10-2020

Multimodal Representation Learning and Set Attention for LWIR In-Scene Atmospheric Compensation

Nicholas M. Westing

Kevin C. Gross
Resonant Sciences

Brett J. Borghetti
Air Force Institute of Technology

Christine M. Schubert Kabban
Air Force Institute of Technology

Jacob Martin
Air Force Research Laboratory

See next page for additional authors

Recommended Citation

N. Westing, K. C. Gross, B. J. Borghetti, C. M. S. Kabban, J. Martin and J. Meola, "Multimodal Representation Learning and Set Attention for LWIR In-Scene Atmospheric Compensation," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 127-140, 2021, doi: https://doi.org/10.1109/JSTARS.2020.3034421

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Multimodal Representation Learning and Set Attention for LWIR In-Scene Atmospheric Compensation

Nicholas Westing, *Student Member, IEEE,* Kevin C. Gross, Brett J. Borghetti, Christine M. Schubert Kabban, Jacob Martin, and Joseph Meola

***Abstract*—A multimodal generative modeling approach combined with permutation-invariant set attention is investigated in this paper to support long-wave infrared (LWIR) in-scene atmospheric compensation. The generative model can produce realistic atmospheric state vectors ($\mathrm{T}, \mathrm{H_2O}, \mathrm{O_3}$) and their corresponding transmittance, upwelling radiance, and downwelling radiance (TUD) vectors by sampling a low-dimensional space. Variational loss, LWIR radiative transfer loss and atmospheric state loss constrain the low-dimensional space, resulting in lower reconstruction error compared to standard mean-squared error approaches. A permutation-invariant network predicts the generative model low-dimensional components from in-scene data, allowing for simultaneous estimates of the atmospheric state and TUD vector. Forward modeling the predicted atmospheric state vector results in a second atmospheric compensation estimate. Results are reported for collected LWIR data and compared against Fast Line-of-Sight Atmospheric Analysis of Hypercubes - Infrared (FLAASH-IR), demonstrating commensurate performance when applied to a target detection scenario. Additionally, an approximate 8 times reduction in detection time is realized using this neural network-based algorithm compared to FLAASH-IR. Accelerating the target detection pipeline while providing multiple atmospheric estimates is necessary for many real-world, time sensitive tasks.**

***Index Terms*—Hyperspectral Imagery, Atmospheric Compensation, Neural Networks, Generative Modeling, Target Detection**

I. Introduction

LONG wave infrared hyperspectral sensors collect data between 8 - 14 μm across hundreds of contiguous bands, providing detailed information about the earth's surface and material temperatures. Accurate characterization of surface constituents is important for a wide range of applications such as urban heat island analysis, search and rescue operations and target detection [1]–[3]. Fully leveraging thermal hyperspectral data for these applications requires precise atmospheric compensation algorithms for accurate material characterization. Additionally, these compensation methods should be efficient and require minimal user input to operate on the large volumes of data collected by modern sensors. This paper extends previous research in efficient long-wave infrared (LWIR) atmospheric compensation [4], investigating new architectures to form a joint representation of atmospheric measurements and their corresponding radiometric quantities. These radiometric quantities are atmospheric transmission, $\tau(\lambda)$, upwelling radiance, $L_a(\lambda)$, and downwelling radiance, $L_d(\lambda)$. The major contributions of this paper are:

- The Multimodal DeepSet Atmospheric Compensation (MDAC) architecture is introduced, predicting both atmospheric state ($\mathrm{T}, \mathrm{H_2O}, \mathrm{O_3}$) and the $\tau(\lambda)$, $L_a(\lambda)$, and $L_d(\lambda)$ vectors to support in-scene atmospheric compensation.
- Variational loss and weighted atmospheric state loss are shown to reduce reconstruction error compared to mean-squared error (MSE) loss functions using a Multimodal Autoencoder (MMAE) architecture.
- Set attention pooling is investigated to understand reflective pixels' role in the MDAC prediction. Emphasis of reflective pixels in the atmospheric compensation prediction is in agreement with the LWIR radiative transfer equation.
- Atmospheric compensation errors are compared from the target detection perspective using collected LWIR data, demonstrating comparable performance to Fast Line-of-Sight Atmospheric Analysis of Hypercubes - Infrared (FLAASH-IR) while reducing total detection time.

The remainder of this paper is structured as follows: in the next section, a review of permutation-invariant neural networks for LWIR atmospheric compensation is discussed, followed by an overview of LWIR hyperspectral data processing. Section III introduces the MMAE and MDAC architectures, the loss functions used to fit these models, and metrics for evaluating model performance. Section IV reports results on synthetic and collected data using the previously defined metrics and Section V summarizes major conclusions of this research.

II. Background

Given N pixels $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, extracted from a data cube collected from an altitude a_s across K bands, $\mathbf{x}_i \in \mathbb{R}^K$, the DeepSet Atmospheric Compensation (DAC) network, $D(\mathbf{X}, a_s)$, predicts a low-dimensional representation, $\mathbf{z}$, of the estimated transmittance, upwelling radiance, and

N. Westing and B. Borghetti are with the Department of Electrical and Computer Engineering, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433, USA e-mail: nicholas.westing.1@us.af.mil

C. Schubert Kabban is with the Department of Mathematics and Statistics, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433, USA

K. Gross is with Resonant Sciences, Beavercreek, OH 45430, USA, and is Adjunct Faculty with Air Force Institute of Technology.

J. Martin and J. Meola are with the Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433, USA.

Manuscript received July 27, 2020; revised September 14, 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

downwelling radiance (TUD) ($\hat{\tau}(\lambda), \hat{L}_a(\lambda), \hat{L}_d(\lambda)$) vector, $\mathbf{y}_T$ [4]. A decoder network, $d(\cdot)$, transforms $\mathbf{z}$ to $\mathbf{y}_T$, such that

$$\hat{\mathbf{y}}_T = d(D(\mathbf{X}, a_s)) \tag{1}$$

The pixel set $\mathbf{X}$ corresponds to a single $\mathbf{y}_T$ vector and the DAC network should provide the same $\mathbf{y}_T$ prediction regardless of the order of the pixels in $\mathbf{X}$. To achieve this functionality, the DAC network is permutation-invariant to pixel order in $\mathbf{X}$, relying on a set transformation operation $\phi(\cdot)$ and a max pooling operation to form a one-dimensional set feature vector. The low-dimensional $\mathbf{z}$ prediction is made with another prediction network, $\rho(\cdot)$, such that the entire DAC network can be expressed by:

$$D(\mathbf{X}, a_s) = \rho\left(\max_{i \in N}\left[\phi(\mathbf{X})\right], a_s\right). \tag{2}$$

Instead of max pooling the transformed set representations created by the $\phi(\cdot)$ network, this research leverages recent advancements in set attention pooling to perform the set decomposition operation [5], [6]. Attention mechanisms are loosely based on how human vision operates: focusing on objects of high importance while blurring background objects. By focusing or attending to the most salient data aspects for a particular task, model performance can be improved while also increasing interpretability [7]. These advantages are achieved through a weighted average where the weights are attention scores that highlight feature importance.

Set attention pooling is a modified attention mechanism used in cases where multiple instances correspond to a single output value [5], [6]. Some samples in the set will contain more information, captured by the set attention scores, and have a stronger influence on the set decomposition operation. Set attention pooling is of interest to the LWIR atmospheric compensation problem because pixels receiving higher attention scores can be further investigated to identify unique spectral properties. This additional interpretability is necessary for validating model performance on a wide range of conditions.

In addition to set attention pooling, this research also extends [4] by investigating a multimodal representation. The decoder network $d(\cdot)$ in [4] utilized the TUD vector data to create the low-dimensional data manifold $\mathbf{z}$, however, this research also utilizes the atmospheric state vector, $\mathbf{y}_A$, creating a MMAE to constrain the data manifold. Evaluating the benefits of these modifications requires a review of LWIR hyperspectral data analysis discussed next.

The observed at-sensor radiance, $L(\lambda)$, consists of two factors: surface-leaving radiance, $L_s(\lambda)$, attenuated by atmospheric transmission, and atmospheric emission directly to the sensor. Assuming a Lambertian surface and monochromatic light, the simplified LWIR radiative transfer equation can be described as [8]:

$$L(\lambda) = \tau(\lambda)L_s(\lambda) + L_a(\lambda) \tag{3}$$

where $L_s(\lambda)$ consists of emissive and reflective contributions:

$$L_s(\lambda) = \underbrace{\epsilon(\lambda)B(\lambda,T)}_{\text{Emissive}} + \underbrace{\left[1-\epsilon(\lambda)\right]L_d(\lambda)}_{\text{Reflective}}. \tag{4}$$

Based on these definitions, the entire simplified at-sensor radiance equation can be described by:

$$L(\lambda) = \tau(\lambda)\Big[\epsilon(\lambda)B(\lambda,T) + \big[1-\epsilon(\lambda)\big]L_d(\lambda)\Big] + L_a(\lambda) \tag{5}$$

λ : wavelength
T : material temperature
$\tau(\lambda)$: atmospheric transmission
$\epsilon(\lambda)$: material emissivity
$B(\lambda, T)$: Planckian distribution
$L_d(\lambda)$: downwelling atmospheric radiance
$L_a(\lambda)$: atmospheric path (upwelling) radiance

The Planckian distribution is:

$$B(\lambda,T) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT}-1} \tag{6}$$

where c is the speed of light, k is Boltzmann's constant and h is Planck's constant.

The signal of interest in LWIR target detection is the material emissivity defined as a ratio between the radiance emitted at temperature T and the radiance emitted by a blackbody ($\epsilon(\lambda) = 1$) at the same temperature [9]:

$$\epsilon(\lambda) = \frac{L(\lambda,T)}{B(\lambda,T)} \tag{7}$$

Retrieving emissivity consists of two steps: atmospheric compensation and temperature/emissivity separation (TES). Atmospheric compensation methods estimate the TUD vector, such that surface leaving radiance can be recovered. Model-based atmospheric compensation approaches rely on radiative transfer models such as MODerate resolution atmospheric TRANsmission (MODTRAN) to predict TUD vectors based on known or estimated atmospheric state information (column water vapor, trace gas content, air temperature) [10], [11]. By generating a look-up table of TUD vectors from expected atmospheric conditions, model-based methods can be implemented efficiently for real-time use [12]. Specifically, methods such as FLAASH-IR modify the surface temperature, water vapor column density and the ozone scaling factor to minimize the error between observed and predicted radiance [10].

In-scene atmospheric compensation methods rely on blackbody pixels to make the compensation problem tractable. The In-Scene Atmospheric Compensation (ISAC) method identifies blackbody pixels allowing at-sensor radiance, $L_{BB}(\lambda)$, to be described by [13]:

$$L_{BB}(\lambda) = \tau(\lambda)B(\lambda,T) + L_a(\lambda) \tag{8}$$

Pixel temperature is estimated through clear bands ($\tau(\lambda) \approx 1$), such that the only remaining unknowns are $\tau(\lambda)$ and $L_a(\lambda)$. A linear fit is performed on each spectral channel to determine these terms. The ISAC procedure does not recover the downwelling radiance, important for accurately characterizing reflective materials.

Next, TES is typically performed to estimate both $\hat{\epsilon}(\lambda)$ and $\hat{T}$. For a sensor with K spectral bands, decoupling these terms is an under-determined problem as there are only K

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

measurements but $K+1$ unknowns $(\hat{\boldsymbol{\epsilon}}, \hat{T})$. A common approach to this under-determined problem is to assume $\epsilon(\lambda)$ is a smooth function of wavelength compared to the atmospheric features [14]. Assuming downwelling radiance was estimated during the atmospheric compensation process, emissivity can be estimated as [9]:

$$\hat{\epsilon}(\lambda) = \frac{\hat{L}_s(\lambda) - \hat{L}_d(\lambda)}{B(\lambda, \hat{T}) - \hat{L}_d(\lambda)} \tag{9}$$

Unfortunately, TES methods recover material temperatures with limited accuracy, leading to increased errors in $\hat{\epsilon}(\lambda)$ [15]. Unique from TES procedures, researchers have investigated methods to determine $\hat{\epsilon}(\lambda)$ with less dependence on $\hat{T}$. The alpha residuals approach introduced in [16] and extended in [17] converts a target emissivity, $\epsilon_t(\lambda)$, to $\alpha_{\epsilon_t}(\lambda)$ by:

$$\alpha_{\epsilon_t}(\lambda_i) = \lambda_i \ln\left[\epsilon_t(\lambda_i)\right] - \frac{1}{K}\sum_{j=1}^{K} \lambda_j \ln\left[\epsilon_t(\lambda_j)\right] \tag{10}$$

The alpha residual formulation presented in [16] and [17] omits the reflective component in the surface leaving radiance. In [18], the reflective component was included allowing improved emissivity estimation for reflective and emissive materials. In both [17] and [18] an estimate of pixel temperature is needed but target signal estimation is robust to temperature estimation errors.

Both TES and alpha residual approaches rely on TUD vector estimates derived from the atmospheric compensation process. This study presents an efficient method for in-scene LWIR atmospheric compensation and compares this method's performance using both TES and alpha residuals from a target detection perspective.

III. Methodology

The MDAC model, $D_m(\cdot)$, predicts a low-dimensional representation, $\hat{\mathbf{z}}$, of both the scene atmospheric state vector, $\hat{\mathbf{y}}_A$, and the TUD vector, $\hat{\mathbf{y}}_T$. A multimodal decoder, $d_m(\cdot)$, is used to reconstruct both outputs from $\hat{\mathbf{z}}$ such that the atmospheric compensation and atmospheric state estimation problem can be described by:

$$\hat{\mathbf{y}}_{\mathrm{A}}, \hat{\mathbf{y}}_{\mathrm{T}} = d_m(D_m(\mathbf{X}, a_s)) \tag{11}$$

This result depends on the ability of the MDAC model to predict the latent space components $\mathbf{z}$ from the set $\mathbf{X}$ and the decoder model to reconstruct $\mathbf{y}_A$ and $\mathbf{y}_T$ from $\mathbf{z}$. The decoder model is a part of the overall MMAE (Figure 1) that is trained prior to fitting the MDAC network. The MMAE model architecture and training is explained in the next section.

A. Multimodal Generative Models

To fit the MMAE model, training data consisting of atmospheric state vectors and corresponding TUD vectors are needed. This research utilizes the Thermodynamic Initial Guess Retrieval (TIGR) database containing 2,311 atmospheric measurements selected from over 80,000 worldwide measurements [19], [20]. Each sample contains temperature,

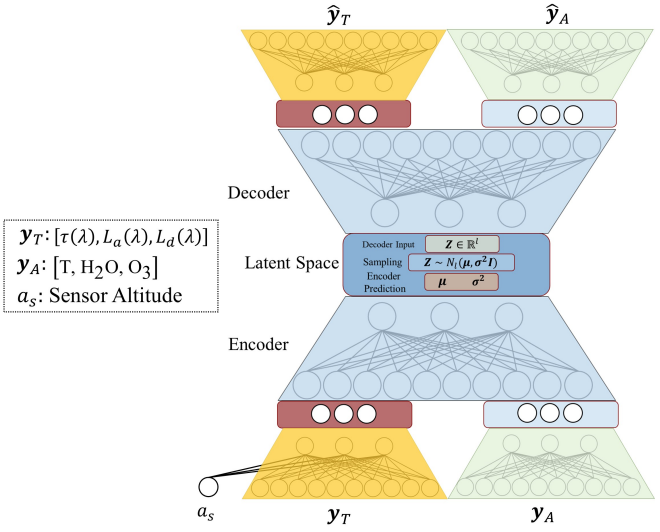

Fig. 1. TUD vectors are compressed by the encoder into the latent space and then reconstructed by the decoder network. Reconstruction error is minimized through weight updates during the training process. Additionally, a scalar altitude input is also presented with the TUD vector allowing the model to scale to multiple altitudes.

water vapor content and ozone content as a function of pressure level starting at the Earth's surface ($1013\,\mathrm{hPa}$) to greater than $30\,\mathrm{km}$ ($< 1\,\mathrm{hPa}$). Additionally, the measurements are grouped by air mass category such as polar, tropical and mid-latitude. The 2,311 measurements are filtered for cloud free conditions using a 96% relative humidity threshold, resulting in 1,790 cloud free measurements.

In [21] reconstruction error was reduced by augmenting the TIGR samples using a dimension reduction approach. This research also leverages the same augmentation approach to increase the number of training samples. First, principal component analysis (PCA) is applied to the concatenated atmospheric measurements (T, H_2O, O_3) using 15 components for each air mass category. Next, a 10 mixture Gaussian mixture model (GMM) is fit to the 15-dimensional space and new atmospheric measurements are created by sampling the GMM. These augmented measurements must still meet the 96% relative humidity threshold to be included in the training data. The result of this process is shown in Figure 2 for the polar air mass. This augmentation process is repeated for each air mass resulting in an additional 8,450 atmospheric state vectors.

These atmospheric state vectors (augmented and TIGR) were forward modeled with MODTRAN 6.0 at $0.005\,\mu\mathrm{m}$ spectral resolution, assuming a nadir sensor zenith angle. In Section IV-C, this research is applied to data cubes collected from altitudes of $0.45\,\mathrm{km}$, $0.92\,\mathrm{km}$ and $1.22\,\mathrm{km}$. To include this altitude variability in the training data, altitudes between $0.15\,\mathrm{km}$ - $3.05\,\mathrm{km}$ were used to forward model the atmospheric state vectors, resulting in 143,640 TUD vectors. The high resolution TUD vectors created by MODTRAN were downsampled to the Spatially Enhanced Broadband Array Spectrograph System (SEBASS) instrument line shape (ILS) to create a sensor-specific TUD database [22]. This downsam-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

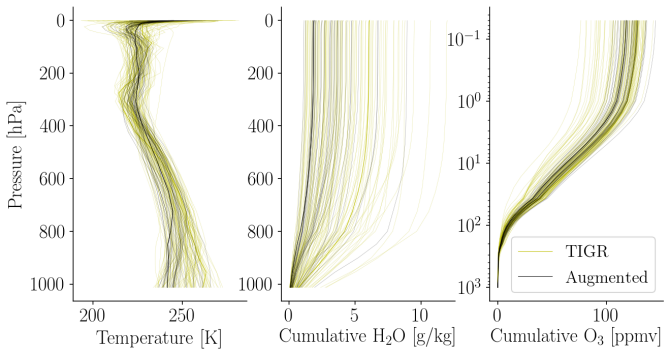

Fig. 2. Polar air mass atmospheric state vectors are shown for the TIGR data and the augmented samples. These augmented samples were created by performing PCA on the atmospheric state vectors and then fitting a GMM to the low-dimensional space as outlined in [21].

pling process assumed a Gaussian lineshape across 92 spectral bands between $8.13\,\mu\text{m}$ and $12.48\,\mu\text{m}$. Validation samples were created from held out atmospheric state vectors at unique altitudes from the training data.

A MMAE is used to compress both the atmospheric state vector and the TUD vector into a joint latent space, $\mathbf{z}$. MMAEs have been investigated in other domains such as audio and video where it is possible to generate one mode from the other [23]. In this research, both modes are always present during training since only the MMAE decoder is used for atmospheric compensation. The MMAE architecture is leveraged to improve feature fusion compared to concatenating the TUD and atmospheric state vectors.

Independent input and output branches combined through joint encoder and decoder networks are used to form the MMAE. The $\mathbf{y}_T$ encoder consists of two layers of 25 and 10 nodes and the $\mathbf{y}_A$ encoder consists of two layers of 20 and 15 nodes. The joint encoder takes the concatenated 10 and 15 node encoder outputs and transforms this representation to the latent space using two layers of 16 and 10 nodes. The latent space is the bottleneck in the representation learning problem, with 6 dimensions considered in this research based on previous results from TUD vector compression [4], [21]. This compression operation is reversed as shown in Figure 1 to create the decoder model.

Interpolations across the latent space should lead to semantically smooth variations in both atmospheric state and TUD outputs. This is a necessary property to support MDAC latent space sampling and is achieved by enforcing a prior distribution on the latent space. This research applies a Gaussian prior such that $\mathbf{z} \sim p(\mathbf{z}) = \mathrm{N}(\mathbf{0}, \mathbf{I})$. This constraint is used in Variational Autoencoders (VAEs) [24] and was extended in [25] for multiple modalities to define a joint multimodal VAE. Given the atmospheric state vector, $\mathbf{y}_A$, and the TUD vector, $\mathbf{y}_T$, the joint multimodal VAE generative processes for these modes are [25]:

$$\mathbf{y}_A, \mathbf{y}_T \sim p(\mathbf{y}_A, \mathbf{y}_T \mid \mathbf{z}) = p_{\theta_A}(\mathbf{y}_A \mid \mathbf{z}) p_{\theta_T}(\mathbf{y}_T \mid \mathbf{z}) \tag{12}$$

where the parameter θ represents the decoder network for each mode. The encoder network, q_ϕ, predicts distribution parameters $\boldsymbol{\mu} \in \mathbb{R}^{1\times c}$, $\boldsymbol{\sigma} \in \mathbb{R}^{1\times c}$ for a latent space with c components. Using the reparameterization trick introduced in [24], the posterior $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{y}_A, \mathbf{y}_T)$ can be sampled according to $\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. To enforce the prior distribution on the latent components, the Kullback-Leibler (KL) divergence is calculated according to [24]:

$$\mathscr{L}_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{y}_A, \mathbf{y}_T) \parallel p(\mathbf{z})\big) = \frac{1}{2}\sum_{j=1}^{c}\left(1 + \log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2\right) \tag{13}$$

While $\mathscr{L}_{KL}$ enforces a prior distribution on the latent components, atmospheric state and TUD vector reconstruction error must also be minimized to provide a useful model. Similar to previous work [4], [26], the TUD vector reconstruction error is minimized using

$$\mathscr{L}_T(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{3K}\sum_{i=1}^{3K}(\hat{y}_i - y_i)^2 + \frac{\gamma}{MK}\sum_{j=1}^{M}\sum_{i=1}^{K}\left(L_{\hat{\mathbf{y}}}(\lambda_i, \epsilon_j) - L_{\mathbf{y}}(\lambda_i, \epsilon_j)\right)^2 \tag{14}$$

where $\mathbf{y}$ is the truth TUD vector and $\hat{\mathbf{y}}$ is the reconstructed vector. K is the number of spectral channels, $L_{\hat{\mathbf{y}}}(\lambda_i, \epsilon_j)$ and $L_{\mathbf{y}}(\lambda_i, \epsilon_j)$ are the at-sensor radiance values for a grey-body emissivity ϵ_j. A linear sampling of M grey-body emissivity values between 0 and 1 are used to calculate loss, improving reconstruction error for reflective and emissive materials. The hyperparameter γ is a regularization term controlling the relative importance between the TUD MSE and the at-sensor radiance MSE within the loss function.

Atmospheric state error is minimized using a weighted MSE loss function described by:

$$\mathscr{L}_A(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{3p}\sum_{i=1}^{3p} w_i(\hat{y}_i - y_i)^2 \tag{15}$$

where the weights $\mathbf{w} \in \mathbb{R}^{1\times 3p}$ are derived from the atmospheric pressure levels leading to the largest deviation in at-sensor radiance. To identify these pressure level dependent deviations, a Jacobian matrix is calculated between at-sensor radiance and each measurement vector. Each pressure level measurement is modified by 1% of the training data mean value resulting in the Jacobian matrix described in Equation 16:

$$\mathbf{J}_{\mathbf{L}}(\mathbf{M}) = \begin{bmatrix} \frac{\partial L(\lambda_1)}{\partial M(a_1)} & \cdots & \frac{\partial L(\lambda_K)}{\partial M(a_1)} \\ \vdots & \ddots & \vdots \\ \frac{\partial L(\lambda_1)}{\partial M(a_p)} & \cdots & \frac{\partial L(\lambda_K)}{\partial M(a_p)} \end{bmatrix} \tag{16}$$

where $\mathbf{M}$ represents the particular measurement $(\mathrm{T}, \mathrm{H_2O}, \mathrm{O_3})$. This calculation is performed using the imager bandcenters, specifically 92 bands between $8.13\,\mu\text{m}$ - $12.48\,\mu\text{m}$. The mean absolute change in at-sensor radiance across all bands for a particular pressure level, p, and measurement vector M is calculated according to:

$$w_p^{\mathrm{M}} = \frac{1}{K}\sum_{i=1}^{K}|\mathbf{J}_{L_i}(\mathbf{M}_p)| \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

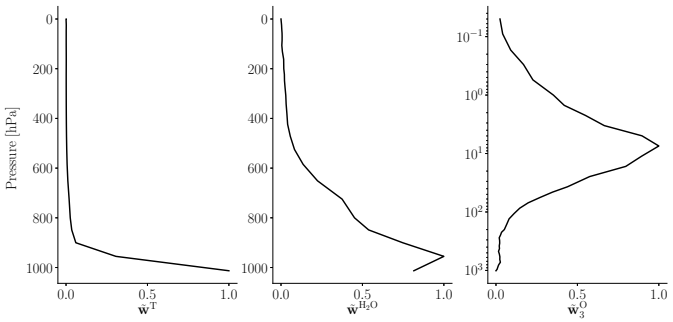

Fig. 3. The atmospheric state weighted MSE loss function utilizes the concatenated weight vectors shown. These weights allow the model to accurately predict atmospheric measurements that have the largest impact on the generated TUD vector. Both temperature and water vapor content must be reconstructed correctly at low altitudes (high pressure levels), while ozone concentration has the largest impact at high altitudes.

Next, $\mathbf{w}^M$ is normalized between 0 and 1 across p pressure levels to form $\tilde{\mathbf{w}}^M$. Each normalized measurement weight vector is concatenated to create $\mathbf{w}$ in Equation 15 such that $\mathbf{w} = [\tilde{\mathbf{w}}^T, \tilde{\mathbf{w}}^{\mathrm{H_2O}}, \tilde{\mathbf{w}}^{\mathrm{O_3}}], \mathbf{w} \in \mathbb{R}^{1\times 3p}$. Multiple atmospheric state vectors were selected from each air mass in the TIGR data to form multiple estimates of $\mathbf{w}$. The average of these estimates are shown in Figure 3 agreeing with typical concentration variation of water vapor and ozone at the altitudes shown. Similarly, temperature profiles can often be fit using only surface temperature and lapse rate [13]. The weight $\tilde{\mathbf{w}}^T$ captures this behavior by emphasizing only the measurements closest to the surface.

The total MMAE network loss is calculated by combining each mode loss and the latent space KL loss:

$$\begin{aligned}\mathscr{L}(\hat{\mathbf{y}}_A, \mathbf{y}_A, \hat{\mathbf{y}}_T, \mathbf{y}_T) = \\ \mathscr{L}_A(\hat{\mathbf{y}}_A, \mathbf{y}_A) + \mathscr{L}_T(\hat{\mathbf{y}}_T, \mathbf{y}_T) \\ + \beta \mathscr{L}_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{y}_A, \mathbf{y}_T) \parallel p(\mathbf{z})\big)\end{aligned} \tag{18}$$

where β is used to trade off reconstruction accuracy against enforcing the prior distribution. The inclusion of β is based on [27] where interpretable latent space components can be recovered if the data generating processes are understood. This research leverages this modification to create an interpretable latent space, capturing variables such as atmospheric water vapor content and atmospheric temperature, allowing new samples to be generated with known properties.

Each layer in the MMAE performs a transform with the function $y = f(\mathbf{w}\mathbf{x} + \mathbf{b})$ where $f(\cdot)$ is the activation function, $\mathbf{w}$ is the layer weight matrix and $\mathbf{b}$ is the layer bias vector. The MMAE implemented here utilizes the exponential linear unit (ELU) activation function:

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if x} > 0 \\ \alpha(\exp(x) - 1), & \text{if x} \leq 0 \end{cases}$$

The activation for predicting $\boldsymbol{\mu}$ is linear and the activation for predicting $\boldsymbol{\sigma}$ is $\mathrm{ELU}(x) + 1$ to guarantee positive variances. Additionally, each mode's output layer utilizes a linear activation function.

1) Generative Model Metrics: Evaluating the MMAE performance on hold out samples is necessary to determine if the model has generalized to the underlying relationships in the data or over fit to the training samples. The hold out samples considered here consist of TUD vectors and atmospheric state vectors never encountered in the training data. Additionally, the validation sensor altitudes were never observed in the training set. To measure hold out sample performance with respect to at-sensor radiance error, a range of grey-body emissivity values, ϵ, with an assumed pixel temperature of 300 K are used to create simulated at-sensor radiance spectra, $L(\lambda, \epsilon)$. Since this study is focused on the LWIR domain, spectral radiance values were converted to brightness temperature, $T_{BB}(\lambda, \epsilon)$:

$$T_{BB}(\lambda, \epsilon) = \frac{hc}{\lambda k \ln\left(\frac{2hc^2}{\lambda^5 L(\lambda,\epsilon)} + 1\right)}. \tag{19}$$

Using $\mathbf{y}_T$ and $\hat{\mathbf{y}}_T$ to create $L(\lambda, \epsilon)$ and $\hat{L}(\lambda, \epsilon)$ respectively, the root mean square error (RMSE) in degrees Kelvin can be calculated with:

$$E(\epsilon) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\Big(T_{BB}(\lambda_i, \epsilon) - \hat{T}_{BB}(\lambda_i, \epsilon)\Big)^2} \tag{20}$$

The grey body emissivity is varied from 0 to 1 producing an RMSE curve describing overall performance between reflective and emissive materials. Additionally, MODTRAN [11] can be used to convert $\hat{\mathbf{y}}_A$ to a TUD vector, resulting in the same error metric for the atmospheric state prediction. When multiple models are compared at once, the brightness temperature RMSE area under the curve (AUC-BT) is reported to capture reflective to emissive performance with a single scalar value:

$$\text{AUC-BT} = \int_{0.0}^{1.0} E(\epsilon) d\epsilon \tag{21}$$

Since the AUC-BT metric measures RMSE across reflective to emissive materials, lower values represent better reconstruction performance with perfect reconstruction represented by AUC-BT $= 0$.

B. Set Attention for In-Scene Atmospheric Compensation

The MDAC model utilizes the MMAE decoder model to predict $\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_T$, from a set of pixels, $\mathbf{X}$. This set-input learning has been investigated in domains such as point cloud classification where a set of points correspond to a single target value or class label [6], [28], [29]. An important characteristic of methods solving set-input learning problems is permutation-invariance to the points in the set. Regardless of pixel selection order, the MDAC algorithm must still provide the same TUD and atmospheric state prediction.

Permutation-invariant predictions are made by the MDAC network using two operations: set transformation and set decomposition. In this study, the set transformation operation is a neural network consisting of an input transform and feature transform as shown in Figure 4. The input transform consists of a K node layer to transform each pixel identically, followed by a set centering operation. The weights in the K node layer are shared across all pixels, maintaining permutation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

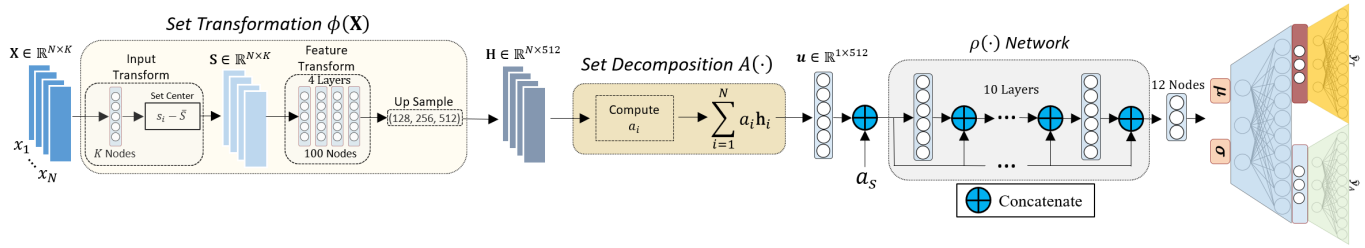

Fig. 4. The MDAC network consists of a set transformation, set decomposition and a network $\rho(\cdot)$ for predicting the MMAE latent components. The set transformation converts the input set $\mathbf{X}$ to the set $\mathbf{H}$ using the input transform and feature transform shown. The set $\mathbf{H}$ is converted into the set representation vector $\mathbf{u}$ with the attention pooling operation $A(\cdot)$. Sensor altitude, a_s, is concatenated to $\mathbf{u}$ before entering the $\rho(\cdot)$ network.

invariance. The feature transform utilizes 4 layers each with 100 nodes, again sharing weights across all pixels. The set transformation concludes with pixel representation upsampling to create the set $\mathbf{H}$:

$$\mathbf{H} = \phi(\mathbf{X}), \quad \mathbf{H} \in \mathbb{R}^{N\times M} \tag{22}$$

where $M = 512$ from the upsampling layer. The rows of $\mathbf{H}$ correspond to transformed pixel representations $\mathbf{h}_i$ which must be pooled together by the set decomposition operation. To understand the role each pixel plays in the overall model prediction, this study investigates set attention pooling [5]:

$$\mathbf{u} = \sum_{i=1}^{N} a_i \mathbf{h}_i, \quad \mathbf{u} \in \mathbb{R}^{1\times M} \tag{23}$$

where $\mathbf{u}$ is the set representation vector and a_i is the attention score for pixel i calculated according to:

$$a_i = \frac{\exp\left(\mathbf{w}^T\left(\tanh\left(\mathbf{V}\mathbf{h}_i^T\right) \odot \text{sigm}\left(\mathbf{U}\mathbf{h}_i^T\right)\right)\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{w}^T\left(\tanh\left(\mathbf{V}\mathbf{h}_j^T\right) \odot \text{sigm}\left(\mathbf{U}\mathbf{h}_j^T\right)\right)\right)} \tag{24}$$

The trainable parameters are $\mathbf{w} \in \mathbb{R}^{1\times L}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$, where L corresponds to the attention pooling dimension. The value of L is varied as part of the overall network hyperparameter sweep with the results in this study using $L = 512$. In Equation 24, $\tanh(\cdot)$ corresponds to the hyperbolic tangent function, $\text{sigm}(\cdot)$ is the sigmoid function and $\odot$ is an element-wise product. The set pixel representations are initially transformed by matrix $\mathbf{V}$ which is learned through the training process. The $\tanh(\cdot)$ operation is approximately linear between -1 and 1 and so the $\text{sigm}(\cdot)$ function is used as a gating function to model more complex dependencies [5], [30]. The matrix $\mathbf{U}$ controls the gating mechanism and is also learned through the training process. The vector $\mathbf{w}$ converts the pixel representation into a scalar value that is used in the overall softmax function to create the attention weights, a_i, which sum to 1.

The set representation vector $\mathbf{u}$ captures information necessary to predict $\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_T$, however, to create a multi-altitude model the sensor altitude a_s is concatenated to $\mathbf{u}$. This concatenated vector forms the input to the $\rho(\cdot)$ network, which predicts the low dimensional components of the MMAE model, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$. The $\rho(\cdot)$ network consists of 10 layers each with 100 nodes utilizing skip connections to propagate the set representation vector to deeper layers as shown in Figure 4. Similar to the MMAE model, the $\rho(\cdot)$ output layer utilizes a linear activation for predicting $\boldsymbol{\mu}$ and $\text{ELU}(x)+1$ for predicting $\boldsymbol{\sigma}$. The output layer has 12 nodes because the first 6 outputs are for $\hat{\boldsymbol{\mu}}$ and the last 6 are for $\hat{\boldsymbol{\sigma}}$. Denoting the attention weighted sum in Equation 23 as A, the MDAC network can be specified as:

$$D_m(\mathbf{X}, a_s) = \rho\big(A(\phi(\mathbf{X})), a_s\big). \tag{25}$$

The network configuration shown in Figure 4 was the result of a hyperparameter sweep over possible set transformation networks, $\rho(\cdot)$ networks and the number of attention nodes in the set decomposition. Additionally, batch size, learning rate, and activation functions were varied in the hyperparameter sweep. The results presented here utilize a learning rate of 1×10^{-3} and a batch size of 512. The number of pixels in each training set was $N = 50$ and so for a single batch, 512 sets were presented to the network (25,600 pixels). The Adam optimization algorithm was used for calculating weight updates [24]. Networks were constructed using Python 3.6.8, Keras version 2.2.4, Tensorflow 1.15 and hyperparameter sweeps were conducted across 20 graphics processing units (GPUs) using Ray Tune version 0.7.6 [31] [32].

C. Algorithm Training

The MDAC algorithm is trained using sets of at-sensor radiance data $\mathbf{X}$ created from an underlying TUD vector and atmospheric state vector. The same TUD and atmospheric state data are used to fit MMAE and MDAC models. Training the MDAC algorithm follows the strategy outlined in [4], with the exception that MDAC has multiple outputs requiring additional loss calculations. Emissivity profiles are sampled from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) database with 200 emissivity samples held out for model validation and 978 different material profiles used during training. Emissivity selection and pixel temperature assignment follows the set generation algorithm outlined in [4]. During training, the at-sensor radiance set $\mathbf{X}$ contains $N = 50$ pixels resulting in $\binom{978}{50} = 3 \times 10^{84}$ possible training emissivity sets.

Only the MDAC weights are updated during training, leaving the MMAE weights unchanged. The MDAC weights are updated based on the $\mathbf{y}_A$ and $\mathbf{y}_T$ error using the loss functions $\mathscr{L}_A$ and $\mathscr{L}_T$, respectively. The same atmospheric weights, w_i, are again used to calculate the loss on $\mathbf{y}_A$ reinforcing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

atmospheric state reconstruction at pressure levels impacting the predicted TUD vector.

D. Pixel Selection

Accurate MDAC prediction is predicated on access to a set of diverse pixels with respect to emissivity and temperature. To select N diverse pixels from a collected data cube, this study follows the pixel selection strategy outlined in [4] where the spectral angle, θ_i, between pixel i and the cube mean, $\bar{L}(\lambda)$, is calculated according to:

$$\theta_i = \cos^{-1}\left(\frac{L_i(\lambda)\cdot\bar{L}(\lambda)}{\|L_i(\lambda)\|\|\bar{L}(\lambda)\|}\right) \tag{26}$$

An iterative pixel selection strategy is employed starting with the 90th percentile pixel with respect to sorted cube spectral angles. A one pixel guard band is applied spatially removing all neighboring pixels from being included in the set $\mathbf{X}$. A uniform sampling of the 10% highest spectral angles is conducted following this procedure resulting in N diverse pixels with respect to the cube mean. Prior to pixel selection, anomalous pixels such as those from dead pixels, are removed from the sorting process. These noisy pixels may not follow the simplified radiative transfer model leveraged in this work and are eliminated from atmospheric compensation consideration.

E. Target Detection Analysis

After sampling a collected data cube using the method presented in Equation 26, the MDAC predictions can be used to compensate a data cube and perform target detection. The target detection method used in this study is the Adaptive Coherence/Cosine Estimator (ACE) detector defined by [33]:

$$r_{ACE}(\mathbf{x}) = \frac{(\mathbf{s}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x})^2}{(\mathbf{s}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{s})(\mathbf{x}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x})} \tag{27}$$

where $\mathbf{x}$ is a sample pixel, $\mathbf{s}$ is the target, and $\hat{\mathbf{\Sigma}}$ is the estimated background covariance. To estimate $\mathbf{\Sigma}$, a Mahalanobis anomaly detector is applied to filter background pixels from possible targets. The Mahalanobis detector can be described by:

$$r_{MD}(\mathbf{x}) = (\mathbf{x}-\hat{\boldsymbol{\mu}})^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}) \tag{28}$$

where $\hat{\boldsymbol{\mu}}$ is the cube mean and $\hat{\mathbf{\Sigma}}$ is the cube covariance. The detection statistic, $r_{MD}(\mathbf{x})$, is sorted and pixels below the 90th percentile are classified as background. These background pixels are used to form $\hat{\mathbf{\Sigma}}$ for the ACE detector. Target detection results can be compared using the Signal-to-Clutter Ratio (SCR) defined as:

$$\text{SCR} = \frac{\mu(r_t)-\mu(r_b)}{\sqrt{\sigma(r_t)^2+\sigma(r_b)^2}} \tag{29}$$

where $\mu(r_t)$ is the mean detection statistic for target pixels and $\mu(r_b)$ is the mean detection statistic for background pixels. Similarly the standard deviations of these two classes are calculated with $\sigma(\cdot)$. Large SCR values imply higher detection statistics on target pixels compared to background pixels with little variance among both classes.

IV. Results

This section first presents the MMAE results and demonstrates the model's ability to generate new atmospheric states and TUD vectors. Next, the MMAE is used as part of the overall MDAC algorithm to perform in-scene atmospheric compensation and atmospheric state estimation. Results are presented for synthetic data to demonstrate model characteristics followed by analysis on SEBASS collected data cubes spanning multiple days and sensor altitudes. Atmospheric compensation results are compared to FLAASH-IR through a target detection study.

A. Multimodal Generative Model Results

Models utilizing $\mathscr{L}_{KL}$, $\mathscr{L}_A$, $\mathscr{L}_T$ are first compared against models using MSE to demonstrate the benefit of these loss functions in minimizing model reconstruction error. The pairwise model comparisons considered for the MMAE network outputs $(\mathbf{y}_A, \mathbf{y}_T)$ respectively are: (MSE, MSE), (MSE, $\mathscr{L}_A$), ($\mathscr{L}_T$, MSE), ($\mathscr{L}_T$, $\mathscr{L}_A$). Additionally, for each model configuration, $\mathscr{L}_{KL}$ is investigated by varying β from 0.0 to 1.0. Each loss and β configuration result is based on 10 randomly initialized models to provide estimates of model mean performance.

The AUC-BT results are shown in Figure 5 for all loss configurations and considered β values. Reconstruction errors on $\mathbf{y}_T$ are reduced by using either $\mathscr{L}_A$ or $\mathscr{L}_T$ compared to MSE with the lowest reconstruction error observed when both $\mathscr{L}_A$ and $\mathscr{L}_T$ are used. The $\mathbf{y}_A$ error is not reduced for the ($\mathscr{L}_T$, MSE) case compared to the baseline MSE model. This is driven by the observation that similar TUD vectors can be created from significantly different atmospheric state vectors. While atmospheric state to TUD vectors is a one-to-one function, TUD vectors to atmospheric state is not.

Figure 5 also highlights the role KL divergence plays in reconstruction accuracy. Increased reconstruction error is observed when $\beta > 10^{-2}$ because the latent components are over-constrained reducing modeling capacity. From Figure 5, it is not clear which β value should be selected or if KL divergence should even be used since $\beta = 0$ has comparable reconstruction error. Next, latent space continuity is evaluated, an important attribute for latent space sampling.

The process to measure latent space continuity is outlined in Algorithm 1 beginning with the MMAE encoder model, e, transforming N input samples, $(\mathbf{y}_T, \mathbf{y}_A)$, to the latent code $\mathbf{z} \in \mathbb{R}^{N\times c}$ for a model with latent dimension c. The decoder model, d, reconstructs the input resulting in $\hat{\mathbf{y}}_T, \hat{\mathbf{y}}_A$. To determine latent space continuity, $\mathbf{z}$ is modified and the output deviation from $\hat{\mathbf{y}}_T$ is measured in terms of AUC-BT, denoted as ΔAUC-BT. The Algorithm 1 input $\mathbf{\Delta} \in \mathbb{R}^{N\times c}$ is the latent space deviation matrix used to modify $\mathbf{z}$. The rows of matrix $\mathbf{\Delta}$ are formed by randomly picking points on a hypersphere using [34]:

$$\mathbf{\Delta}_i = \frac{r}{\sqrt{x_1^2+x_2^2+\cdots+x_n^2}}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix},\ x_l \sim \mathrm{N}(0,1) \tag{30}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

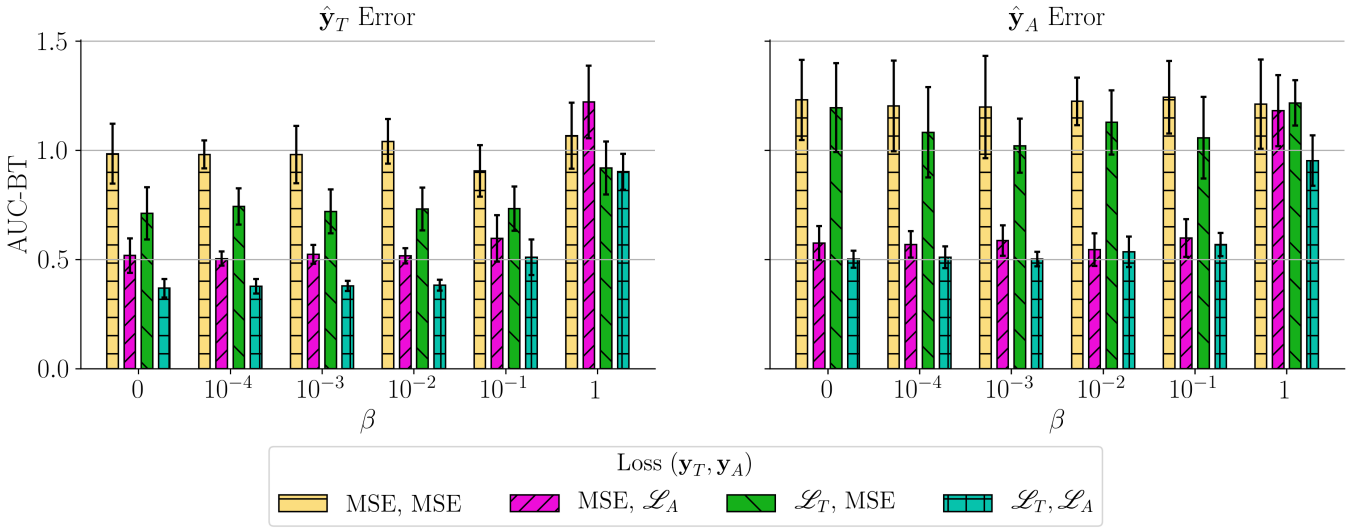

Fig. 5. Loss configuration results are shown using the AUC-BT reconstruction error (lower is better) where the best performance is achieved when both $\mathscr{L}_T$ and $\mathscr{L}_T$ are used. As β is increased beyond 0.01, reconstruction error increases because the latent space is overconstrained and no longer has adequate capacity to capture data variability.

Algorithm 1 Latent Space Variation

Input: e, d, $\mathbf{y}_T$, $\mathbf{y}_A$, $\mathbf{\Delta}$, ϵ
Output: ΔAUC-BT

Modify latent components :

$\mathbf{z} \leftarrow e(\mathbf{y}_T, \mathbf{y}_A)$
$\hat{\mathbf{y}}_T, \hat{\mathbf{y}}_A \leftarrow d(\mathbf{z})$
$\mathbf{\Sigma} \leftarrow \mathrm{E}[(\mathbf{z} - \mathrm{E}[\mathbf{z}])(\mathbf{z} - \mathrm{E}[\mathbf{z}])^T]$
$\mathbf{U}, \mathbf{\Lambda} \leftarrow$ s.t. $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$
$\tilde{\mathbf{z}} = (\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{z})$
$\tilde{\mathbf{z}}_\Delta = \tilde{\mathbf{z}} + \mathbf{\Delta}$
$\mathbf{z}' = \mathbf{U}\mathbf{\Lambda}^{1/2}\tilde{\mathbf{z}}_\Delta$

Measure output deviation

$\mathbf{y}'_T, \mathbf{y}'_A \leftarrow d(\mathbf{z}')$
$E(\hat{\mathbf{y}}_T, \mathbf{y}'_T, \epsilon) \leftarrow \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(T_{BB}(\lambda_i, \epsilon) - \hat{T}_{BB}(\lambda_i, \epsilon)\right)^2}$
$\Delta\text{AUC-BT} \leftarrow \int_{0.0}^{1.0} E(\hat{\mathbf{y}}_T, \mathbf{y}'_T, \epsilon) d\epsilon$
return ΔAUC-BT

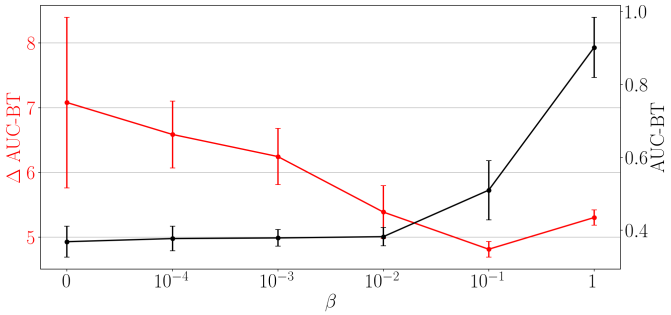

Fig. 6. Making small changes to latent components and measuring the change in the $\hat{\mathbf{y}}_T$ is plotted on the left axis. The model validation performance is shown on the right axis, also shown in Figure 5 for the $(\mathscr{L}_T, \mathscr{L}_A)$ configuration. Increasing β results in a more continuous latent space as shown by the decreasing ΔAUC-BT values. However, increasing β beyond 10^{-2} over-constrains the latent space resulting in poor validation performance (right axis). By selecting $\beta = 10^{-2}$, the MMAE has both a continuous latent space and low reconstruction error.

where $\|\mathbf{\Delta}_i\| = r$. To make comparable changes to each β model latent space $\mathbf{z}$, Algorithm 1 applies PCA whitening to $\mathbf{z}$ resulting in $\tilde{\mathbf{z}}$. After adding $\mathbf{\Delta}$ to $\tilde{\mathbf{z}}$, the whitening process is reversed and the decoder transforms the new latent samples to $\mathbf{y}'_T$ and $\mathbf{y}'_A$. Output deviations are measured between $\hat{\mathbf{y}}_T$ and $\mathbf{y}'_T$ resulting in ΔAUC-BT. If small changes to $\mathbf{z}$ lead to large ΔAUC-BT values, sampling the latent components will be challenging as greater sampling accuracy is needed.

Applying Algorithm 1 to each β model results in the ΔAUC-BT shown in Figure 6 where smaller output deviations are observed for larger β values. The right axis of Figure 6 shows the validation reconstruction error for the $(\mathscr{L}_T, \mathscr{L}_A)$ loss configuration from Figure 5. When $\beta > 10^{-2}$, KL divergence loss begins to negatively affect reconstruction error as the latent space is over-constrained. In this research, $\beta = 10^{-2}$ is selected to trade off a continuous latent space and low reconstruction error.

Many generative model studies have investigated latent space attribute vectors allowing for new samples to be generated with certain properties such as images of faces wearing sunglasses or smiling [35], [36]. Varying the MMAE latent space components reveals analogous attribute vectors allowing atmospheric state conditions to be precisely controlled. One latent component is varied from -1.0 to 1.0 (the domain of this component) while all other components are unchanged resulting in the atmospheric measurements and TUD vectors shown in the Appendix. The predicted atmospheric measurements show significant changes in the total water vapor content and ozone content as a single component is varied with corresponding changes in the predicted TUD output. Sampling additional points in this region of the data manifold is useful for a range of applications such as radiative transfer modeling and data augmentation. Next, the joint, low-dimensional representation created by the MMAE will be used for in-scene

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

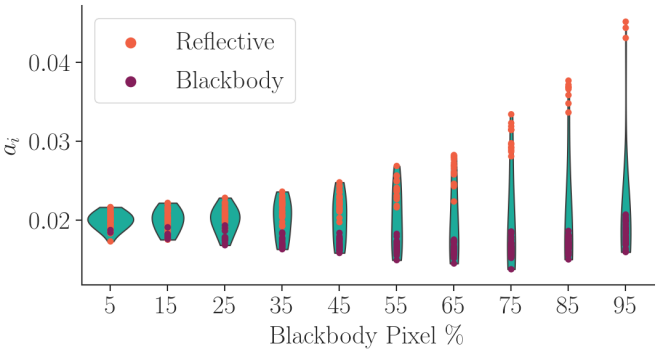

Fig. 7. At-sensor radiance sets were created with an increasing percentage of blackbody pixels. The attention scores for reflective scenes (low blackbody pixel %) are small and clustered together while scenes containing only a few reflective pixels have larger attention scores to emphasize the importance of the reflective pixels. The violin plots show the attention score density for the 50 points displayed at each blackbody pixel percentage.

atmospheric compensation.

B. Atmospheric Compensation with Synthetic Data

Using the previously fit MMAE network, the MDAC network was trained to predict the low-dimensional representation $\mathbf{z}$ from a set of at-sensor radiance samples, $\mathbf{X}$. At-sensor radiance sets were generated based on the set generation algorithm presented in [4]. Using a batch size of 512 and set size of $N = 50$, training executed for 50 epochs. At the conclusion of 50 epochs, new training data was generated, with this process repeated 60 times. During each 50 epoch training iteration, error was gradually reduced as the model fit to the new data. We found that 60 iterations of this training process resulted in stable errors, even when the model was presented new at-sensor radiance sets.

The MDAC network relies on attention pooling to convert the pixel set $\mathbf{X}$ into the set representation vector, $\mathbf{u}$. The attention weights, a_i, represent the importance of each pixel in forming the set representation. To evaluate data characteristics the attention pooling operation has learned, at-sensor radiance sets were generated with varying blackbody pixel percentage within the scene. These synthetic scenes were used to evaluate the attention weights with the results shown in Figure 7. The violin plots in Figure 7 represent the attention score density as there are 50 points displayed for each blackbody pixel percentage.

Reflective material dominated scenes (low blackbody percentage), result in tightly clustered, low attention scores because multiple reflective pixels contain information necessary for recovering the scene TUD vector. As the generated scenes change from reflective material dominated to emissive material dominated (large blackbody percentage), the overall attention score increases. The remaining reflective pixels are important for downwelling radiance estimation and receive a larger attention score. This observation is supported by the LWIR radiative transfer equation where downwelling radiance can only be estimated if reflective materials are present. This dependence on reflective pixels is an important characteristic of the MDAC model, specifically when applying the model to globally diverse data.

TABLE I
ALTITUDE, COLLECTION TIME, WEATHER AND COLLECTION DAY ARE REPORTED FOR THE 3 DATA CUBES INVESTIGATED.

	Cube 1	Cube 2	Cube 3
Altitude (km)	0.45	0.92	1.22
Time (UTC)	1856	1819	1638
Weather	Clear	Clear	Clouds
Collection Day	1	1	6

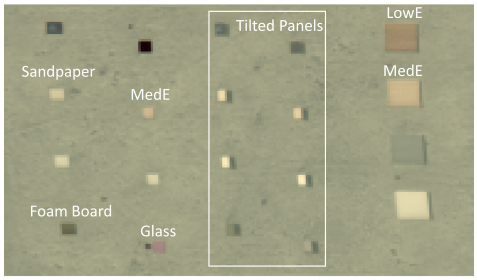

Fig. 8. The panel configuration for the 5 considered materials is shown. The tilted panel section was not considered in this study to evaluate the downwelling radiance accuracy. The unlabeled panels are other materials not considered in this work.

C. Collected HSI Data Results

This study considers three data cubes collected by the SEBASS LWIR imager at altitudes of $0.45\,\mathrm{km}$, $0.92\,\mathrm{km}$ and $1.22\,\mathrm{km}$. The first two cubes were collected on the same day and the third cube was collected 5 days later as shown in Table I. The collected data contains varying size material panels at different tilt angles, however, only flat panels within the scene are considered to evaluate downwelling radiance accuracy. The labeled materials considered are: Foam Board, Low Emissivity Panel (LowE), Glass, Medium Emissivity Panel (MedE) and Sandpaper with the configuration shown in Figure 8. The ground truth emissivity for each material was measured with a D&P spectrometer and downwelling radiance was measured using an infragold sample.

Predictions for FLAASH-IR, DAC [4], and each output of MDAC are shown in Figure 9 when applied to cube 1. The downwelling radiance provided by FLAASH-IR also contains atmospheric transmission. To compare downwelling radiance quantities, the FLAASH-IR downwelling radiance is divided by atmospheric transmission resulting in the $L_d(\lambda)$ shown in Figure 9.

It is important to note the $\mathbf{y}_A$ prediction in Figure 9 is based on the model's atmospheric state prediction (T, H_2O, O_3) converted to a TUD vector using MODTRAN. This atmospheric state prediction is shown in Figure 10, highlighting the complexities of predicting pressure level measurements. While no radiosonde data is available to directly compare the atmospheric state prediction, this atmospheric state estimate does result in a comparable TUD estimate to DAC and $\mathbf{y}_T$ using only in-scene data. Next, the TUD estimates are compared from a target detection perspective, using both TES [14] and improved alpha residuals (AR) [18].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

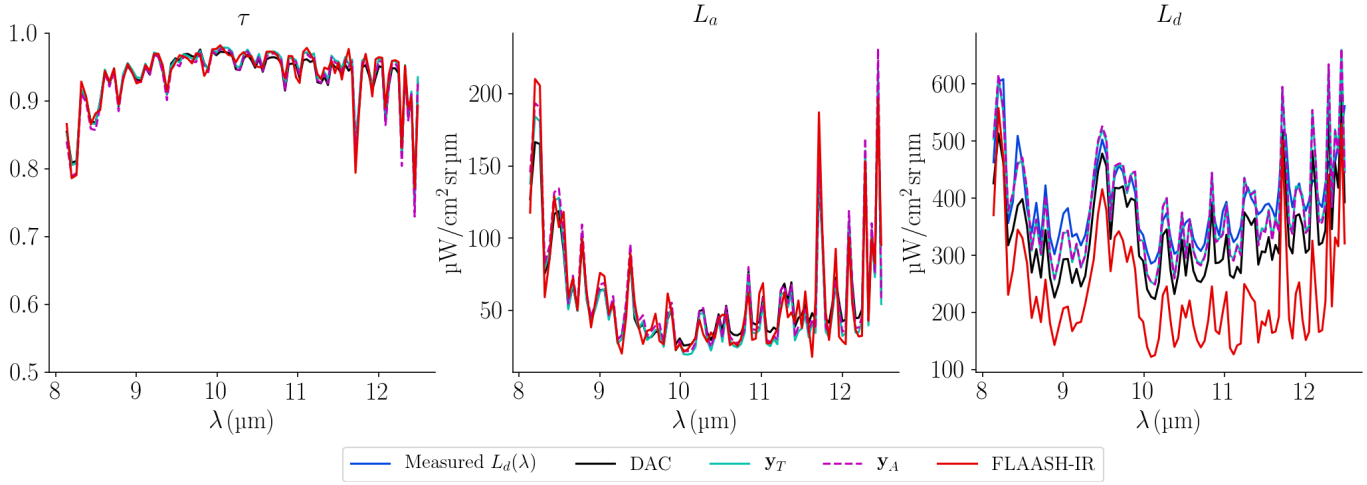

Fig. 9. Applying MDAC to collected data (cube 1 in Table I) results in the two TUD predictions $\mathbf{y}_A$ and $\mathbf{y}_T$ shown. The $\tau(\lambda)$ and $L_a(\lambda)$ estimates are comparable for all methods. As expected, the largest model discrepancy is in the downwelling estimate, which relies on the selection of reflective pixels to estimate this term.

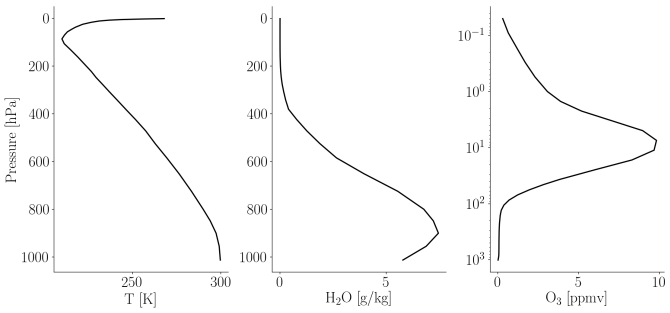

Fig. 10. The atmospheric state prediction corresponding to the $\mathbf{y}_A$ TUD prediction shown in Figure 9 is plotted showing the complexity of predicting each atmospheric level from in-scene data only. No truth data is available at the time of collect to compare against.

D. Target Detection Results

Many target detection applications require an efficient pipeline to resolve targets as quickly as possible. To support these applications, the improved AR approach outlined in [18] is used for comparing detection performance. Additionally, the commonly investigated maximum smoothness TES approach presented in [14] is also considered, however, this approach is significantly more time-consuming compared to improved AR. For all considered materials, the recovered signals are shown in Figure 11 based on the TUD predictions shown in Figure 9. Close agreement is observed between all AR results for this data cube, while the TES results contain larger biases. These biases are derived from incorrect temperature estimates made during the TES process, but the distinctive signal features are still clearly evident. Additionally, the emissivity measurements will have some spectral variability, however, we do not consider the impact of spectral variability in this study. The results presented thus far are for a single data cube. To further compare performance, two additional data cubes are considered and aggregated target detection results are reported.

For each of the three investigated data cubes, the ACE background covariance matrix, $\mathbf{\Sigma}$, was estimated using the Mahalanobis anomaly detector with a threshold of 90% to classify pixels as background or anomaly. To compare ACE detector performance, Receiver Operating Characteristic (ROC) curves are used to show the relationship between probability of detection (P_D), and probability of false alarm (P_{FA}) for varying operating points. Applying the ACE detector in AR space or emissivity space results in the average ROC curves shown in Figure 12 with comparable results observed for all methods and materials. To further illustrate this point, SCR mean and standard deviation results are shown for each material across all three cubes in Figure 13. With few exceptions, comparable SCR results are observed for all materials and methods.

Many target detection scenarios are time-sensitive, requiring an efficient data pipeline to convert measured at-sensor radiance to a detection statistic. Atmospheric compensation with MDAC takes on average $0.3\,\mathrm{s}$ including pixel selection. Combining MDAC with the improved AR approach and the Mahalanobis anomaly detector for background statistic estimation allows for target detection in $8.5\,\mathrm{s}$ using the data cubes reported in this study. Replacing MDAC with FLAASH-IR in this processing chain results in $75\,\mathrm{s}$ target detection, which may be significant for some detection applications.

V. Conclusion

This study has presented a new LWIR in-scene atmospheric compensation approach, producing both an atmospheric state vector and TUD vector from in-scene data only. The compensation approach takes advantage of a pretrained generative model that jointly maps atmospheric state vectors and TUD vectors to a low-dimensional space using LWIR radiative transfer loss, variational loss and a weighted atmospheric state loss. Sampling the generative model yields physically plausible outputs with correct dependencies between atmospheric constituents, transmission and radiance. Given a set of in-scene data, the permutation-invariant MDAC method produces low-dimensional components which map through the generative model to compensate the data cube.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

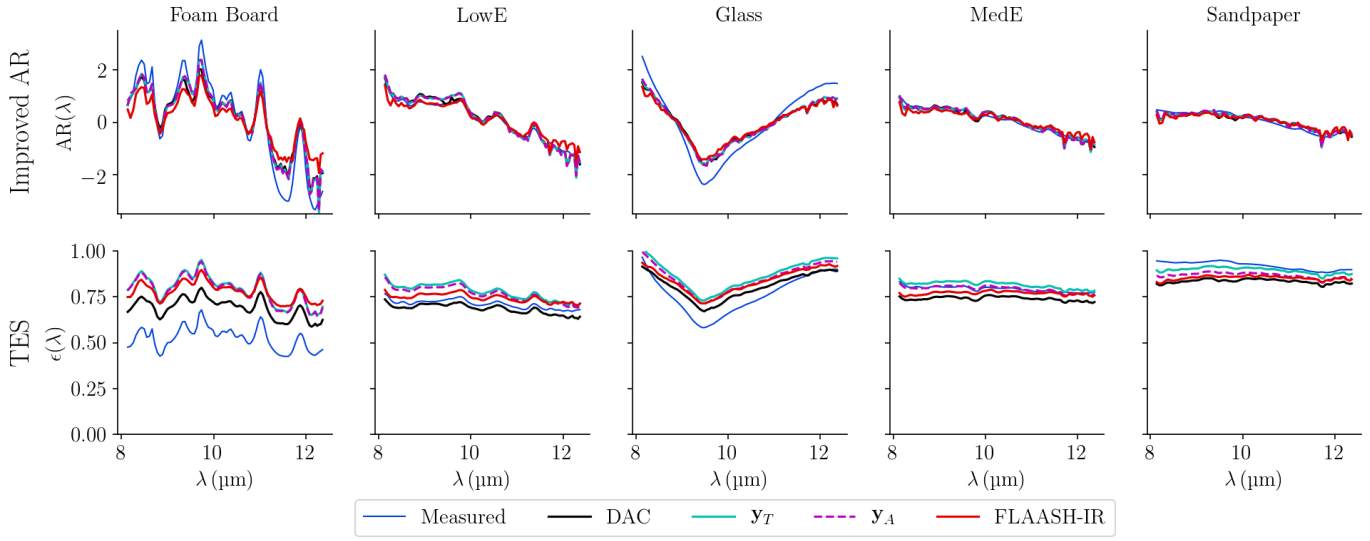

Fig. 11. Predicted alpha residual curves and emissivity profiles are shown for FLAASH-IR, DAC and the two MDAC outputs for cube 1. Alpha residual estimates were made using the improved alpha residual method discussed in [18] and the emissivity estimates were made using the maximum smoothness TES procedure from [14]. Materials are organized by increasing mean emissivity from left to right.

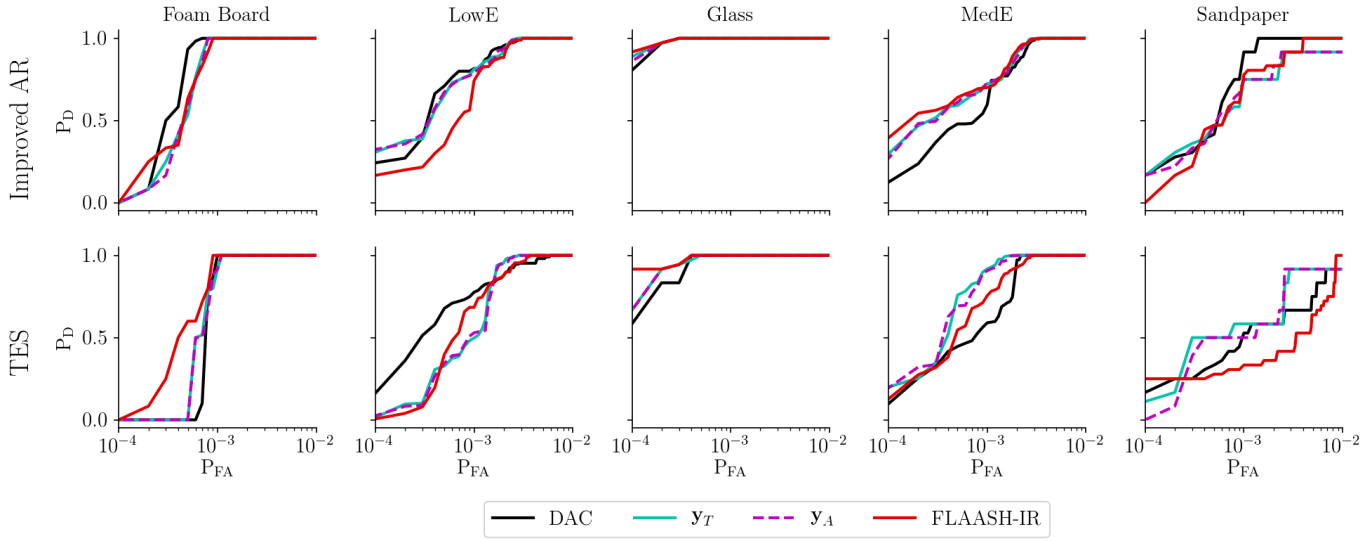

Fig. 12. Mean ROC curves are shown for DAC, each MDAC output and FLAASH-IR for all considered materials across three collected cubes. The probably of false alarm axis utilizes a logarithm scale because of the low false alarm rates for all methods and materials.

Both of the MDAC predictions were compared against FLAASH-IR and DAC on collected data cubes, demonstrating commensurate detection performance, with a significant reduction in processing time. The use of attention set pooling in the MDAC network revealed the model's use of reflective pixels, agreeing with the LWIR radiative transfer equation. This is an important model property, as fully understanding the mechanisms governing network prediction is necessary for dealing with diverse data. While not a primary goal of this study, the atmospheric state predictions of the MDAC network demonstrated that limited atmospheric sounding can be performed. The comparable detection results using the atmospheric state vector prediction suggest the model prediction was a reasonable estimate of the actual atmospheric state.

Applying this approach to higher resolution sensors is an area of future work that will identify how increased sensor resolution impacts target detection performance. Increasing sensor resolution is expected to improve the atmospheric state estimate, supporting the in-scene atmospheric sounding results presented in this study. Also, applying this atmospheric compensation method to additional data cubes is necessary to better understand how emissivity and temperature diversity affects target detection results.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

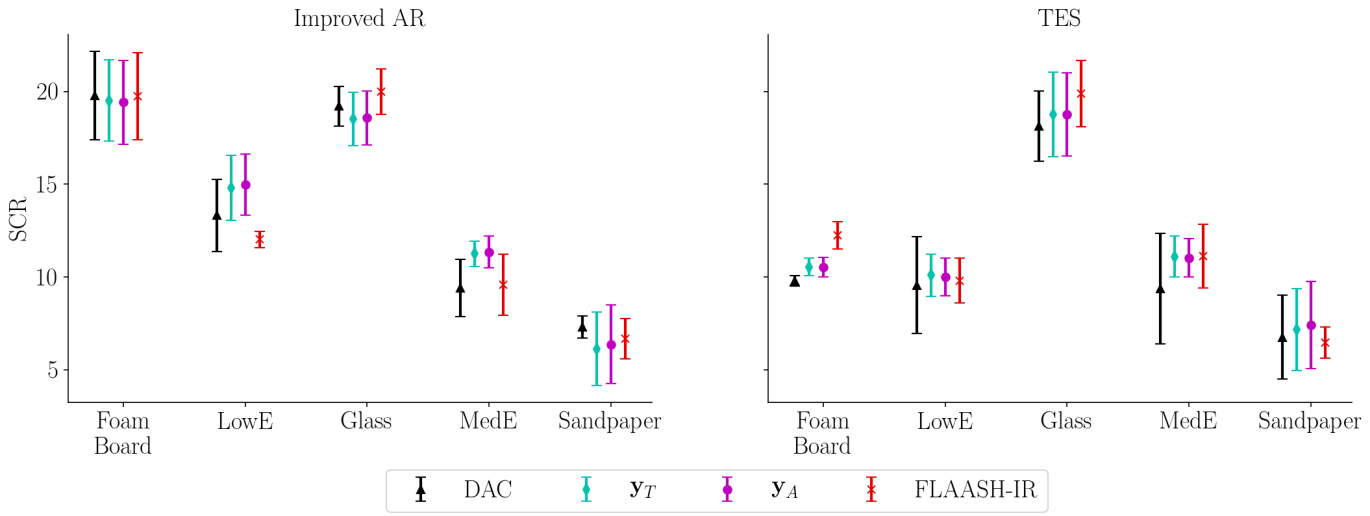

Fig. 13. Considering three collected data cubes, the SCR results are shown based on multiple atmospheric compensation approaches. Similar performance is observed for all compensation methods, however, DAC and the MDAC outputs $\mathbf{y}_A$ and $\mathbf{y}_T$ reduce the compensation time allowing for faster target detection.

REFERENCES

[1] G. Camps-Valls, J. Munoz-Mari, L. Gomez-Chova, L. Guanter, and X. Calbet, "Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1759–1769, 2011.

[2] J. A. Sobrino, R. Oltra-Carrió, J. C. Jiménez-Muñoz, Y. Julien, G. Soria, B. Franch, and C. Mattar, "Emissivity mapping over urban areas using a classification-based approach: Application to the dual-use european security IR experiment (DESIREX)," *International Journal of Applied Earth Observation and Geoinformation*, vol. 18, pp. 141–147, 2012.

[3] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proceedings of the IEEE*, vol. 97, no. 6, pp. 1031–1055, 2009.

[4] N. Westing, K. C. Gross, B. J. Borghetti, J. Martin, and J. Meola, "Learning set representations for LWIR in-scene atmospheric compensation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1438–1449, 2020.

[5] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 10–15 Jul 2018, pp. 2127–2136.

[6] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer," *arXiv preprint arXiv:1810.00825*, 2018.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[8] M. T. Eismann, *Hyperspectral Remote Sensing*. Bellingham, WA USA: SPIE, 2012.

[9] D. Manolakis, M. Pieper, E. Truslow, R. Lockwood, A. Weisner, J. Jacobson, and T. Cooley, "Longwave infrared hyperspectral imaging: Principles, progress, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 72–100, 2019.

[10] S. Adler-Golden, P. Conforti, M. Gagnon, P. Tremblay, and M. Chamberland, "Long-wave infrared surface reflectance spectra retrieved from telops hyper-cam imagery," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, vol. 9088. International Society for Optics and Photonics, 2014, p. 90880U.

[11] A. Berk, P. Conforti, R. Kennett, T. Perkins, F. Hawes, and J. van den Bosch, "MODTRAN6: a major upgrade of the MODTRAN radiative transfer code," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, vol. 9088. International Society for Optics and Photonics, 2014, p. 90880H.

[12] B. D. Bue, D. R. Thompson, M. L. Eastwood, R. O. Green, B.-C. Gao, D. Keymeulen, C. M. Sarture, A. S. Mazer, and H. H. Luong, "Real-time atmospheric correction of AVIRIS-NG imagery." *IEEE Trans. Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6419–6428, 2015.

[13] S. J. Young, B. R. Johnson, and J. A. Hackwell, "An in-scene method for atmospheric compensation of thermal hyperspectral data," *Journal of Geophysical Research: Atmospheres*, vol. 107, no. D24, 2002.

[14] C. Borel, "Error analysis for a temperature and emissivity retrieval algorithm for hyperspectral imaging data," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII*, vol. 6565. International Society for Optics and Photonics, 2007, p. 65651Q.

[15] M. L. Pieper, D. Manolakis, E. Truslow, T. W. Cooley, M. Brueggeman, J. Jacobson, and A. Weisner, "Performance limitations of temperature-emissivity separation techniques in long-wave infrared hyperspectral imaging applications," *Optical Engineering*, vol. 56, no. 8, pp. 1 – 11, 2017.

[16] P. S. Kealy and S. J. Hook, "Separating temperature and emissivity in thermal infrared multispectral scanner data: Implications for recovering land surface temperatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 6, pp. 1155–1164, 1993.

[17] A. Gillespie, "A new approach for temperature and emissivity separation," *International Journal of Remote Sensing*, vol. 21, no. 10, pp. 2127–2132, 2000.

[18] M. Diani, M. Moscadelli, and G. Corsini, "Improved alpha residuals for target detection in thermal hyperspectral imaging," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 779–783, 2018.

[19] A. Chedin, N. Scott, C. Wahiche, and P. Moulinier, "The improved initialization inversion method: A high resolution physical method for temperature retrievals from satellites of the TIROS-N series," *Journal of Climate and Applied Meteorology*, vol. 24, no. 2, pp. 128–143, 1985.

[20] F. Chevallier, F. Chéruy, N. Scott, and A. Chédin, "A neural network approach for a fast and accurate computation of a longwave radiative budget," *Journal of Applied Meteorology*, vol. 37, no. 11, pp. 1385–1397, 1998.

[21] N. M. Westing, B. J. Borghetti, and K. C. Gross, "Analysis of LWIR hyperspectral classification performance across changing scene illumination," in *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXV*. International Society for Optics and Photonics, 2019, pp. 239 – 249.

[22] J. A. Hackwell, D. W. Warren, R. P. Bongiovi, S. J. Hansel, T. L. Hayhurst, D. J. Mabry, M. G. Sivjee, and J. W. Skinner, "LWIR/MWIR imaging hyperspectral sensor for airborne and ground-based remote sensing," vol. 2819. International Society for Optics and Photonics, 1996, pp. 102 – 107. [Online]. Available: https://doi.org/10.1117/12.258057

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, p. 689–696.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[26] N. Westing, B. Borghetti, and K. C. Gross, "Fast and effective techniques for LWIR radiative transfer modeling: A dimension-reduction approach," *Remote Sensing*, vol. 11, no. 16, p. 1866, Aug 2019. [Online]. Available: http://dx.doi.org/10.3390/rs11161866
[27] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework." *ICLR*, vol. 2, no. 5, p. 6, 2017.
[28] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3391–3401.
[29] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.
[30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, p. 933–941.
[31] F. Chollet *et al.*, "Keras," https://keras.io, 2015.
[32] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.
[33] S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Transactions on signal processing*, vol. 49, no. 1, pp. 1–16, 2001.
[34] M. E. Muller, "A note on a method for generating points uniformly on n-dimensional spheres," *Commun. ACM*, vol. 2, no. 4, p. 19–20, Apr. 1959. [Online]. Available: https://doi.org/10.1145/377939.377946
[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 20–22 Jun 2016, pp. 1558–1566.

Nicholas Westing received the B.S. degree in electrical and computer engineering from the University of Minnesota-Duluth, Duluth MN, USA in 2010 and the M.S. degree in electrical engineering from the Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA in 2014.

He is currently a Ph.D. student at the Air Force Institute of Technology, investigating methods for hyperspectral data analysis and exploitation.

Brett Borghetti earned the Ph.D. degree in computer science in 2008 from the University of Minnesota, Twin Cities, MN; a M.S. degree in computer systems in 1996 from the Air Force Institute of Technology (AFIT) in Dayton, OH; and a B.S. in electrical engineering in 1992 from Worcester Polytechnic Institute (WPI), Worcester, MA.

Dr. Borghetti is an Associate Professor in the Department of Electrical and Computer Engineering in the Graduate School of Engineering Management at the Air Force Institute of Technology. His research interests focus on improving human-machine team performance in complex environments using artificial intelligence and machine learning. He has research experience in estimating human cognitive performance, statistical machine learning, genetic algorithms, self-organizing systems, neural networks, game theory, information theory and cognitive science.

Kevin C. Gross received the B.S. degree in chemistry and mathematics and the M.S. degree in chemistry both from Wright State University, Dayton, OH, USA in 1998 and 2001 respectively. He received Ph.D. degree in engineering physics from the Air Force Institute of Technology (AFIT), Wright-Patterson Air Force Base OH, USA in 2007.

He is currently the Director of EO/IR Technologies at Resonant Sciences in Dayton OH, USA and Adjunct Associate Professor at AFIT. His academic interests are remote sensing, spectroscopy, radiative transfer, and the development of physics-based algorithms to extract information content from electro-optical data.

Christine Schubert Kabban received the B.S. degree in Mathematics from the University of Dayton, M.S. in Applied Statistics from Wright State University and a PhD in Applied Mathematics from the Air Force Institute of Technology (AFIT). She started as an Assistant Professor at Virginia Commonwealth University and is currently a Professor of Statistics at AFIT. She has been researching and practicing statistics for over 20 years in clinical, engineering, and statistical fields. Her current work focuses in applications to structural health monitoring, target detection, and autonomous systems and networks with hierarchical and complex multi-dimensional data.

Jacob Martin received the B.S. degree in physics from Michigan State University, East Lansing MI, USA in 2010 and the M.S. and Ph.D. degrees in applied physics from the Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA in 2012 and 2016, respectively.

He currently works in the hyperspectral tech area with the Electo-Optic Target Detection and Surveillance Branch, Sensors Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH. His research interests include hyperspectral target detection, machine learning, sensor characterization, and performance modeling.

Joseph Meola received the B.S. and M.S. degrees in electrical engineering from the University of Dayton, Dayton OH, USA, in 2004 and 2006, respectively, and the Ph.D. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2011.

He is currently the Hyperspectral Technical Area Lead with the Electro-Optic Target Detection and Surveillance Branch, Sensors Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA, where he is also an Adjunct Faculty Member with the Air Force Institute of Technology. His research interests include hyperspectral data modeling, sensor calibration and characterization, data exploitation, target detection, and change detection.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3034421, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

APPENDIX

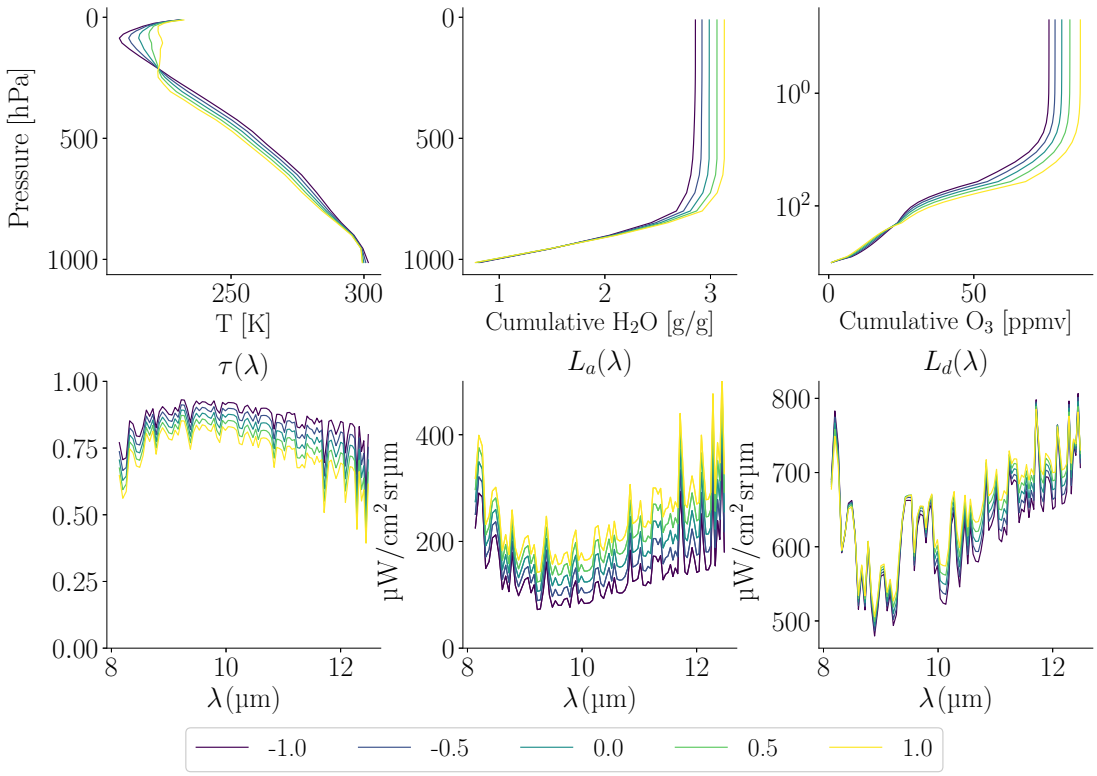

Fig. 14. Modifying one latent component from -1.0 to 1.0 results in the generated atmospheric state vectors and TUD vectors. Warping the latent space in this range allows samples to be created varying from dryer atmospheric conditions (-1.0) to more humid conditions (1.0). By increasing the total water vapor content, more radiation can be absorbed (lower transmittance) and more radiation can be emitted (higher path and downwelling radiance).