *Applied Sciences* **12**(13), 6588 (2022).

[20] Baltrušaitis, T., Ahuja, C., and Morency, L.-P., "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence* **41**(2), 423–443 (2018).

[21] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE transactions on medical imaging* **35**(5), 1299–1312 (2016).

[22] Hussain, M., Bird, J. J., and Faria, D. R., "A study on cnn transfer learning for image classification," in [*Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*], 191–202, Springer (2019).

[23] Pérez-Pellitero, E., Catley-Chandar, S., Leonardis, A., and Timofte, R., "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 691–700 (2021).

[24] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).