SANDIA REPORT
SAND2022-12264
Printed September 2022

Sandia National Laboratories

Data Fusion via Neural Network Entropy Minimization for Target Detection and Multi-Sensor Event Classification

Dylan Anderson, Jorge Garcia, Lisa Linville, Joshua Michalenko

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods/


DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
NNSA
National Nuclear Security Administration

ABSTRACT

Broadly applicable solutions to multimodal and multisensory fusion problems across domains remain a challenge because effective solutions often require substantive domain knowledge and engineering. The chief questions that arise for data fusion are in when to share information from different data sources, and how to accomplish the integration of information. The solutions explored in this work remain agnostic to input representation and terminal decision fusion approaches by sharing information through the learning objective as a compound objective function. The objective function this work uses assumes a one-to-one learning paradigm within a one-to-many domain which allows the assumption that consistency can be enforced across the one-to-many dimension. The domains and tasks we explore in this work include multi-sensor fusion for seismic event location and multimodal hyperspectral target discrimination. We find that our domain-informed consistency objectives are challenging to implement in stable and successful learning because of intersections between inherent data complexity and practical parameter optimization. While multimodal hyperspectral target discrimination was not enhanced across a range of different experiments by the fusion strategies put forward in this work, seismic event location benefited substantially, but only for label-limited scenarios.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page left blank

ACRONYMS AND TERMS

Acronym/Term	Definition
CNN	Convolutional Neural Network
DNN R&D	NNSA Defense Nuclear Nonproliferation Research and Development Office
EMIN	Entropy Minimization loss objective
HSI	Hyperspectral Imagery
RGB	Red-Green-Blue
SWIR	Short Wave Infrared
UNESE	Underground Nuclear Explosion Signatures Experiment
VNIR	Visible through Near Infrared

1. INTRODUCTION

Data fusion remains a persistent challenge for many domains because the optimal exploitation of data across scales, resolutions, and phenomenologies depends uniquely on data characteristics. Ideal data fusion strategies maximize the available information across available data modalities and effectively prioritize it for informed decision making. Data sharing early in the modelling process (such as feature fusion and parameter sharing) reduces downstream complexity because the effective combination of disparate data sources is calibrated through the modeling process. Alternatively, independent modeling can solve limitations that early data sharing experiences at the expense of more nuanced and challenging decision fusion later. Both approaches and the myriad variations of each that have evolved can be highly effective for specific use cases. Despite the abundance of ideas and the rapid pace of research in data fusion, the most effective solutions on one problem or domain rarely provide similar gains broadly. The goal of this work was to explore a model training paradigm that could be applied broadly across domains by remaining agnostic to input representation and output decision-making. Conceptually, our goal was to maximize the predictive competence for situations, samples, and phenomena where multiple examples were available and when shared context would result in more robust and informative representations and task performance.

The training method we explored reduces the disagreement (or entropy) between model outputs on a given sample when multiple data sources, channels, or modalities exist for that sample. We call the objective entropy minimization or EMIN henceforth and apply it to problems in two domains. The first domain is a set of Hyperspectral Imagery (HSI) collections over a specific geographic location. The data collection in this region is characteristic of temporal, geographic, and dynamic ambient conditions that make data fusion challenging. The second domain we explore is the challenging task of single-station event location from different seismic sensor types. In the first domain EMIN is applied directly with labeled loss in a multiclass prediction task. In the second domain EMIN is applied to a transform of the multitask regression targets. Our experiments inform how modelling paradigms, multiple objective weighting, and inherent data complexity and quality all influence

the practical value of EMIN on data fusion problems that resemble the scenarios explored here.

2. ENTROPY MINIMIZATION

The goal of multi-modal decision fusion is to make better overall predictions by utilizing multiple observations of the same event/phenomena to corroborate evidence towards a single prediction. One popular method of achieving this is to have independent models that have been optimized to give high performance for each modality. For a given event, all model predictions can then be fused together to form a better estimate. One drawback of this method is that without an additional calibration scheme, fusion of the individual model predictions is an additional issue that must be solved; a 'high' estimate from one model may be within a different range of another model's 'high' estimate, resulting in a non-trivial fusion problem which has been addressed in the literature (Anderson et al., 2007; Simonson, 1998). Alternatively, fusion could be a much easier task if individual models tended to agree with each other when appropriate (e.g., be similarly calibrated). We hypothesized that encouraging the agreement of models when individual modalities observe the same phenomena as an objective during the learning process would improve fused predictive performance.

2.1. EMIN formulation

Expanding on previous work (Michalenko et al., 2020), we matured the concept of EMIN by formulating a generic loss function and variants based on the principle that multi-modality outputs should typically agree with one other. Our approach assumes that multi-modality outputs are independent identically distributed draws from a gaussian distribution with known mean and unknown variance. We then used standard parameter estimation techniques to estimate the unknown parameters and compute distributional entropy as a proxy variable of agreement between models. The resulting proxy variable is then used as a regularization loss term to a supervised loss while training gradient-based models such as neural networks.

More concretely, let $\mathrm{A} = \alpha_0,\alpha_1,\ldots,\alpha_M$ with $\alpha_m \in \mathbb{R}^D$ and let $D = 1$ (for simplicity, but extended in section 2.2) be the scalar outputs from a set of M multi-modal models. This is the case for a regression function f where the output of the $\mathrm{m_{th}}$ model is scalar value $\alpha_m = f(x_m \mid \Theta_m)$ where x_m and Θ_m are the inputs and parameters of the m_{th} model respectively. We want to place assumptions over A such that the M

models must adhere to what we believe will result in better overall model performance. If A is a random variable (RV) and αs are realizations of A, we can compute statistics of A that are of interest. To make modeling easier, we also say the αs are independent draws of the same distribution, which can make an initial formulation more tractable. We note that α_m has a deterministic relation to x_m and although it is likely that x_* are not independent, we will assume this for simplicity.

We desire that A has low entropy $\mathcal{H}(\mathrm{A})$, or 'surprise'. If draws of A are similar and typically on the same order as each other, decision fusion becomes easier because separate models will output similar values for different x_* inputs. Calculating entropy typically requires knowledge of the full distribution of A. For a well-behaved parametrized distribution such as $\mathrm{A} \sim \mathcal{N}(\mu = \mu_o, \sigma^2)$ with known μ_o and unknown σ we can estimate the unknown σ using standard parameter estimation techniques like maximum likelihood and subsequently compute $\mathcal{H}(\mathrm{A})$.

One of the easiest estimation techniques is to evaluate the likelihood of $\mathcal{L}(\sigma^2|A) = p\left(\mathrm{A} = \{\alpha_0, \alpha_1, \ldots . \alpha_M\}\right) \mid \sigma^2, \mu = \mu_o)$ and find the value $\widehat{\sigma^2}$ that maximizes it. Since $\mathrm{A} \sim \mathcal{N}(\mu = \mu_o, \sigma^2)$ we can write out the likelihood function, take the gradient w.r.t. σ^2, set $\nabla_{\sigma^2}\left(p\left(\mathrm{A} \mid \sigma^2, \mu = \mu_o\right)\right) = 0$ and solve for σ^2. The solution, $\widehat{\sigma^2} = \frac{1}{\mathrm{n}}\sum(\alpha_i - \bar{\alpha})^2$ is the Maximum Likelihood Estimate (MLE) of σ^2 where $\bar{\alpha}$ is the sample mean.

Once we have $\widehat{\sigma^2}$, we know $\mathrm{A} \sim \mathcal{N}\left(\mu = \mu_o, \sigma^2 = \widehat{\sigma^2}\right)$ and can now calculate $\mathcal{H}(\mathrm{A}) = \frac{1}{2}\log\left(2\pi\widehat{\sigma^2}\right) + \frac{1}{2}$. In this case, we can see that treating entropy as a loss term in the minimization problem $\min \mathcal{H}(\mathrm{A})$ means minimizing $\widehat{\sigma^2} = \frac{1}{n}\sum(\alpha_i - \bar{\alpha})^2$since the log is monotonic in $\widehat{\sigma^2}$. Therefore, under the Gaussian assumption, minimizing entropy is the same as minimizing variance across different model outputs. This is shown in Figure 1 for a decision fusion problem with 2 separate models. Each plot represents the same model and data trained 200 epochs on a toy decision fusion problem with varying levels of EMIN enforced upon the models. Each point in the scatter plot represents two model predictions over the same event for two inputs. The left plot represents no EMIN applied to the model, the middle plot represents a moderate penalty, and the right plot represents a high EMIN penalty applied to the model. As

the EMIN penalty is increased, individual model outputs become more correlated which is the expected effect when minimizing entropy. Simple use case examples and EMIN objectives as formulated and used in this work can be found at https://innersource.sandia.gov/portal.

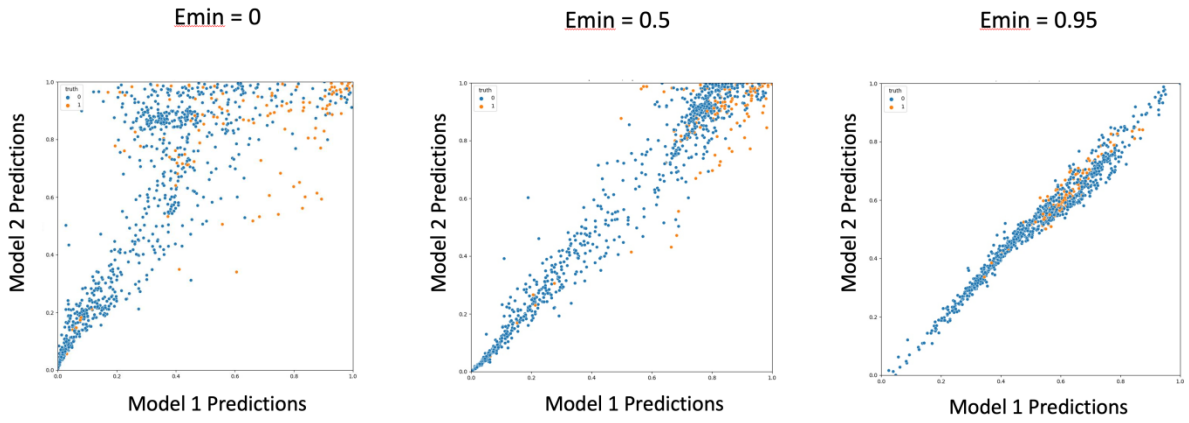

Figure 1. Effects of minimizing entropy across two models for 3 different penalty values.

2.2. Multi-Attribute derivative EMIN

The $D = 1$ formulation can be extended to the multidimensional case. A straightforward formulation leveraging uncorrelated multivariate assumptions treats $A \sim \mathcal{N}(\mu = \underline{\mu_o}, \underline{I\sigma^2})$ where parameter estimation of each dimension can be performed independently. This formulation is feasible when M is small as it keeps the number of parameters to estimate less than the number of observations. This formulation was used in most of the seismic and HSI experiments described in following sections.

In the UNESE HSI experiments, up to 6 models estimate a classification value in a 10-class prediction task. In the seismic event location experiments, 2 models estimate 2 regression targets each. EMIN is applied to a differentiable transform of the model predictions where $D = 2$. The transform conveys directional and distance attributes to positional attributes, specifically latitude and longitude (in radians), which EMIN then is applied to (see Equations 2-3 in Section 4.5).

2.3. Student-Teacher EMIN

In practice, we may encounter scenarios in which one or a subset of modalities are much higher performing than the complement set. This may be the case in which one sensing modality is able to capture more event information than other modalities at our disposal. In such settings the practical effect of EMIN is an increase in performance for the poor performing modality at the cost of decreasing performance for the higher performing modalities. A student-teacher framework can be used to address this setting, where the teacher model(s) predictions $\alpha_{teacher}$ are used in the estimate of $\bar{\alpha}\ in\ \widehat{\sigma^2_{\mathrm{MLE}}} = \frac{1}{n}\sum(\alpha_i - \bar{\alpha})^2$ but are excluded from the set α_i. With this method, EMIN gradients with respect to model parameters from the teacher models are zero, while EMIN gradients with respect to student model parameters are still non-zero. The effect is that teacher models are unaffected by the EMIN training process and student models are pulled in the direction of teacher model outputs with the assumption teacher models are higher performing. Student-teacher learning experiments are explored for both HSI and seismic cases.

3. UNESE HYPERSPECTRAL IMAGERY MULTISENSOR TARGET CLASSIFICATION

3.1. Introduction

This section describes experiments for multi-sensor target classification leveraging the UNESE U12 Hyperspectral Imagery (HSI) collections focused around the P-tunnel apron and mesa above the Disko Elm test in area 12 of the Nevada National Security Site (UNESE, 2018). These data were collected as part of the Underground Nuclear Explosion Signatures Experiment (UNESE) multi-lab venture funded by DNN R&D. The UNESE data used in the EMIN experiments described below consists of six total modalities: three visible-through near infrared (VNIR) and three short wave infrared (SWIR) pseudo-reflectance hyperspectral images. Each image was collected at a different time and has a unique spatial footprint that does not completely overlap with the five other collections. This chapter describes curation of a multi-modality, multi-class classification dataset from this source data and a sequence of EMIN fusion experiments conducted against this dataset.

There were five experiments for the UNESE classification dataset. The first experiment (Section 3.3) executed against the UNESE dataset consisted of treating each of the six HSI flights (3 VNIR, 3 SWIR) as independent modalities. For each of the six modalities, an independent feedforward neural network was trained to make optimal predictions over that modality. This experiment represented a direct application of the EMIN methodology to an independent multi-sensor scenario for multi-class classification. The second experiment (Section 3.4) developed a single, cross-modality model for all six modalities to decrease the hyperparameter tuning difficulty. The third experiment (Section 3.5) tested how well EMIN performs in different label scarce scenarios. The fourth experiment (Section 3.6) explored how EMIN performed as a fine-tuning step in model training. The final fifth experiment (3.7) tested how a student-teacher version of the EMIN objective performed.

3.2. Dataset curation

3.2.1. Data Labeling

The UNESE HSI data collection was unlabeled, providing only spatial-spectral measurements without annotations. The scene was manually annotated with a series of polygons indicating known or discernable materials of interest (asphalt road, aluminum, stemming sand, etc.). In total, 29 individual targets across 206 distinct polygons were manually annotated. All annotations were performed against the RGB orthomosaic, which provided the greatest spatial accuracy of all collected imagery products. The bulk of annotations were performed around the P-tunnel and T-tunnel aprons, which featured several anthropogenic artifacts and structures (see Figure 2). Photos collected from on the ground as part of the data collection campaign were referenced in manually identifying annotated polygons.

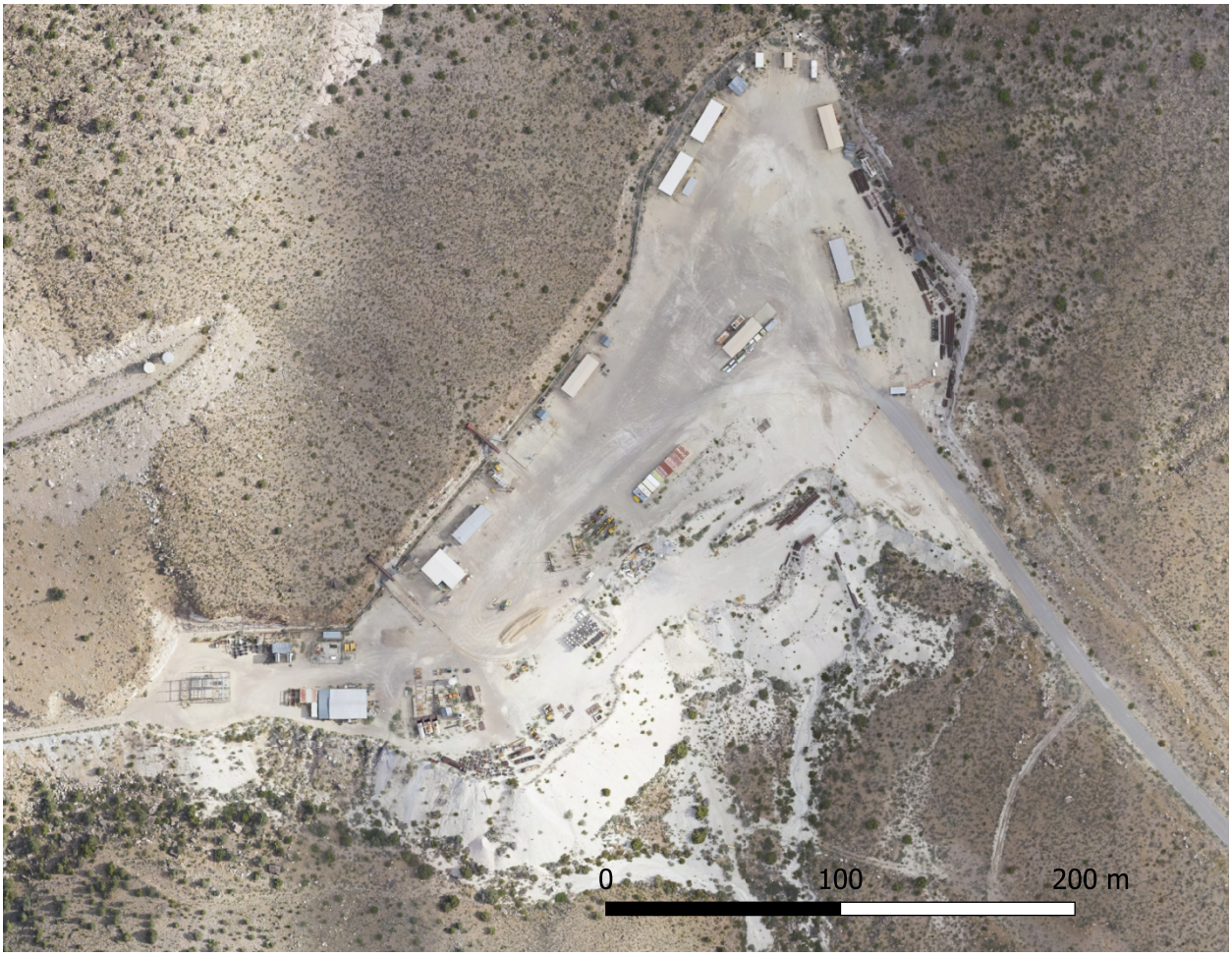

Figure 2. Example manually annotated polygons around the P-tunnel apron.

3.2.2. Spatial Point Sampling

The labeled polygons and original hyperspectral source rasters were conflated to generate a much smaller and more digestible spectral classification dataset. First, the 10 most populous class types were selected for sampling. The polygons corresponding to these 10 class types were then used to generate random points within the polygons. Target points were generated by sampling a point density of 5 pts / m^2 within target polygons and at a density of 0.025 pts / m^2 outside polygons to generate "background" points. Each target class was then downsampled to 2,000 pts/target. Using these randomly generated points, the spectra were sampled at each point for the 6 hyperspectral raster images. Many points did not overlap with specific hyperspectral images and these "empty" spectra were dropped from the analysis, resulting in a non-uniform distribution of target classes per raster. See Table 1 for statistics on the sampled dataset, resolved by raster and class.

Table 1. UNESE multiclass sample statistics.

Target	VNIR1	VNIR2	VNIR3	SWIR1	SWIR2	SWIR3
Asphalt road	66	790	1531	27	1519	1515
Background	747	1206	514	504	1488	694
Concrete	1506	503	9	1401	1635	8
Graded Dirt Road	1220	293	55	598	1318	139
Pad Muck	580	2000	232	507	580	0
Rusted Metal	333	1073	916	518	416	72
Sheet Metal	114	1886	873	568	114	0
Tailings	91	1485	0	956	1237	1485

Target	VNIR1	VNIR2	VNIR3	SWIR1	SWIR2	SWIR3
Stemming Gravel	0	411	0	63	0	0
Stemming Sand	0	697	0	595	0	0
Total	4657	10344	4130	5737	8307	3913
% Missing	70.9	35.4	74.2	64.1	48.1	75.5

3.2.3. Spectral Preprocessing

Two preprocessing steps were used to condition the data for classification. The data were notionally corrected to reflectance, meaning theoretically that it should take on values between [0, 1]. However, necessary assumptions made during the reflectance calibration process mean this was not the case. The differences in collection times-of-day as well as the reflectance calibration methodology used suggest that there are meaningful and systematic differences, even between collections made with the same sensors. The preprocessing steps employed were as follows:

1. Mask “bad” bands, with poor sensor responsivity, low atmosphere transmittance (see Figure 3), etc. Specifically:
    a. Limit VNIR data to 400 – 900 nm
    b. Limit SWIR data to 900 – 2500 nm, drop bands with atmospheric transmission < 0.25, drop bands for 1900 nm and 2050 nm.
2. Clip data to a range of [0, 1.5]. Clipping the data mitigates the effects of outliers from poor calibration but retains the overall shapes and reflectance profiles. This is desirable over standard normalization since the data is already primarily within the requisite range.

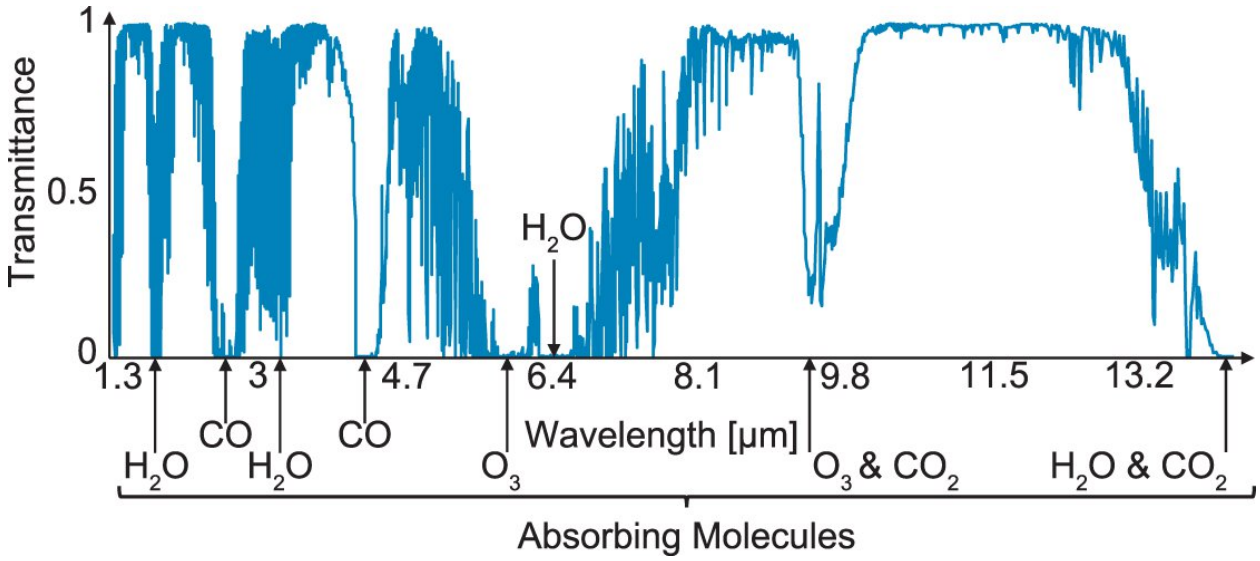

Figure 3. Sample atmospheric transmission profile.

3.2.4. Partitioning and Known Data Issues

Data were split into stratified (by class) 5-fold cross validation, with folds 1-4 used for leave-one-set-out cross validation and fold 5 withheld for test evaluation after training and model selection.

The resulting processed multi-class classification dataset consists of several known issues. The first and biggest issue is that of spatial alignment, which can degrade the correctness of labels. Each underlying hyperspectral image has variable spatial alignment errors to the labeled polygons. For example, see Figure 4, which shows how the same “concrete” label polygon sits atop the original RGB, a VNIR image, and the SWIR image. Steps were taken to partially mitigate this by applying an interior buffer, preventing sampling of points along polygon edges.

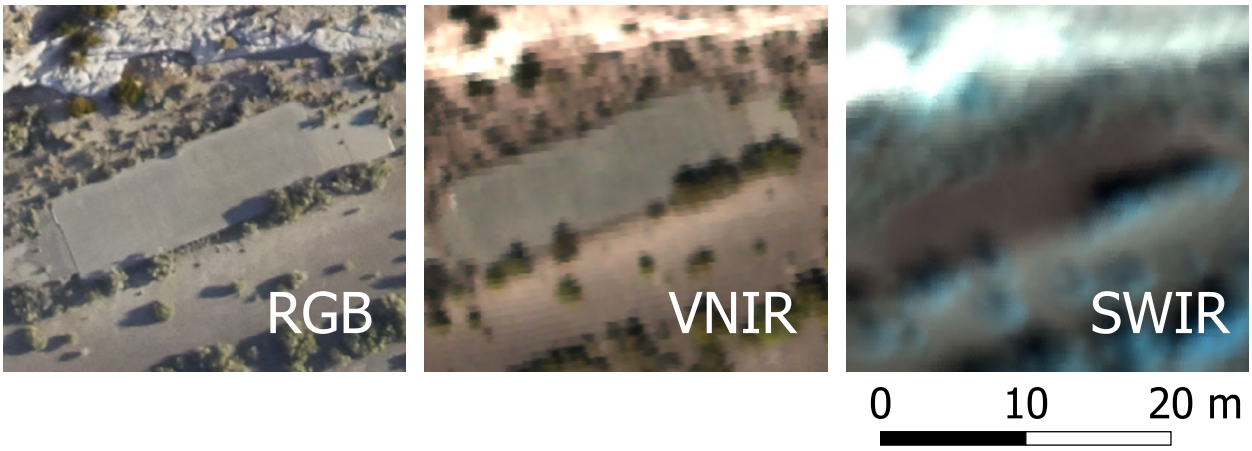

Figure 4. Variable spatial alignment of annotated target polygon with VNIR and SWIR data modalities.

The second known issue is that of data calibration and shadows. The data reflectance calibration is poor, and the data does not align well to known spectral libraries. These issues are further compounded by the exceptionally steep topography in the region, which induces angular bi-directional reflectance distribution function (BRDF) differences into measured spectra, as well as deep shadows dependent on collection time.

As described previously, the collection footprints for each of the 6 hyperspectral images vary widely, resulting in a sparse set of modalities per spectral point used in the dataset. The histogram of the number of modalities present in each data sample is shown in Figure 5; there are no samples observed by more than 4 modalities. The combinations of modalities present in samples is highly biased, as it is based on the spatial overlap of collection footprints. Thus, not all combinations of all modalities are observed in the data.

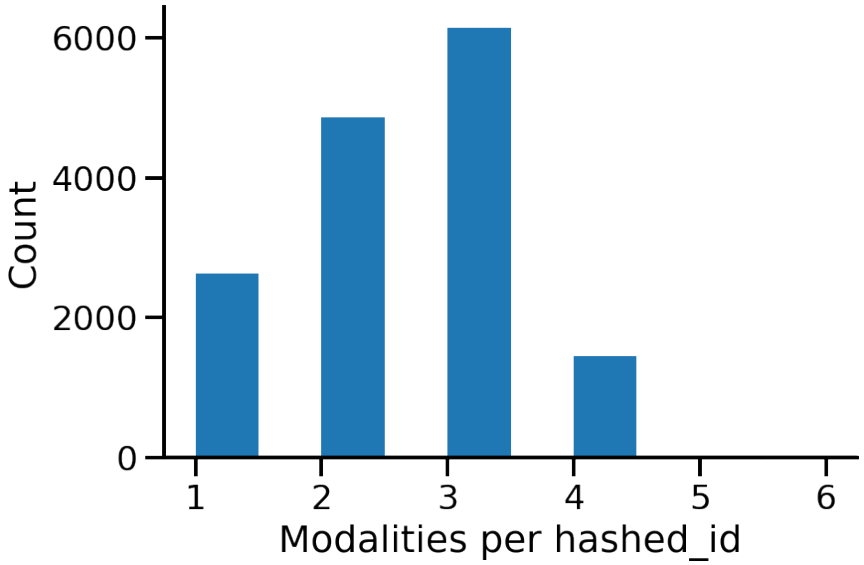

Figure 5. Distribution of the number of modalities available per sample.

3.3. UNESE experiment 1: EMIN multitarget classification

A baseline fusion approach was established by averaging softmax-normalized predictions from six independently trained models. Each model was selected from a fully connected network architecture, which consisted of 2-4 layers with 10-120 units per layer. Dropout was applied to each layer (dropout probability was sampled as a hyperparameter) and ReLU non-linear activations and batch normalization were applied between each layer. A summary of the hyperparameter distributions is shown in Table 2. Runs were executed using early stopping with a patience of 500 epochs applied to validation macro accuracy. Two hundred and fifty runs were sampled for each modality.

Table 2. Baseline hyperparameter ranges and distributions.

Parameter	Sampling Distribution
Number of layers	Uniform integers on [2, 4]
Layer sizes	Uniform integers on [10, 120]
Dropout probability	Uniform on [0, 0.35]
Batch size	Uniform integers quantized to 50 on [50, 1000]
Learning rate	Log-uniform on [1e-5, 1e-2]
Weight decay	Log-uniform on [1e-8, 1e-1]

Analysis of the validation accuracy during hyperparameter sampling revealed that dropout and weight decay both negatively impacted performance. Both were disabled for a second round of 250 runs/modality hyperparameter tuning that was used for final fused results. The best run per modality was selected using validation macro accuracy performance (across 4 replicate cross-validation iterations). Hyperparameter configurations for each best selected run were then re-run 5 times with different seeds. This resulted in 6 modalities x 5 replicate runs / modality to define the baseline. Cross-modality runs were randomly selected to fuse into 5 distinct baselines. For a given baseline set of models, post-softmax predictions were averaged by sample ID (for the modalities that observed a particular sample) and argmax was used over the averaged softmax scores for class predictions.

Entropy minimization was applied very similarly as in the baseline, with the core difference being that all six modalities were trained simultaneously and linked via the Gaussian Entropy Minimization method. These EMIN based runs used the same range of model architecture parameters as were used for the baseline. Dropout and weight decay were disabled based upon baseline results, and batch size, learning rate, and network sizes were sampled from the same distributions as in the baseline.

The EMIN runs balanced the EMIN and cross entropy loss terms using an α parameter as follows:

$$L = \alpha L_{ce} + (1 - \alpha) L_{emin} \qquad \textit{Equation 1}$$

Exploratory experiments highlighted the importance of carefully selecting this α: too large and the models were identical to the baselines (EMIN had no effect) and too small and the models experienced mode collapse by optimizing only for entropy minimization and not classification performance. To address this, two different selection strategies were implemented and subsequently sampled over. First, a

simple sampling strategy employed by simply randomly sampling α. Second, a strategy for Pareto front optimization was used to dynamically adjust α during each training step using the Multiple Gradient Descent Algorithm (MGDA) (Equation 4 for two-term multi-objective optimization (Sener & Koltun, 2018)). The use of gradient normalization in the MGDA formulation was additionally sampled with a 25% probability. The decision to use the random sampling strategy or MGDA was sampled over during hyperparameter tuning with equal weight for each.

The MGDA-based Pareto front optimization provided an elegant solution to the selection of α by selecting a value and taking optimization steps that are guaranteed to improve both the cross entropy and entropy minimization losses. However, this selection strategy is predicated on the idea that optimizing cross entropy and entropy minimization are *equally important*. The overarching goal of EMIN based regularization was improved prediction performance in fusion-based scenarios: it does not follow that it was appropriate to equally weight these loss terms to achieve this goal. Indeed, many of the training loss curves from MGDA showed a very minimally decreasing cross entropy loss to simultaneously decrease cross-modality entropy. These runs often resulted in inferior final performance to models trained by simply randomly sampling a fixed value of α.

The best runs for both baseline and EMIN were selected (per modality in the case of baseline) based upon validation set performance. Test set macro error rates from selected models are shown in Figure 6. These results showed no statistically significant difference in overall performance between baseline and EMIN and showed a degradation in some individual modalities from EMIN. However, the overall performance from the baseline was quite good and may reflect the Bayes error rate due to data issues inherently present in the data (see Section 3.2.4). This experiment appeared to provide little to no headroom for EMIN to improve fusion performance over the baseline.

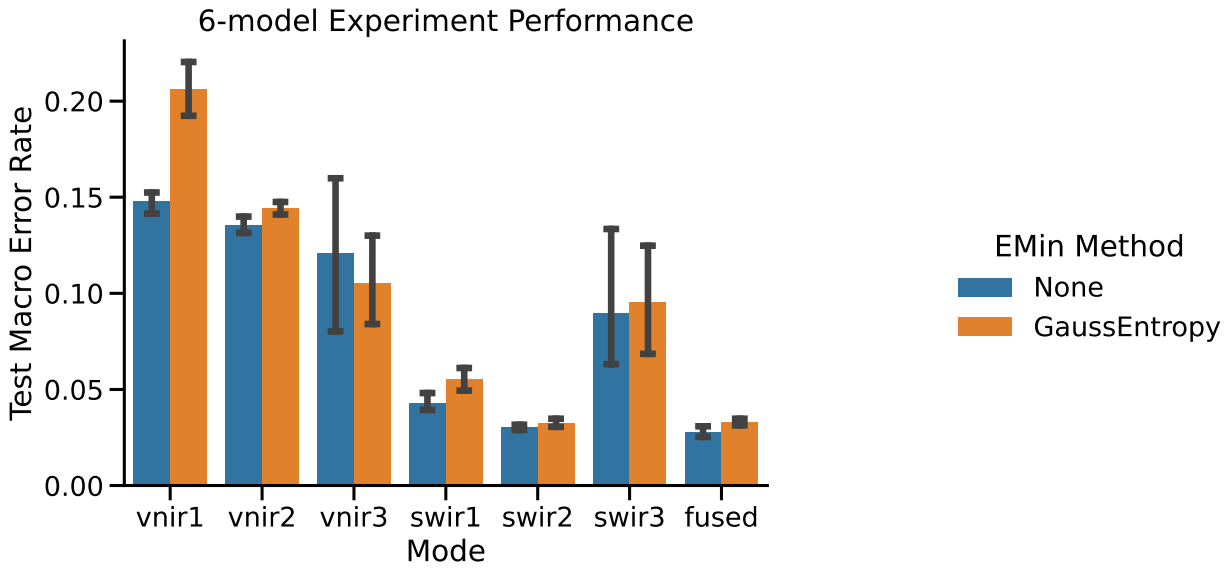

Figure 6. Performance and results from UNESE experiment 1.

3.4. UNESE experiment 2: Hydra Models

Throughout the hyperparameter tuning of UNESE experiment 1 it became apparent that training all six modalities at the same time greatly increased the tuning difficulty. Hyperparameter search expands combinatorically in the number of parameters and linking the parameters across modalities vastly expanded the search. In total, 1000 unique hyperparameter runs were executed for EMIN. To address this issue, a new multi-headed feed forward architecture was designed such that it could accept input from any of the six spectral modalities while also allow for systematic differences between the modalities. The resultant architecture, named the “Hydra model”, employed a modality-specific input (only the requisite input layer was used for any given spectra), with a common feedforward backbone to all input layers. This architecture, illustrated in Figure 7, allowed for modality specific learning in the input layer but aggregate learning in the backbone. Multi-modality predictions (outputs) were averaged into a single softmax normalized output for final class predictions.

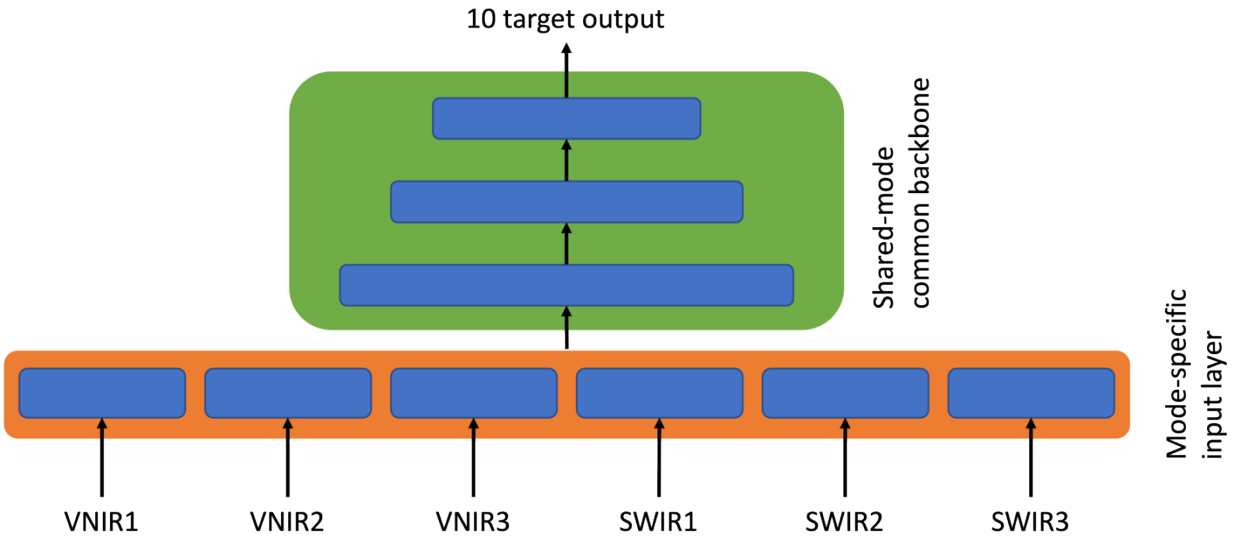

Figure 7. Hydra model architecture.

The Hydra model was hyperparameter tuned over the same parameter ranges used in UNESE experiment 1. Each baseline and EMIN regularized run was defined by a single hydra model rather than 6 individual models as before. MGDA was not employed for EMIN α selection as it did not produce useful results in UNESE experiment 1. Hyperparameter tuning of the singular Hydra baseline and EMIN based models yielded the test macro error rate shown in Figure 8. The EMIN regularized model did not improve results over the baseline. The overall hydra performance was commensurate with observed performance during UNESE experiment 1, whilst proving far easier to conduct effective hyperparameter tuning.

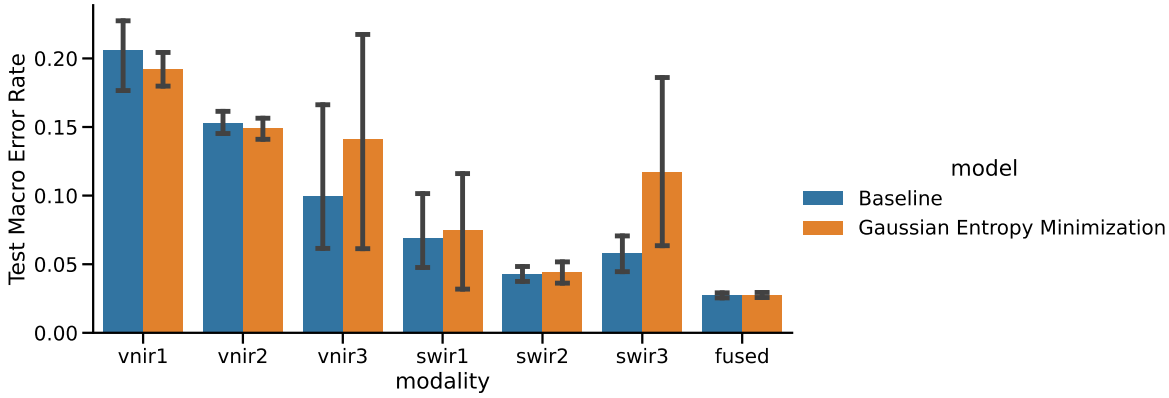

Figure 8. UNESE Hydra Baseline and Gaussian Entropy Minimization results.

Given the observed difficulties in the selection of the α during UNESE experiment 1, the learning dynamics of the best performing hyperparameter settings were observed over a grid of α values (see Figure 9). By examining the evolution of model predictions for specific samples dynamically during the training process, it was observed that for a large range of α values EMIN seemed to have little to no effect on training, had a very fine range in which it changed the dynamics, and then quickly fell off into a degenerate condition that prevented any model learning whatsoever from cross entropy (the supervised classification loss), which strongly regularized models to the initial predictions. From these observations, it seemed possible that the resultant selected best hyperparameters from UNESE experiment 2 had selected a mode for which the value of α yielded no difference from EMIN regularization, hence the equivalent performance with baselines.

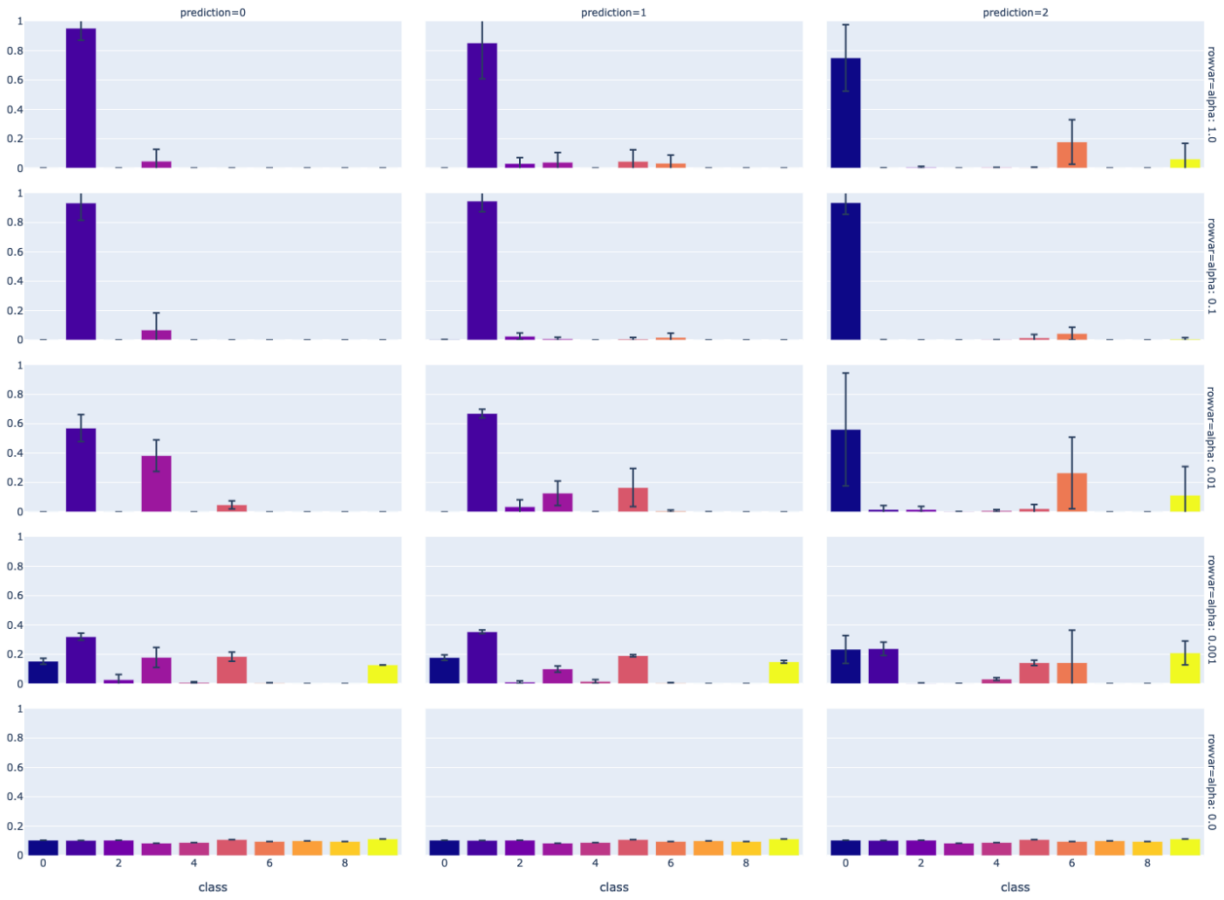

Figure 9. Learning dynamics for variable α after 2,000 training epochs.

3.5. UNESE experiment 3: Label limited EMIN

During UNESE experiments 1 and 2 the developed baselines were so performant that there was minimal headroom for EMIN regularization to improve results. To make the evaluation task more difficult and better highlight differences between approaches, training data was subsampled to fractions of the available training data. The rest of the training data was either withheld from training or was provided as unlabeled data for semi-supervised learning.

Hyperparameter tuning was conducted over the same parameter ranges as in UNESE experiment 2 for a training data fraction of 0.1. The best selected hyperparameter settings were re-trained against a training data fraction of 1.0 and compared to the best results from UNESE experiments 1 and 2. They were found to be slightly better than UNESE experiment 1, which suggested that the best hyperparameters had minimal sensitivity to the fraction of training data and did not need to be retuned for varying fraction. The selected parameters were then used to retrain models over a grid of training data fractions.

The label attrition experiments (Figure 10) highlighted expected behavior in that final model performance was directly correlated with the amount of available training data. In particular, the low training data fractions showed baseline error rates increases to as much as 30% (from a full training data error rate of 3-4%). However, EMIN was not shown to provide a performance increasing regularization benefit (GaussEntropy in Figure 10) or a strong semi-supervised learning mechanism (GaussEntropySSL in Figure 10) even as baseline performance degraded. Label attrition clearly made the resulting problem more challenging, but the baseline performance still represented a strong showing of the possible-to-achieve resultant model error rate.

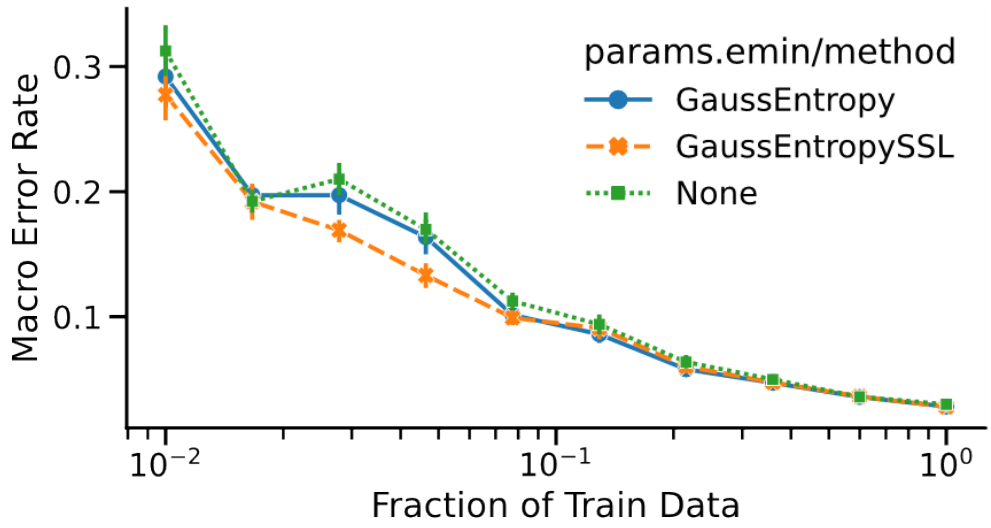

Figure 10. UNESE label attrition results.

3.6. UNESE experiment 4: EMIN as model finetuning

From observations on the difficulty in finding the narrow α band, a new, less sensitive approach for the application of EMIN was designed. In this approach, a model is first trained for a time using $\alpha = 1$ (no EMIN). After reaching a predefined epoch, the parameter is changed to $\alpha = 0$ (no cross entropy, only EMIN) and trained for an additional period. In this manner, EMIN is applied as "model finetuning" to a traditionally trained, supervised model. By training as finetuning, EMIN benefits can be realized on an already well-tuned baseline model.

The best hyperparameter settings found in UNESE experiment 3 were used to explore the EMIN finetuning approach. The model was trained as supervised only for 1,000 epochs, yielding a resultant model that had performance commensurate with a well-tuned baseline. Next, this model was fine-tuned as EMIN-only for an additional 1,000 epochs. The observed error rates after initial training and EMIN finetuning are shown in Figure 11. Finetuning degraded the overall performance.

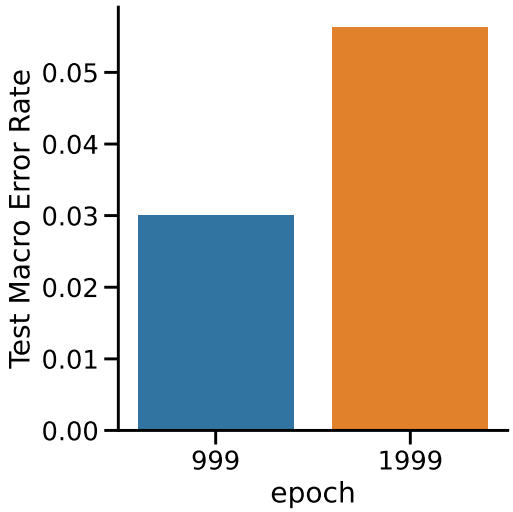

Figure 11. Overall test performance for pre- and post-EMIN finetuning.

The distribution of cross-modality predictive variance was compared between initial training and EMIN finetuning. As expected, the EMIN finetuning process reduced the variance of the highest-variance samples to be more in line with the rest of the dataset. This indicates that EMIN finetuning promoted the expected variance-reducing behavior. However, while the predictive variance analysis illustrated that the EMIN regularization provided the expected mathematical behavior, this regularization did not yield an improvement in overall predictive performance. Comparison of performance per modality (Figure 12) showed that while EMIN finetuning did not change the overall fusion performance, it made the individually worst performing modes better and the best performing modes worse.

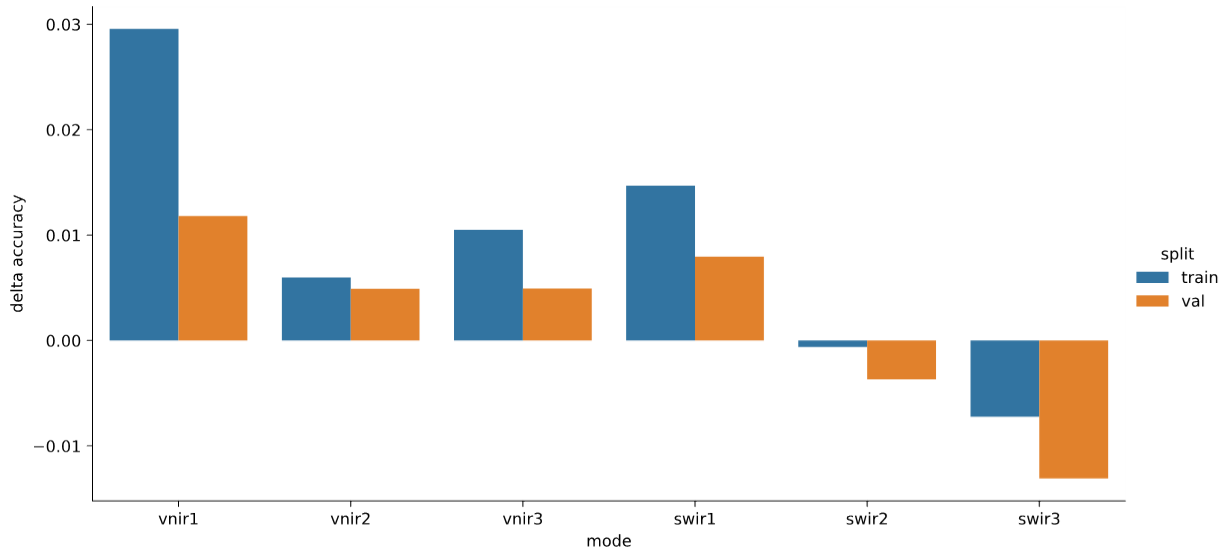

Figure 12. Per modality delta accuracy resulting from EMIN finetune.

3.7. UNESE experiment 5: Student-Teacher Experiments

Throughout the course of UNESE experiments, EMIN-based regularization was continually observed to have statistically insignificant impact on final model performance. From the model finetuning experiments, it was observed that while overall performance was not improved, EMIN provided improvements to the worst performing modalities. Based on these observations, we hypothesized that an alternative, student-teacher EMIN regularization (see Section 2.3) may yield improvement.

In this framework, the modalities were designated *a priori* as being either “student” or “teacher” modalities. In the computation of the EMIN loss, statistics (mean/variance) were computed using the set of teacher modalities with gradient blocking. The EMIN loss using the teacher statistics was then applied over the student modalities. We hypothesized that this would allow the teacher modalities to “teach” the student modalities, thereby applying EMIN regularization to the subset of modalities for which there may be improvement and leaving the others unimpacted by EMIN.

In the application of the student-teacher paradigm to the UNESE dataset, the three SWIR modalities were designated as the teachers and the three VNIR modalities were designated as the students. This was based upon observations that the SWIR

modalities individually provided better performance over the VNIR (Figure 6, Figure 8), and that the VNIR modalities could be improved by EMIN regularization (Figure 12). The student-teacher framework was re-applied in the semi-supervised learning label attrition experiment described in UNESE experiment 3, with a critical difference: both EMIN and baseline models were trained for many more epochs and early stopping of training based on validation performance was disabled.

The results from student-teacher EMIN regularization are shown in Figure 13. The rationale for training longer with student-teacher than in previous experiments was based upon manual examination of learning dynamics of the student-teacher trained model, which showed oscillatory behavior in validation loss that limited the effectiveness of early stopping. Initially, student-teacher based EMIN appeared to show a small but consistent performance improvement over the baseline (denoted as “Early Stopping” in the figure). However, while the baseline did not obviously exhibit this oscillatory behavior, a long-trained baseline (labeled as “Long Baseline” in the figure) yielded the same performance as the student teacher EMIN regularized model. As in other UNESE experiments, when sufficient attention was given to the tuning and setup of the baseline model there were no observed statistically significant differences between EMIN regularization and baseline.

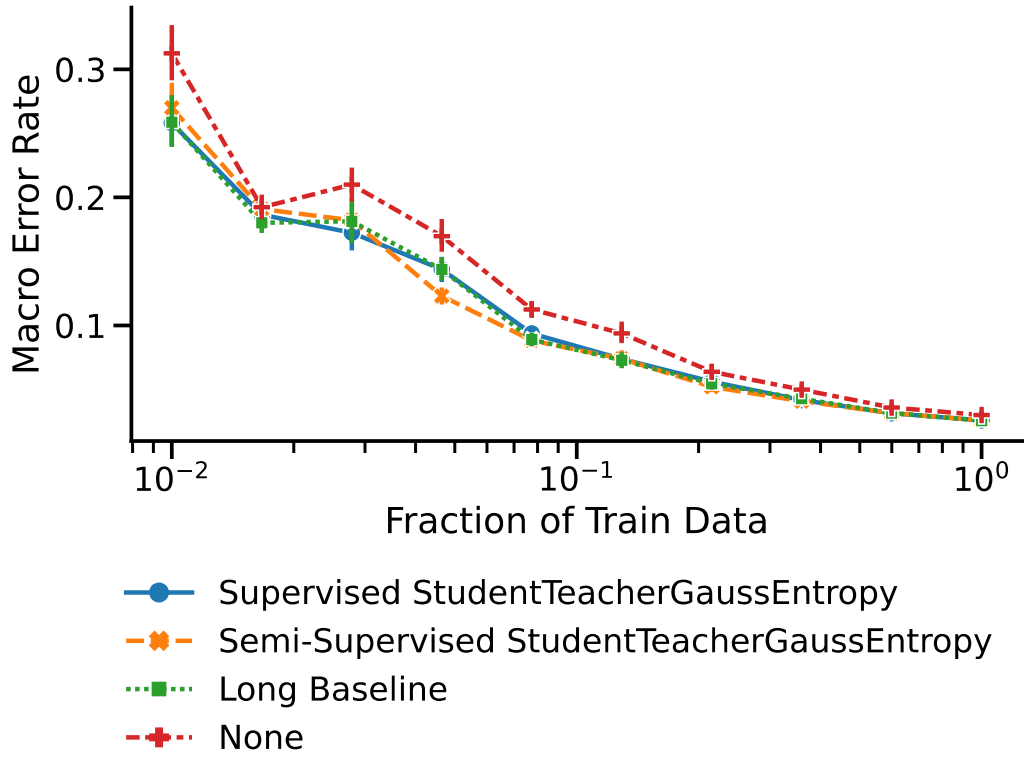

Figure 13. Student-Teacher based EMIN regularization results.

3.8. UNESE Discussion

Within the UNESE multi-modality, hyperspectral classification experiments entropy minimization-based regularization never showed statistically significant improvements compared to baselines in overall fused performance. In experiment 1, overall performance of tuned baselines was very good. At the same time, the joint hyperparameter tuning of 6 models required for EMIN regularization was cumbersome and unwieldy, representing a combinatorial expansion in the hyperparameter space for joint tuning. In response, the multi-headed hydra model was developed which proved far easier for hyperparameter tuning. The hydra architecture was evaluated under subsampled training labels, and while performance was correlated with available training data as expected, EMIN did not provide benefit over the degraded baseline in this regime. Analysis of these results highlighted a highly sensitive dependence on the value of α, the parameter that controlled the relative weighting between the supervised and EMIN loss terms. A less α-sensitive method of training was developed, instead leveraging EMIN as a finetuning step after normal supervised learning as pretraining. This two-step

approach illustrated that while fused performance between EMIN and baseline was flat, EMIN appeared to improve the performance of some individual modalities at the cost of others. To exploit this behavior, a student-teacher framework in which the best performing modalities (SWIR1-3) were used to “teach” the worst performing modalities (VNIR1-3). At first, this appeared to provide small but consistent improvement over baselines, but a further extra-long training cycle baseline closed the observed performance gap.

This dataset was characterized by a per-sample modality overlap that had most samples observed by 2-3 modalities. It is not clear that a statistical cross-modality regularization provides a useful target for learning in this regime. Furthermore, high performing baselines showed that often, per modality differences were insignificant and effectively quashed under simple mean-aggregation fusion. On a per-sample basis, errors between baselines and EMIN were similar, but EMIN often sacrificed the performance of the best performing modalities to improve the worst.

A strong recommendation from the UNESE experiments is that significant investment and experimentation should be paid into underlying baselines prior to developing advanced fusion approaches. Early stopping should be used with the utmost caution; models continue to learn and improve for long after validation metrics exhibit flat trends. The developed hydra model architecture is an elegant solution to data fusion while still using classic mean-aggregation fusion approaches, particularly when data modalities are similar enough to justify sharing most model weights. The hydra model resolves issues of inter-modality model score calibration, which can plague and degrade simple fusion aggregation techniques such as mean, by outputting all modality predictions for a sample into the same space.

Tuning of multi-term loss functions remains a significant challenge which cannot be decoupled from issues of loss and gradient scaling, learning dynamics, and training stability. There remains no obvious strategy to jointly optimize two or more loss terms at the same time. Sophisticated Pareto-front optimization strategies, such as MGDA explored in the UNESE experiments, require a quantitative understanding of relative loss importance. This is rarely possible in all but the most contrived examples, as direct interpretation of smooth, differentiable loss functions is

challenging and data dependent. Other strategies, such as random sampling of α, require a fixed or described schedule of relative loss weighting which can be unresponsive and ill-tuned to observed learning dynamics. Furthermore, random sampling-based strategies can struggle to find the often-narrow bands of α which produced desired results.

Under the EMIN project, the UNESE hyperspectral imagery collects were used to produce a realistic benchmark dataset for evaluating multi-modality multi-class classification algorithms. Appropriate data labels were developed, and the source rasters were subsequently sampled. The conditions of sparse and variable modality overlap, data alignment, etc. in this dataset are realistic exactly because they stem from real data collection processes. This dataset is available for future investigations into multi-modality fusion and is a lasting resource in support of algorithm development (Linville, 2022).

4. SEISMIC EVENT LOCATION REGRESSION

The problem of locating the source of an incoming signal from a single sensor remains a challenge for seismic event processing in part because directional information with traditional processing requires triaxial sensors which are often unavailable. Less expensive vertical axis sensors on the other hand can help constrain distance relationships owing to their relative abundance compared to other sensor types. We explore the advantages for direction and location prediction on vertical channel and triaxial sensors for a seismic network in the state of Utah. We first develop a location model through multitask learning to obtain a baseline for event location in Sections 4.1- 4.4. We then develop models that are trained with EMIN loss and evaluate the impact on directional and distance prediction for triaxial and vertical sensors in Section 4.5. We finish in Section 4.6 with an assessment of EMIN modelling in label sparce scenarios (semi-supervised learning) using student-teacher EMIN.

4.1. Direction (back azimuth) prediction

The direction of energy arriving from a seismic event observed at a specific location is typically measured in degrees from north looking toward the origin and is referred to as back azimuth or BAZ. High quality BAZ prediction with seismic data traditionally requires an array of stations. Arrays provide directional information when there are measurable differences in the arrival times of incoming energy across the geographically separated elements of the array. Less robust measurements are possible through polarization analysis on co-located orthogonal measurements available from individual triaxial sensors (Frohlic and Pullium, 1999). Observed accuracies for triaxial BAZ estimates are often on the order of 30° (Davenport et al., 2021), which can result in large, potentially unusable location errors if based off a small number of distant observations. Initial tests on single-station BAZ prediction with both triaxial and single vertical orientation sensors suggested that deep neural networks learn a sin/cos encoding of the BAZ attribute at a level only slightly better than existing methods for triaxial sensors (~18 km). However, when coupled with distance information in multitask learning BAZ improved markedly for both triaxial and single channel sensors.

4.2. Distance prediction

The distance to an observed phenomenology combined with the estimated BAZ provides a geographic (surface) location for a seismic event from a single sensor measurement. While there are no formal traditional methods to estimate distance with high precision using a single station, common 'rules of thumb' such as travel time distance between the arrival time of p and s waves can provide reasonable estimates for local and regional events. Distance is therefore usually considered the easier attribute for determining location because the errors that results from distance-only estimates from azimuthally distributed sensors result in higher location qualities than if BAZ is included at 18-to-30-degree resolution. Coupling distance and BAZ prediction benefits BAZ while not substantively inhibiting distance prediction with convolutional neural network (CNN) models for single stations. Therefore, because of the performance advantage multitask learning offers, distance and BAZ (sine/cos embedding) are predicted together in a single CNN model.

4.3. Dataset

The dataset for single-station location was developed by analysts and the University of Utah and has been leveraged extensively for algorithm development in the past (Linville et al, 2019, Linville, 2021). The regional extent of the dataset includes the monitoring region around the state of Utah with event locations from 34.7-46.3° N Latitude and 117.0-106.7° W Longitude. Observations come from a network of 280 stations of which 66% are single vertical channel sensors (111,606 observations) and 34% are from 3 orthogonal channel sensors (58,202). The maximum observation distance is 445km, however most events are observed at local distance scales < 150km (catalog mean=51km). The mean reported horizontal error reported in the catalog is 1km. Partitioning for training, testing, and validation was performed randomly over events (events remained grouped during sampling) and 10% of the samples were used for testing and validation, respectively.

4.4. Baseline single-station location models

Baseline models utilize 3-channel (zero padded for vertical sensors) time-frequency (spectrogram) input representations. The spectrograms have 1sec time resolution and 1Hz frequency resolution. While we explored CNN models of various depths, the

largest models were typically the most competent at multitask learning for event location. Final models used the VGG16 architecture (Simonyan and Zisserman, 2014) and resulted in median location errors over the randomly sampled test catalog of 3.2 – 3.5 km from 1.1-1.3 km distance errors and 3.5-3.7° BAZ errors. Median event level errors were 1.69km (mean=4.4). Median location errors for single channel models are 4.1km compared to 2.8km in triaxial models. There are substantive differences between the average and the median for reported location errors (17km vs 3.5km) because a majority of the events are close (< 51 km distant) and maximum location error is controlled by distance. Therefore, in selecting (through validation) the models with the smallest median error we are preferentially selecting models that are best for short distances, but these models also perform the best in aggregate over all distance samples. These experiments were run on a subset of the catalog: earthquake sources with reported horizontal error < 1km (38% of the catalog). For validation several full catalog runs were performed, which increased performance of the best models to 2.6 km median location error.

4.5. Experiment 1: Seismic EMIN models

Seismic EMIN models were trained with architectures, hyperparameters, and loss in the same manner as baseline models, with the exception that triaxial and single channel modalities comprised separate models and took in their respective inputs. The EMIN objective was a function of the predicted BAZ (in radians) and distance (km normalized the earth radius = 6378.1) and required a starting location (*startlat*) for each observation *i* in set of stations (for a single event, for simplicity):

$$X_i = \arcsin\left(\sin\left(startlat_i\right)\cos(dist_i) + \cos(startlat_i)\sin\left(dist_i\right)\cos\left(BA\right)\right. \quad \textit{Equation 2}$$

$$Y_i = startlon_i + \arctan\frac{\left(\sin\left(BAZ_i\right)\sin(dist_i)\cos(startlat_i)\right)}{\cos(dist_i) - \sin(startlat_i)\sin\left(X_i\right)} \quad \textit{Equation 3}$$

EMIN minimized the variance of the resulting distributions of X_i,Y_i from each model *and* for each event. In the case of student-teacher EMIN for seismic, distributions

were stabilized using the median values of X_i,Y_i for set of stations in an event but EMIN was only applied to the lowest performing model (the single channel modality, the student in the student-teacher paradigm).

EMIN models in aggregate did not outperform baseline models within statistically meaningful margins unless considering observations within 40 km (1 std of the catalog distance distribution; Figure 14a). Triaxial models remained better on average than vertical channel sensors, although the performance differences per modality depended on distance.

There was a difference between behaviors for both baseline and EMIN models on the best and worst performing modalities on event level classification. Namely, triaxial performance improved for most events but worsened for the 'hardest' cases. In contrast, the EMIN single channel models were better, even if very marginally (Figure 14e).

When the dataset was extended beyond the most well constrained earthquake locations to include all cataloged events, the performance for EMIN models was statistically equal to that of baseline models (2.6 km median error). Performance differences between sensor types at different distance ranges also became nearly equivalent for both very near and very distant observations. For all models, predictive variance was an unreliable estimate of predictive accuracy for both BAZ and distance attributes. In order to be informative for decision making, the least stable distance and BAZ estimates should clearly relate to the largest prediction errors, which they did not.

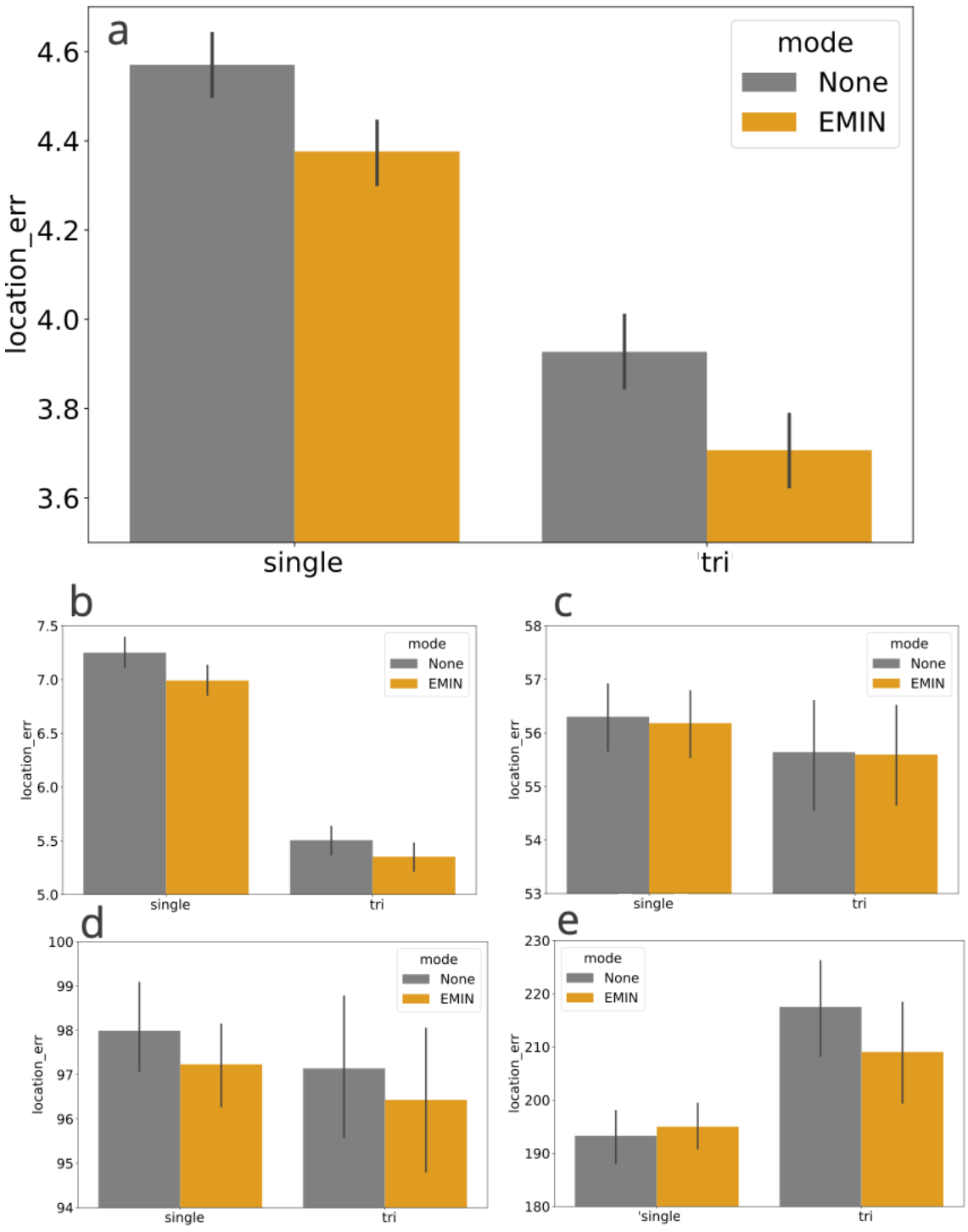

Figure 14. Location error in km for catalog subsets. Panels are broken into subsections of the catalog based off standard deviation of location errors in km. a: 1 std = ~40 km, b: 1std = 88.9% of the catalog, c: 2std = 5.6%, d: 3std = 2.2%, e: 4std = 3.1%.

4.6. Experiment 2: Semi-supervised seismic location

In contrast to supervised learning loss, formulations of the EMIN objective do not require distance and BAZ labels as input. EMIN can therefore be used in semi-supervised learning (SSL) tasks where the requirements of learning are a sparse number of known labels and some additional quantity of unlabeled data. The SSL experiments performed here use 25 event labels which equates to 262 labeled samples. For the randomly sampled label set explored here, SSL EMIN reduced the km location error by 15.5 km (median over test set) and the km location error per km distance by .8 - 1.4 km, on average half compared to non-EMIN models (Figure 15a).

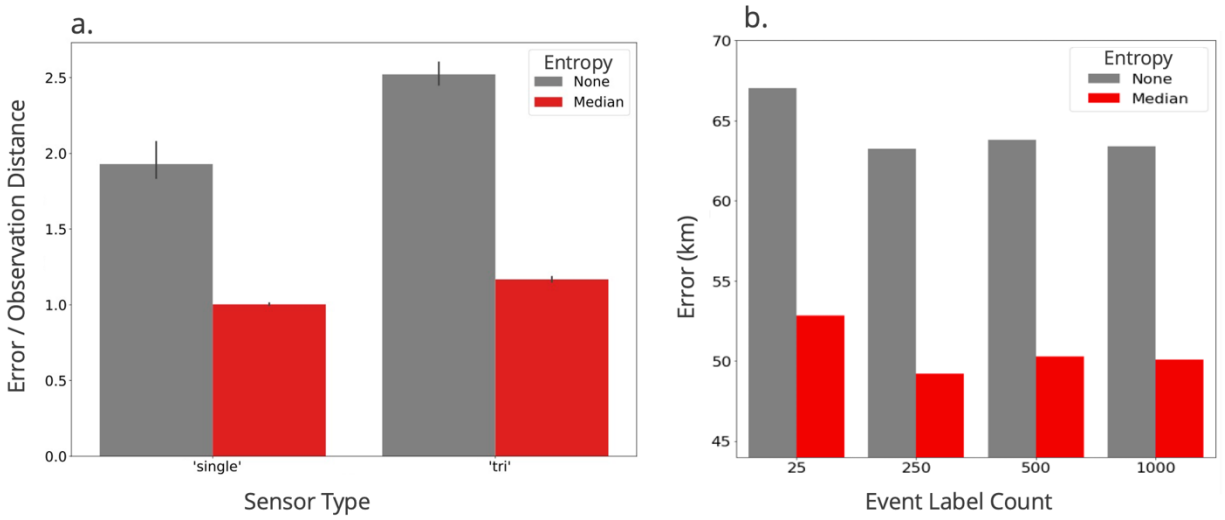

Figure 15. SSML EMIN location error normalized by observation distance (a) and with different label counts (b). Models with labels counts above 500 are assumed to have sub-optimally explored HP based on their lack of decrease in model error given additional labels. This is due to the time available to continue to perform model experiments for this study.

In geographic terms, aggregate errors over the catalog reduce from ~67 to 53 km (Figure 15b) and estimates for both modalities improve dramatically at the event level (Figure 16; red compared to grey error ellipses). While clear performance gains have been demonstrated, the error achieved in these examples is sufficiently high that with or without EMIN these approaches may be only minimally useful for event location in practice.

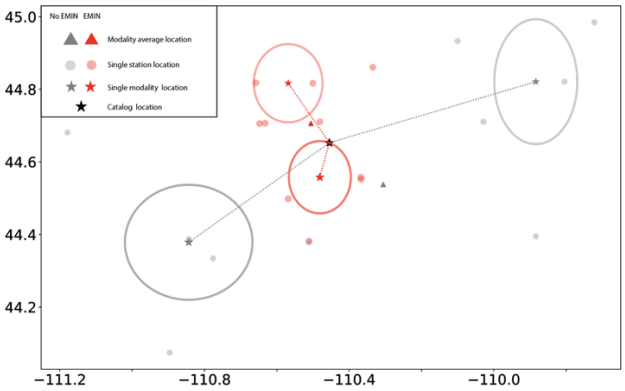

Figure 16. Model location ellipses for baseline models (grey) and EMIN models (red).

4.7. Seismic Discussion

In the best models, triaxial sensor median error was 2km compared to 3km for single channel sensors or 9.5km compared to 16km when using the mean. While single channel models were more numerous at 66% of the catalog, triaxial estimates were more accurate, with the exception of long- distance observations. Recognizing the disparate aptitudes across modalities, EMIN can be formulated within a student-teacher paradigm where only the teacher (the triaxial modality) influences the decisions of the student (single channel sensors). In these experiments regardless of the EMIN objective formulation (enforced across the entire distribution instead of just the per-modality median, using the modality median to stabilize outliers, or as the modality median of the teacher) EMIN was either helpful or at least not damaging for learning. The stability of the performance for EMIN in seismic experiments was likely the result of consistent azimuthal coverage for each event. Consistent azimuthal coverage provides distributed error averaging and helps minimize the impact of poorly performing individual sensors on location estimates and the loss objectives. Therefore, while it is possible to formulate the EMIN objective function to be less sensitive to outliers with maximum a posteriori probability or Bayesian formulations, these are less likely to be advantageous for this dataset and therefore testing of these formulations remained limited. When EMIN objectives were optimized using the student-teacher methods, loss occurred more stably (monotonically decreases) but when trained to completion with appropriate hyperparameters performance remained equivalent. The only cases where EMIN generated substantive performance differences were in severely label-limited cases, where performance is doubled relative to baselines (errors are halved when normalized by distance). This suggests that a redundant precision forcing, which is what location entropy minimization does in multitask supervised regression settings, adds little to no value in accuracy or predictive confidence. The non-redundance of the information provided by unlabeled samples through EMIN proved to be powerful, however, the gains are unlikely to scale linearly with label size under random sampling. Optimal label selection for these experiments is the next step in understanding the practical use cases for SSL EMIN for specific problems (e.g. how accurate do single-station locations need to be to add value beyond distance-based

predictions). However, label selection remains a foundational question in the field of SSL broadly and is beyond the scope of this work.

Lastly, enforcing consistency when doing so is mostly redundant (as in the supervised seismic location modelling) increases the HP search space without the commensurate benefits of reducing training time or significant increases in performance. In this scenario we expected mostly the ‘hard cases’ to be improved by model clamping, and the more difficult attribute (BAZ) did in fact improve compared to unclamped models, but often at the expense of precision across distance, and with gains that decrease as training epoch increased. After 4 days of model training there was no recognizable difference between clamped and unclamped models for this task. The model gained insight about how it performed on current samples based on how it performed on related samples, but this is the same insight accessible through labeled loss given sufficient training times (8 days for baseline models on the full dataset compared to 41 days for EMIN).

5. DISCUSSION

The EMIN objective is a theoretically simple idea: if models consistently and correctly assign sample x a set of attributes, other models with less descriptive data should incorporate x into the representative space of what those attributes can look like. Intuitively it seems that EMIN loss would be increasingly valuable when instances of x are rare in one modality, and even more valuable when instances of x exist but for which no ground truth is available. While theoretically simple, the complexity introduced in real data generally drives performance in far greater ways than EMIN has opportunities to leverage. For example, enforcing EMIN loss over arbitrary (non-physical attributes) is more likely to lead to poor generalization, and reduced inherent (and perhaps desired) uncertainty for phenomenologies with differential detectability across modalities. A more concretely example would be enforcing agreeance over time-of-day for modelling quarry blasts in a region where blasting is not restricted by daylight. This is not likely to be valuable. When applying EMIN for a specific phenomenology, the hard question is whether detectability for some samples is poor because no signal exists, or because the signal is weak or previously unrecognized (i.e., hard cases with valuable aspects that traditional processing has failed to recognize). If no signal exists, enforcing a noise sample into positive class may not be ideal. At the very least the expectation is that the adversarial relationship between class loss and consistency loss results in low predictive confidence for these samples. This was true for training samples but does not reliably translate to new (test) samples in the experiments explore in this work.

Complexity introduced by the domain is also coupled to learning dynamics. For example, weighting labeled loss as equivalent in importance to EMIN loss was generally detrimental to learning. Learning suffers because models struggle to escape low entropy initialization states or can be driven to mode collapse (another low entropy state). This means that elegant solutions to multi-objective optimization are not helpful and instead brute force search space approaches are required. Problems like this compound as the number of modalities increase because joint model optimization becomes increasingly intractable. Shared parameters with independent model heads for each modality (the HSI hydra model) are an efficient

solution to unwieldy parameter spaces, but when EMIN and labeled loss are at odds for a given sample late in learning, it is likely because there are few or no helpful features left to exploit beyond labeled learning for a given dataset. In these cases, EMIN at best can induce tradeoffs in prediction for specific samples but given a labeled learning loss of sufficient strength, these differences will not be substantively different from what is achieved in the absence of EMIN from different weight initialization and training pathways. This is essentially what we observe in the first two HSI experiments. In cases where we optimize labeled learning and subsequently apply EMIN, we observed expected behaviors given the training objective (sample entropy decreases) which had positive impacts on poor performing modalities but those came at the expense of higher performing modalities. Simple mitigation strategies follow a student-teacher learning scenario but likewise failed to offer substantive benefit even in label-limited experiments likely because as in previous experiments, EMIN forcing did not help in developing additional predictive features.

The optimization challenges discussed above were not limiting for seismic experiments as they were for HSI because the number of modalities were limited to 2. Instead, complexity in applying EMIN was shifted to the development and stable application of EMIN on a non-trivial transform of the model output. Physically based learning objectives must be carefully normalized and learning dynamics investigated extensively in order to verify gradient pathways and convergence as expected. Once accomplished, we found that EMIN forcing on seismic regression tasks is a superfluous precision forcing term in fully labeled domains. Given sufficient time, labeled loss alone was able to arrive at the same location accuracy, and did so more rapidly (8 days vs 41) under a broader HP range. This is not to say EMIN did not demonstrate value. As in the HSI experiments, there were cases for which EMIN improved performance, but in aggregate across the dataset those gains were a trade-off. The one case where EMIN provided clear and substantive gains was for semi-supervised regression where unlabeled data was able to decreases error rates to half their non-EMIN values. Not only did average and median station values improve, but event level estimates for each modality became substantially more

aligned with each other (calibrated). The reason this outcome is not considered a substantive win for general data fusion is that engineering complexity required to arrive there is formidable and will be required for each new problem, making it a highly specialized rather than a general solution to label-limited but data rich problems.

6. CONCLUSIONS

This work developed a new training method (EMIN) for gradient based predictive models that expects low entropy when multiple data sources, channels, or modalities exist for a given decision. Conceived as a general method of data fusion that maintains input representation flexibility between modalities, EMIN helps calibrate decisions from automated models across modalities and can increase predictive performance substantially in very specific label-limited cases. Despite generalizable theory, EMIN in practice requires substantive engineering for each specific problem and is not guaranteed to enhance performance when the labels that exist for learning already exploit the features of the data available for the predictive tasks.

REFERENCES

Anderson, D. N., Fagan, D. K., Tinker, M. A., Kraft, G. D., & Hutchenson, K. D. (2007). A mathematical statistics formulation of the teleseismic explosion identification problem with multiple discriminants. *Bulletin of the Seismological Society of America*, *97*(5), 1730-1741.

Davenport, K., Linville, L., & Young, C. (2021, December). Single-Station Backazimuth Estimations for Earthquake & Explosive Sources. In *AGU Fall Meeting Abstracts* (Vol. 2021, pp. S15C-0254).

Frohlich, C., & Pulliam, J. (1999). Single-station location of seismic events: a review and a plea for more research. *Physics of the Earth and Planetary Interiors*, *113*(1-4), 277-291.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, *14*(1), 28-44.

Linville, L., Pankow, K., & Draelos, T. (2019). Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, *46*(7), 3643-3651.

Linville, L. (2021). Utah Discrimination Datasets 2012-2020, ISC Seismological Dataset Repository, https://doi.org/10.31905/RDQW00CT

Linville, L. A., Dylan; Michalenko, Joshua; Garcia, Jorge. (2022). *Multi-modal-hyperspectral-classification-dataset*: https://opendata.sandia.gov/datasets/?category=satellite.

Michalenko, J. J., Linville, L. M., & Anderson, D. Z. (2020). Multimodal Data Fusion via Entropy Minimization. IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium.

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, *31*.

Simonson, K. M. (1998). *Probabilistic fusion of ATR results*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

UNESE. (2018). *Underground Nuclear Explosions Signatures Experiment (UNESE) Project FY2018 Report* (DOE/NV/03624—0355).

Vidale, J. E. (1986). Complex polarization analysis of particle motion. *Bulletin of the Seismological society of America*, *76*(5), 1393-1405.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Technical Library	1911	sanddocs@sandia.gov

This page left blank

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.