X-Net: Bimodal Feature Representation Learning in Satellite Imagery

Kenneth Tran[*^], Wesam Sakla[*]

Lawrence Livermore National Laboratory[*], NCSU[^]

Abstract/Motivation

In our work, we are interested in data fusion between visible color imagery and multispectral imagery in satellite data. We use a deep learning model that learns a joint representation between the two modalities. By using an autoencoder, we can train the model in an unsupervised fashion and transfer the learned encoder weights to perform downstream classification tasks. As a result, we have a semi-supervised method of learning bimodal data that can use unlabeled data, which is very common for remote sensing images.

We apply our method to the multispectral and visible image data from the novel Function Map of the World (fMoW) data set. This data set came from a competition released by IARPA in September 2017. The top three performing contestants only used the RGB data in their models. We hypothesize that by learning a meaningful representation of both RGB and multispectral data, we can enhance classification performance.

Functional Map of the World (fMoW) Data Set

- 1,047,691 images covering a majority of countries (over 80%)
- Contains 62 named categories and an additional false detection category
- Includes metadata features and statistics such as the ISO Country Code, UTM Zone, and Off-Nadir Angle
- Each data point contains both RGB and multispectral imagery
- Each area of interest is imaged at multiple times, making temporal analysis possible

Examples (Resized from original):

Educational Facility

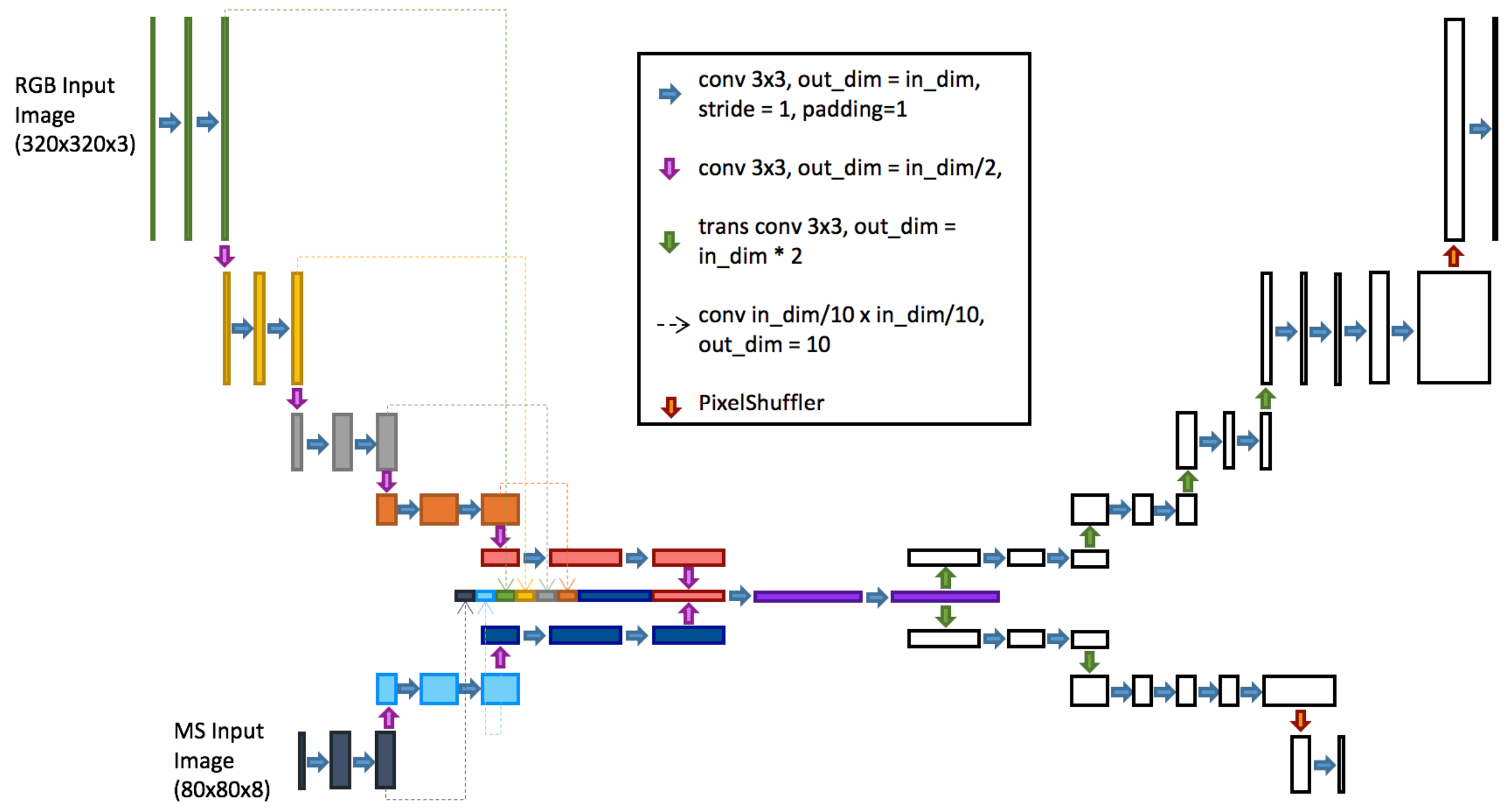
Military Facility

Place of Worship

Ground Transportation Station

Method

We aim to create a meaningful joint representation between the RGB and multispectral imagery. To do so, we draw inspiration from early multimodal autoencoders, along with more recent developments in deep learning, such as feed forward connections from U-Nets and densely connected CNNs. We design an architecture named X-Net which consists of two U-Net like autoencoders from separate modalities that are fused in the final stage of encoding. Since we are interested in the joint representation, we drop feed forward connections from the encoder to decoder and instead feed activations from each stage of the encoder directly to the fusion stage.


RGB Input Image (320x320x3)
MS Input Image (80x80x8)
conv 3x3, out_dim = in_dim, stride = 1, padding=1
conv 3x3, out_dim = in_dim/2,
trans conv 3x3, out_dim = in_dim * 2
conv in_dim/10 x in_dim/10, out_dim = 10
PixelShuffler

X-Net Architecture

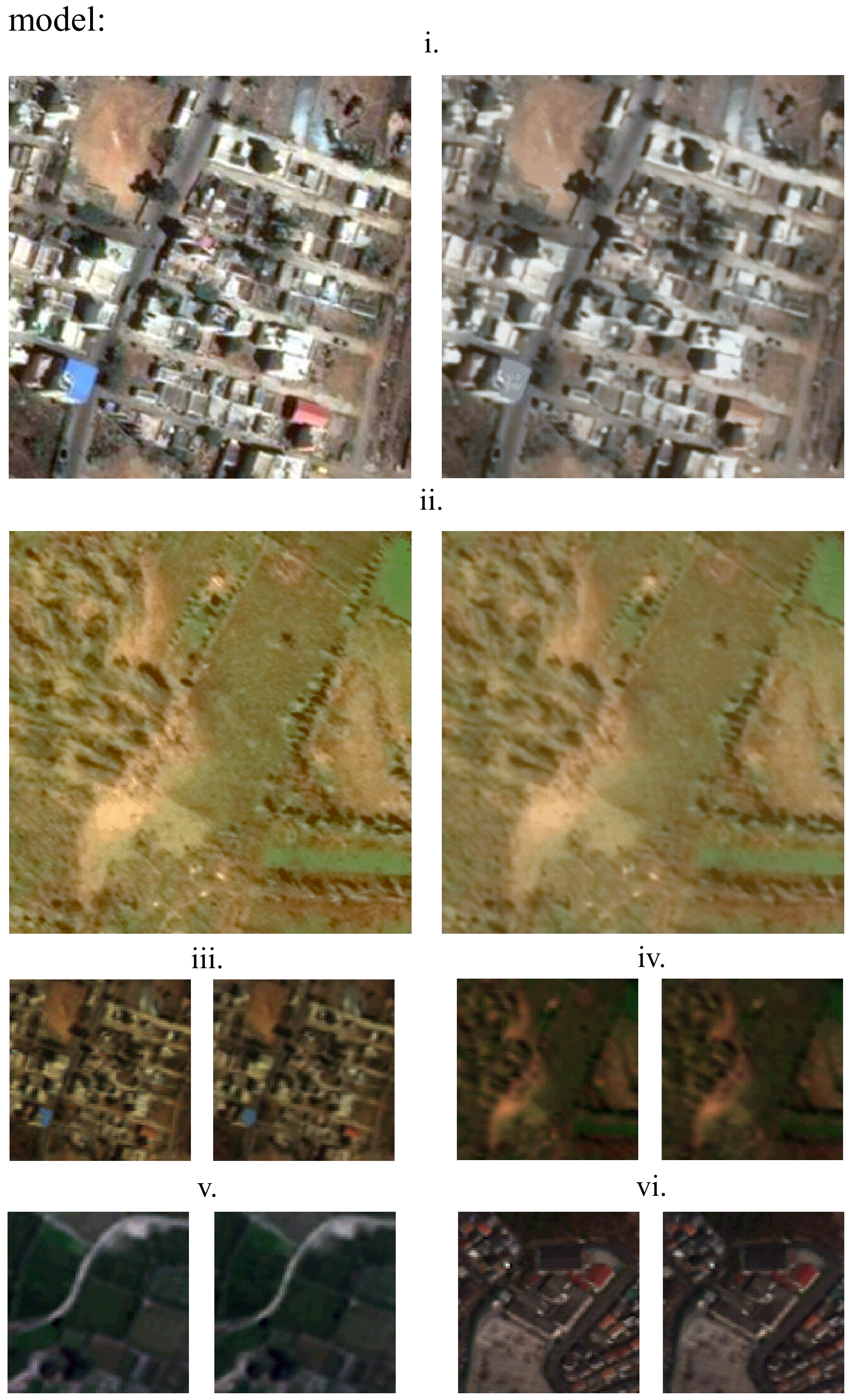
We make several design choices in building our model, including:

- Using a fully convolutional neural network to learn the filter weights for down-sampling convolutions instead of a pooling layer
- Down-sampling the output activations at each individual stage (color coded) and feeding it directly to the input of the fusion layers (purple activations)
- Replacing the last up-sampling layer in the decoders with a pixel shuffler operation, which has shown promising results for constructing images in super resolution applications
- Using a perceptual loss to train our model. This can be thought of as taking the mean squared error between activations of intermediate layers of a pre-trained model (in our case VGG on ImageNet). Mathematically, this can be written as follows:

$$l_{VGG}^{SR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{HR}))_{x,y})^2$$

Results

We use reconstruction error as a measure for how well our model performs. Below, we show some examples of reconstruction from our model:



RGB: i and ii, MS (RGB channels shown): iii-vi

Left: Ground Truth Right: Reconstructed Image

Peak Signal to Noise Ratio			
	Bimodal Autoencoder	Bidirectional DNN	X-Net (our architecture)
RGB	28.086	27.947	30.384
Multispectral	28.799	28.294	33.820

A table comparing the peak signal to noise ratio (PSNR) between the three different neural network architectures

Conclusion

Our contributions to multimodal learning include:

- Enhancing the basic multimodal autoencoder architecture by adopting the U-Net structure for increased learning capability
- Introducing forward propagation from each encoding stage to the fusion layers, which helps retain information from previous convolutions
- Applying tricks from super resolution applications that can re-create more realistic and less blurry images as output of the network

By using unsupervised learning, we can still utilize the abundant amount of unlabeled remote sensing data available to learn the joint representation of visible color and multispectral data. The encoder then acts as a pre-trained feature extractor for supervised tasks such as classification or change detection.

Future Work

Currently, we only evaluate the joint representation using reconstruction performance. In the near future, we will also test other tasks for performance measures, including:

- Classification of the annotated satellite image patches provided by the fMoW data set
- Generating "missing data" in images by using the temporal component of the data set. "Missing" data could be patches of images covered by cloud, shadows, etc.

We would like to enrich our model to be more general for other data sets as well as improve its current performance by:

- Using a model pre-trained on remote sensing data instead of ImageNet for better perceptual loss
- Modify the architecture to accommodate more modalities

References

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.

Christie, G. et al (in press), "Functional Map of the World," Proc. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, (2018).

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In International Conference on Machine Learning (ICML), Bellevue, USA, June 2011.

O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervention*, pp. 234-241, 2015.

V. Vukotić et al. Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications. In *Proceedings of the 2016 ACM International Conference on Multimedia Retrieval (ICMR)*, pages 343–346. ACM, New York, 2016.

IM-POST-XXXXX
tid.llnl.gov/print

Views expressed here do not necessarily reflect the opinion of the United States Government, the United States Department of Energy, the Remote Sensing Laboratory, or the Lawrence Livermore National Laboratory. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52 07NA27344. Lawrence Livermore National Security, LLC