Meta-Learning and Self-Supervised Pretraining for Storm Event Imagery Translation

Ileana Rugina* *MIT EECS* irugina@mit.edu
Rumen Dangovski* *MIT EECS* rumenrd@mit.edu
Mark Veillette *MIT Lincoln Lab* mark.veillette@ll.mit.edu
Pooya Khorrami *MIT Lincoln Lab* pooya.khorrami@ll.mit.edu
Brian Cheung *MIT CSAIL & BCS* cheungb@mit.edu
Olga Simek *MIT Lincoln Lab* osimek@ll.mit.edu
Marin Soljačić *MIT Physics* soljacic@mit.edu

***Abstract*—Recent advances in deep learning have provided impressive results across a wide range of computational problems such as computer vision, natural language, or reinforcement learning. However, many of these improvements are constrained to problems with large-scale curated datasets which require a lot of human labor to gather. Additionally, these models tend to generalize poorly under both slight distributional shifts and low-data regimes. In recent years, emerging fields such as meta-learning and self-supervised learning have been closing the gap between proof-of-concept results and real-life applications of machine learning by extending deep learning to the semi-supervised and few-shot domains. We follow this line of work and explore spatiotemporal structure in a recently introduced image-to-image translation problem for storm event imagery in order to: *i)* formulate a novel multi-task few-shot image generation benchmark in the field of AI for Earth and Space Science and *ii)* explore data augmentations in contrastive pretraining for image translation downstream tasks. We present several baselines for the few-shot problem and discuss trade-offs between different approaches. Our implementation and instructions to reproduce the experiments, available at https://github.com/irugina/meta-image-translation, are thoroughly tested on MIT SuperCloud, and scalable to other state-of-the-art HPC systems.**

***Index Terms*—few-shot learning, self-supervised learning, meta-learning, generative adversarial networks**

I. Introduction

Deep learning techniques have gained popularity and demonstrated their power through benchmarks such as ImageNet [6] in computer vision and SQuAD [22] in NLP. More recently, works such as ObjectNet [3] in vision have shown impressive performance on these established benchmarks does not necessarily translate to robust performance in real-world situations, where the datasets might be less structured or more diverse. There is significant interest in devising more challenging datasets, both of general interest as well as domain-specific applications, that more closely resemble real-world situations practitioners might encounter when trying to deploy machine learning models. Growing fields such as self-supervised [18] and multi-task learning [12] reflect these interests and provide promising solutions to the aforementioned issues.

However, the problem of model evaluation remains: for example, in few-shot learning model evaluation is currently largely constrained to Omniglot [16, 15] (which has essentially been saturated), Miniimagenet [32] and Metadataset [30]. Similarly, contrastive pretraining techniques are generally evaluated on ImageNet.

We address known limitations in our field by introducing a new computer vision multi-task problem. Instead of focusing on classification problems, we turn our attention to image generation. We tackle a challenge highlighted in [9] and employ the weather dataset from [31] to define a novel few-shot image-to-image translation task. Deep learning for weather prediction has become a well-established research area [24], and recently such research yielded state-of-the-art results in the domains of weather forecasting [21, 17]. Unlike the approaches in [9, 31], we make use of the dataset's spatiotemporal structure for our few-shot tasks. In a preliminary study [26] we have investigated the applications of meta-learning [13] and self-supervised learning [2] to exploit the spatiotemporal structure. Building on this structure, we employ meta-learning and contrastive pretraining with innovative data augmentations, leading to consistent improvements in sample quality. Our research here offers three main contributions:

- we introduce a novel few-shot image translation benchmark and provide several baselines for this problem.
- we train generative adversarial networks using model-agnostic meta-learning (MAML) [8] and discuss the advantages and drawbacks of this approach.
- we pretrain part of the generator parameters using contrastive learning and show consistent improvements in downstream image-generation performance.
- we explore methods of improving forecasting [28] and contribute to the challange posed in [9].

II. Background and Related Work

A. Storm Event Imagery

The Storm Event Imagery (SEVIR) [31] is a radar and satellite meteorology dataset (Figure 1). The dataset comprises over 10,000 weather events, each tracking 5 sensor modalities within 384 km $\times$ 384 km patches for 4 hours. The events are uniformly sampled so that there are 49 frames for each 4 hour period, and the 5 channels consist of: *i)* 1 visible

* Equal contribution.

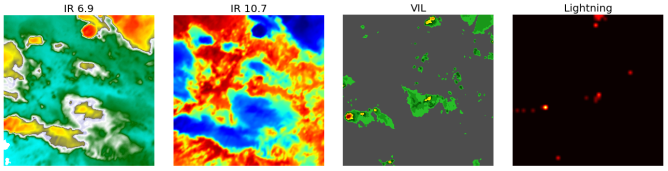

Fig. 1. **Frame from The Storm Event Imagery (SEVIR) dataset.** We use four of the five available modalities: 2 IR, VIL, and lightning information.

and 2 IR sensors from the GOES-16 advanced baseline [27], *ii)* vertically integrated liquid (VIL) from NEXTRAD *iii)* lightning flashes from GOES-16 . Figure 1 shows examples of the two IR and the VIL modalities. We disregard the visible channel because it often contains no information as visible radiation is easily occluded. Veillette et al. [31] suggested several machine learning problems that can be studied on SEVIR and provided baselines for two of these: nowcasting and synthetic weather radar generation. In both cases they train U-Net models to predict VIL information and experiment with various loss functions.

a) Evaluation: We review common evaluation metrics used in the satellite and radar literature to analyse artificially-generated VIL imagery. They all compare the target and generated samples after binarizing them with an arbitrary threshold in $[0, 255]$ and look at counts in the associated confusion matrix. Let H denote the number of true positives, M denote the number of false negatives and F the number of false positives. Veillette et al. [31] define four evaluation metrics: Critical Success Index **(CSI)** is equivalent to the intersection over union $\frac{H}{H+M+F}$; Probability of detection **(POD)** is equivalent to recall $\frac{H}{H+M}$; Success Ratio **(SUCR)** is equivalent to precision $\frac{H}{H+F}$.

B. Generative Adversarial Networks in Low Data Regimes

There is significant interest in training GANs in low-data settings. In such scenarios, a key challenge arises: the discriminator network can easily memorize the training set and achieve perfect performance on training examples[35]. Consequently, training becomes unstable, and the generator fails to produce realistic samples. Additionally, the discriminator performs poorly when evaluated on held-out validation or test splits.

Clouâtre and Demers [5], Sridhar [29] also look at few-shot multi-task image generation using second-order gradient updates. Clouâtre and Demers [5] optimize using Reptile [19], a first-order approximation to MAML, and evaluate on the MNIST and Omniglot datasets. They also introduce a dataset which presents a very clear delimitation between different tasks and more generally does not exhibit the challenges of modeling real-world phenomena because the examples are icons rather than realistic images. Sridhar [29] analyze both MAML and Reptile and perform experiments on MNIST and SVHN datasets datasets. They propose alternate algorithms for applying MAML gradient updates during adversarial training.

Zhao et al. [35] apply augmentations to both real and generated samples. They require differentiable transformations in order to backpropagate to the generator, and obtain good results using as few as 10% of the available samples. Consistency Regularization (CR) is a semi-supervised training technique introduced to GANs by [34]. Designed as a discriminator regularization technique, it can be paired with spectral normalization methods. The goal is to ensure the discriminator's predictions remain consistent, even when arbitrary transformations are applied to real samples. Zhao et al. [36] extend this work to balanced Consistency Regularization (bCR) and latent Consistency Regularization (zCR), and combine the two into Improved Consistency Regularization (ICR). These techniques regularize the discriminator and generator network using data augmentations on the generated samples or latent variables.

Wang et al. [33] leverage pretraining and adversarial training to improve diffusion models's performance on image-to-image translation [11].

C. Machine Learning for Forecasting

We focus on building upon the work in [31] due to its associated extensive public dataset. However, there has been wider interest in applying deep learning to improve the efficiency and performance of weather nowcasting systems. Ravuri et al. [23] introduced deep generative models to nowcast radar data over 90 minutes, training them to minimize three loss functions: one temporal and one spatial discriminator terms to ensure spatiotemporal consistency, as well as a grid cell regularization term that improves performance by penalizing errors at the grid cell resolution level. Conversely, we delve into various approaches to achieve similar outcomes, promoting spatiotemporal consistency through explicit task construction or contrastive pretraining.

III. Few-Shot Benchmark and Our Baselines

A. Benchmark Construction

Utilizing the SEVIR [31] dataset, we have constructed a few-shot multi-task image-to-image translation problem, with each task corresponding to a single event. From the 49 available frames we keep the first N_{support} frames to form the task's support set and the next N_{query} to be the query. Throughout the following experiments we set $N_{\text{support}} = N_{\text{query}} = 10$.

Assuming we've rescaled all input modalities to the maximum observed resolution of 384 × 384, we can perceive SEVIR as a straightforward input tensor $\mathcal{D}_1 \in \mathbb{R}^{N_{\text{event}} \times N_{\text{frames}} \times C \times w \times h}$, where: *i)* $N_{\text{event}} = 11479$; *ii)* $N_{\text{frames}} = N_{\text{support}} + N_{\text{query}}$; *iii)* $C = 4$; *iv)* $w = h = 384$. The four input channels are split into three input modalities $C_{\text{in}} = 3$ and one target $C_{\text{out}} = 1$. For joint training we ignore the hierarchical dataset structure and collapse the first two axis $\mathcal{D}_2 \in \mathbb{R}^{N \times C \times w \times h}$, where $N = N_{\text{event}} \times N_{\text{frames}}$ — the total number of frames.

B. Methods

We solve the aforementioned task using either first-order or second-order gradient descent methods on U-Nets trained using either reconstruction or adversarial objectives. Note that in the case when we train GANs using MAML we are

searching for a good initialization for multiple related saddle-point problems. Despite this challenging task, we still obtain good performance.

Next, we introduce the meta-train loop for adversarial networks, which is a novel contribution of our work. For simplicity, we only present the variant with a single SGD inner-loop adaptation step. We train a U-Net generator G with model weights w_G jointly with an extranous patch discriminator D with model weights w_D using data $\mathcal{D} \in \mathbb{R}^{N_{\text{event}} \times N_{\text{frames}} \times C \times w \times h}$. We use batched alternating gradient descent as our optimization algorithm and consider batches $\mathcal{B} \in \mathbb{R}^{B \times N_{\text{frames}} \times C \times w \times h}$, where B is the meta-batch size. Each of these can be split along the second axis into the support and query sets, and along the third axis into the source (S) and target tensors (T) to create $S^{\text{support}} \in \mathbb{R}^{B \times N_{\text{support}} \times C_{\text{in}} \times w \times h}$, $S^{\text{query}} \in \mathbb{R}^{B \times N_{\text{query}} \times C_{\text{in}} \times w \times h}$, $T^{\text{support}} \in \mathbb{R}^{B \times N_{\text{support}} \times C_{\text{out}} \times w \times h}$, $T^{\text{query}} \in \mathbb{R}^{B \times N_{\text{query}} \times C_{\text{out}} \times w \times h}$. For any of these tensors $X \in \{S^{\text{support}},\ S^{\text{query}},\ T^{\text{support}},\ T^{\text{query}}\}$ we refer to the four-dimensional tensor given by the i^{th} task or event as X_i. We use such four-dimensional tensor quantities to evaluate the generator and discriminator loss functions:

$$\hat{\mathcal{L}}_G(t^{\text{generated}}, t, s; w_G, w_D) = - \log D(s, t^{\text{generated}}) + \lambda ||t^{\text{generated}} - t||_1 \quad (1)$$

$$\hat{\mathcal{L}}_D(t^{\text{generated}}, t, s; w_G, w_D) = \frac{\log D(s, t^{\text{generated}}) - \log D(s, t)}{2}, \quad (2)$$

where $t^{\text{generated}} = G(s)$ is a generated target sample, t and s are corresponding input and output modalities, $||x||_1$ is the mean absolute error. Note the slight abuse of notation where by $\log D(x, y)$ with $x, y \in \mathbb{R}^{N \times C \times w \times h}$ we mean the average $\frac{1}{N} \sum_{i=1}^{N} \log D(x_i, y_i)$. This formulation also uses the trick of replacing $\max \log (1 - D(G(z)))$ with $\min \log D(G(z))$ to obtain a non-saturating generator objective. We wrote the loss functions above such that both players want to minimize their respective objectives.

For each task within the meta-batch size, we evaluate the losses on the support set frames. Using SGD, we adapt to this event, obtaining parameters ϕ. Following this, we evaluate the same losses on the task's query set with finetuned models. This process is repeated for every event in the meta-batch. Subsequently, a second-order gradient update is applied to the initial parameters to optimize the average loss across all events in the meta-batch. A schematic summary of this procedure can be found in Algorithm 1. The procedure for the reconstruction loss only needs minor alterations from Algorithm 1: we eliminate all lines associated with the discriminator D and adjust Equation 1 by excluding the first term for the discriminator.

for meta-train-batch $\mathcal{B} \in \mathbb{R}^{B \times N_{\text{frames}} \times C \times w \times h}$ **do**
  unpack $\mathcal{B} \in \mathbb{R}^{B \times N_{\text{frames}} \times C \times w \times h}$ along support/query, source/target into:
  $S^{\text{support}},\ S^{\text{query}},\ T^{\text{support}},\ T^{\text{query}}$
  init $l_G^{\text{batch}} = 0$, $l_D^{\text{batch}} = 0$
  for each event i out of B in meta-batch **do**
    forward pass $T_i^{\text{support; generated}} = G(S_i^{\text{support}})$
    $l_G^{\text{adapt}} = \mathcal{L}_G(T_i^{\text{support; generated}}, T_i^{\text{support}}, S_i^{\text{support}}; w_G, w_D)$ from Eq. 1
    $l_D^{\text{adapt}} = \mathcal{L}_D(T_i^{\text{support; generated}}, T_i^{\text{support}}, S_i^{\text{support}}; w_G, w_D)$ from Eq. 2
    task-specific parameters $\phi_G \leftarrow w_G - \eta \nabla_{w_G} l_G^{\text{adapt}}$
    task-specific parameters $\phi_D \leftarrow w_D - \eta \nabla_{w_D} l_D^{\text{adapt}}$
    forward pass $T_i^{\text{query; generated}} = G(S_i^{\text{query}})$
    $l_G = \mathcal{L}_G(T_i^{\text{query; generated}}, T_i^{\text{query}}, S_i^{\text{query}}; \phi_G, \phi_D)$ from Eq. 1
    $l_D = \mathcal{L}_D(T_i^{\text{query; generated}}, T_i^{\text{query}}, S_i^{\text{query}}; \phi_G, \phi_D)$ from Eq. 2
    update rolling sums $l_G^{\text{batch}} += l_G$ and $l_D^{\text{batch}} += l_D$
  end
  backpropagate 2nd order updates $\nabla_{w_G} l_G^{\text{batch}}$ and $\nabla_{w_D} l_D^{\text{batch}}$ to w_G and w_D
end
return good initializations w_G and w_D for both generator and discriminator.

Algorithm 1: One Epoch MAML-Train Loop for U-Net Generator with Adversarial Loss.

C. Experimental Details

We run experiments on a single 32GB Nvidia Volta V100 GPU. For MAML optimization [1] we use meta-batch sizes of 2, 3 or 4 events. For the corresponding joint training baselines we used $N_{\text{support}} + N_{\text{query}}$ frames from each event and comparable number of events to keep comparisons fair. We randomly split all SEVIR events into 9169 train, 1162 validation, and 1148 test tasks. Joint training baselines and MAML outer loop optimizations are both performed using the Adam optimizer [14] with learning rate 0.0002 and momentum 0.5.

We resize all input modalities to have 192×192 resolution and keep the target at 384×384. The generator's encoder has four convolutional blocks, and the decoder has five. All generator blocks, except for the last decoder layer, use ReLU activation functions. The very last layer uses linear activation functions to support z-score normalization for all four image modalities.

D. Results

We conducted tests on our multi-task few-shot formulation, illustrating the empirical benefits of MAML. By comparing models trained using meta-learning algorithms with those under joint training—for both reconstruction and adversarial loss objectives—we consistently observed that meta-learning reduced the reconstruction error. However, superior perfor-

mance during training didn't consistently translate to enhanced weather evaluation metrics.

a) Reconstruction Loss: We compare joint training with MAML that uses a single adaptation step for each event. We evaluate model performance using weather metrics with two different thresholds (74 and 133). We summarize our evaluation results in Figure 2. While U-Nets trained via MAML exhibit superior performance on the optimization objective, this enhancement doesn't always lead to consistent benefits in weather-specific evaluations. Specifically, while finetuning for particular tasks bolsters precision, it adversely affects recall and IOU. Furthermore, issues inherent to training with reconstruction loss, such as the production of blurry outputs, persist.

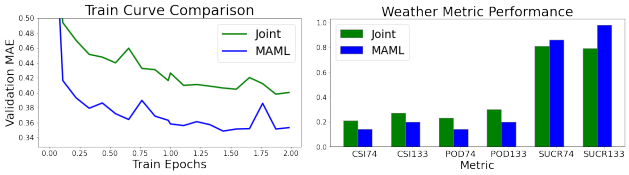

Fig. 2. Validation mean absolute error throughout training and test-set evaluation of weather-specific metrics. MAML optimization leads to better train objective and SUCR but lower CSI and POD.

Figure 3 displays synthetic VIL imagery produced by each method alongside the corresponding ground-truth data. The task adaptation mechanism helps recognize storm events in the lower-left corner. However, it struggles to accurately predict the shape of these low-intensity precipitations at a fine-grained scale.

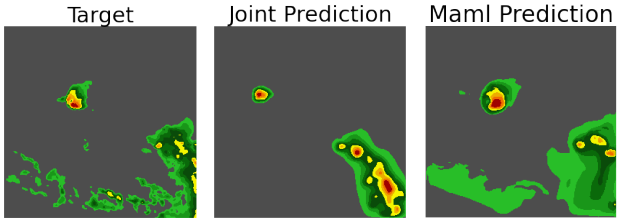

Fig. 3. Target VIL test-frame and generated samples. Task-adaptation helps recognize sparse VIL regions.

b) Adversarial Loss: We trained generative adversarial networks using both the second-order MAML procedure and the joint training baseline. In Figure 4, we analyze the progression of the reconstruction error during training and observe that MAML considerably aids in reducing the training objective. For these curves, we set $\lambda = 10^2$ and $\eta = 10^{-4}$.

Next, we evaluate on meteorological metrics for all values of λ and η, and summarize our results in Table I and II for joint and MAML training, respectively. For joint adversarial training, especially when evaluating with lower thresholds, we see the critical success index is fairly constant as we vary λ while increasing λ leads to lower recall and higher precision. This seems to suggest that placing more weight on the reconstruction loss will lead to predicting fewer high-valued VIL pixels.

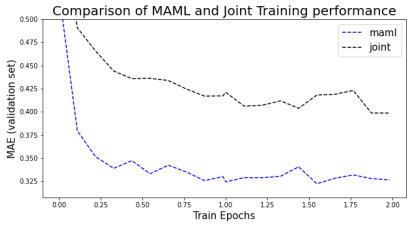

Fig. 4. **Adversarial loss - train curve.** MAML outperforms Joint Training. Evaluation is done on validation set throughout training.

TABLE I
Adversarial Joint - evaluation. TEST-SET EVALUATION ON METEOROLOGICAL METRICS.

thresh.	74			133		
metric	CSI	POD	SUCR	CSI	POD	SUCR
λ: 10^2	0.29	0.50	0.56	0.27	0.30	0.76
λ: 10^3	0.29	0.46	0.58	0.29	0.35	0.71
λ: 10^4	0.29	0.43	0.64	0.29	0.33	0.73

In the context of MAML adversarial training, we don't observe a discernible relationship between hyperparameters λ and η and the test-split meteorological metric values. This suggests increased instability during training. Such instabilities are amplified when attempting to optimize a bilevel Nash equilibrium problem via gradient descent, as per our earlier methodology. A comparison of Tables I and II indicates that, analogous to the reconstruction loss scenario, MAML optimization yields increased precision at the expense of diminished recall. A visual examination of the generated samples reveals instances of mode collapse, where the output doesn't even approximate realistic forms. In contrast, other samples closely mimic the ground-truth. We showcase successful samples generated via MAML-adversarial loss-trained models below, emphasizing significant variability in the proportion of realistic samples across distinct models. This disparity isn't mirrored in any evaluation metrics, reinforcing our hypothesis that in image generation, the association between evaluation performance and actual sample quality remains tenuous.

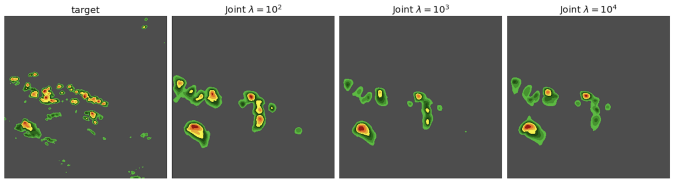

Fig. 5. **Adversarial Joint - generated samples.** Reconstruction loss biases the model towards sparser predictions.

Figures 5 and 6 (on the next page) compare samples generated by models trained on adversarial loss through either joint or MAML-based procedures for different values of λ. The MAML models all used an inner SGD learning rate of 10^{-5}. We see that in this case the intuitions from the reconstruction loss setting are still valid and the task-adaptation inherent

TABLE II
Adversarial MAML - evaluation. TEST-SET EVALUATION ON METEOROLOGICAL METRICS. MAML MODELS HAVE HIGHER PRECISION AND LOWER RECALL AND IOU.

thresh.		74			133		
metric		CSI	POD	SUCR	CSI	POD	SUCR
η: 10^{-4}	λ: 10^2	0.14	0.16	0.93	0.24	0.26	0.90
	λ: 10^3	0.09	0.09	0.98	0.20	0.20	0.99
	λ: 10^4	0.13	0.21	0.91	0.21	0.32	0.87
η: 10^{-5}	λ: 10^2	0.19	0.23	0.87	0.23	0.27	0.84
	λ: 10^3	0.17	0.20	0.90	0.25	0.29	0.87
	λ: 10^4	0.12	0.15	0.93	0.22	0.26	0.91

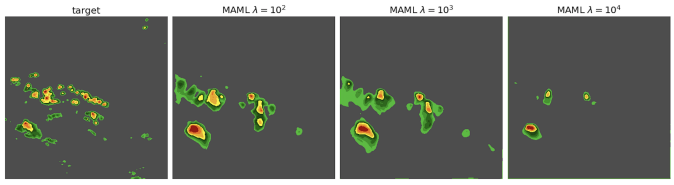

Fig. 6. **Adversarial MAML - generated samples.** Finetuning helps identify low-intensity VIL regions.

to MAML enables it to correctly generate low-intensity VIL data that joint-setting misses out on. We also confirm the aforementioned trend of higher λ values leading to lower VIL values.

IV. SELF-SUPERVISED PRE-TRAINING

A. Method

We follow recent work in self-supervised pretraining which applies contrastive learning to convolutional networks before finetuning on classification tasks and improves downstream performance and data efficiency. We ask if these improvements extrapolate to our image-to-image setup. The main distinction between our scenario and those in previous work is that we can initialize only a fraction of our parameters through contrastive pretraining.

We restrict our attention to the U-Net encoder parameters during the pretraining stage and follow the same network architecture as in Section III. Our experiments are inspired by the large-scale study on unsupervised spatiotemporal representation learning, conducted by [7]. In particular, we focus on MoCoV3 [4], which is a state-of-the-art contrastive learning method, because [7] identify the momentum contrast (MoCo) contrastive learning method as the most useful for our data.

a) Pre-training objective.: For a given representation q of a query frame from the dataset, a positive key representation k^+ and a negative key representation k^-, the loss function increases the similarity between the representations within the positive pair (q, k^+) and decreases the similarity within the negative pair (q, k^-) respectively. All representations are normalized on the unit sphere and the similarity is the dot product (i.e. the cosine similarity, because the representations are normalized). The loss is the InfoNCE loss [20]:

$$\hat{\mathcal{L}}_q = -\log \frac{e^{p(q)\cdot \mathrm{sg}(k^+)/\tau}}{e^{p(q)\cdot \mathrm{sg}(k^+)/\tau} + \sum_{k^-} e^{p(q)\cdot \mathrm{sg}(k^-)/\tau}} \tag{3}$$

for a temperature parameter τ and a predictor MLP p, which is a two layer MLP, with input dimension 128, hidden dimension 2048, output dimension 128, BatchNorm and ReLU in the hidden layer activation, and where "sg" is the stopgradient operation. Following [4], the gradients are not backpropagated through $k^{\{+,-\}}$ and the encoder representations both for keys and queries are obtained after a composition of the backbone and the projector (which is a two layer MLP, with dimensions [256, 2048, 128] with BatchNorm and ReLUs in between the hidden layers, and ending with a BatchNorm with no trainable affine parameters). Additionally, the branch for key representations follows the momentum update policy $\theta_\mathrm{k} \leftarrow m\theta_\mathrm{k} + (1-m)\theta_\mathrm{q}$ from [10] with momentum parameter $m = 0.999$, where θ_k are the weights in the key branch and θ_q are the weights in the query branch.

Data Augmentations. A unique challenge in our setup is selecting appropriate data augmentations. This is especially tricky with weather modalities, which have different invariances than natural images. For instance, the commonly used color jitter transformation isn't suitable in this context, given that image-to-image translation is color-sensitive. Of the standard augmentations, we explore random resized crops, random horizontal flips, gaussian noise, gaussian blur, random vertical flips, and random rotation. Additionally, we harness the temporal structure of SEVIR to derive "natural" augmentations, which we will discuss next.

Natural augmentations. We further consider using the temporal structure of SEVIR for augmentations, as follows. Each event consists of 49 frames, so we anchor every even frame as query frame and use every odd frame as key frame. For each query frame, to obtain q and k^+ we apply the following stochastic transformations to the frame twice: random resized crops using scale (0.8, 1.0); random horizontal flips with probability 0.5, pixel-wise gaussian noise sampled from the normal distribution $\mathcal{N}(0, 0.1)$ with probability 0.5, gaussian blur with kernel size 19, random vertical flips with probability 0.2, random rotation by angle unformly chosen in $(-\pi/6, \pi/6)$. The rest of the augmentation arguments follow the default in the Torchvision library[1]. In Figure 7 we present a conceptual visualization of the transforms. To obtain k^- we apply the above stochastic transformations to the corresponding key frame once.

Training hyperparameters. Our experiments use the following architectural choices: mini-batch, consisting of 3 events with 24 frames for queries and key 24 frames for keys each; 0.015 base learning rate; 100 pre-training epochs; standard cosine decayed learning rate; 5 epochs for the linear warmup; 0.0005 weight decay value; SGD with momentum 0.9 optimizer. We report the joint training reconstruction loss

[1]https://pytorch.org/vision/stable/transforms.html

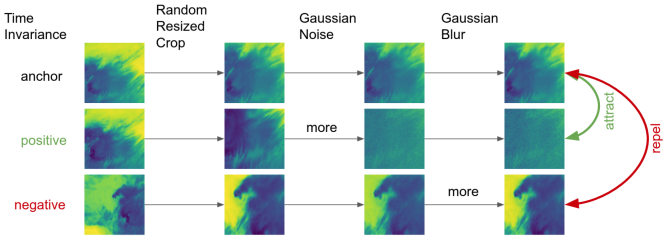

Fig. 7. **Augmentations for the contrastive learning experiment** By indicating "more" we show examples of a larger magnitude of the augmentation being applied.

experimental setup by finetuning the checkpoint obtained from pretraining.

B. Results

In Figure 8 we report our results. Firstly, for mean absolute error we find marginal yet somewhat consistent gains up to level 3 augmentation. Secondly, we also evaluate on meteorological metrics. We find that even though pretraining has a marginal effect on the reconstruction loss train objective, it often provides important gains on domain-specific evaluation criteria. We highlight the large improvement in CSI133 and POD133, which stems mostly from significant improvements in precision. We observe that up to level 4 MoCoV3 augmentation we obtain improvements throughout all measures with the contrastive pretraining. Finally, we show example samples in figure 9 and find that pretraining the U-Net encoder leads to better performance in high-VIL regions.

Lessons Learned from HPC Experimentation. Data augmentation is vital for our self-supervised studies, making preprocessing and sharding the data challenging. The SEVIR dataset on MIT SuperCloud uses individual files in HDF5 format for each video. To avoid slow and low memory utilization from stressing the file system, we limited batch sampling to a few videos, leveraging intra-video diversity, and also controlled the number of CPU workers we use for the data pipeline.

Furthermore, in Figure 10 we benchmark the performance of the data loaders as a function of the number of CPU workers. We observe that the data pipeline benefits from increasing the number of CPU workers. Level 1 (top) performs MoCoV3 without any data augmentation in contrast to Level 7 (below), which performs the *full* list of data augmentation transforms. Thus, we expect that Level 1's experiments will complete faster, as demonstrated in the figure. However, we also notice that increasing the number of workers reduces the latency coming from data augmentation since the red curves at the top and bottom perform similarly. Finally, notice that there is a diminishing return of increasing the number of workers for Level 1 given the small improvement from increasing from 16 to 32 (green to red). More workers speed up pretraining.

V. Conclusion and Future Work

A. Conclusion: novel few-shot multi-task image-to-image translation

We formulated a novel few-shot multi-task image-to-image translation problem leveraging spatiotemporal structure in a large-scale storm event dataset. We provided several benchmarks for this problem and considered two optimization procedures (joint training and gradient-based meta-learning) and two loss functions (reconstruction and adversarial). We trained U-Nets in all these regimes and presented each model's performance, as well as evaluated on various domain-specific metrics. We discussed the advantages and disadvantages of each of these. In this process we also explored a training method unexplored until now to the best of our knowledge: meta-learning adversarial GANs with second-order gradient updates. Additionally, we explored pretraining U-Net encoder parameters using various augmentations in both the spatial and temporal domains.

B. Future work: improving performance and stability

There are numerous tricks for training GANs that have been shown to work well in practice for natural image generation. An interesting research direction would be exploring if these gains extend to our meteorological domain. Two of these techniques are applying spectral normalization to the discriminator network and updating the generator network more often than the discriminator.

We have not fully explored the interplay between adversarial training and MAML's bilevel optimization, and we believe it would also be very interesting to further develop this aspect of our work. The most immediate next step could be meta-learning just a subset of the networks' parameters.

Another interesting direction would be applying importance sampling or even curriculum learning techniques to the training schedule. An important difference between SEVIR and the natural images datasets we are more accustomed to is that not all events are equally informative: our models can presumably learn much more from complex storms than from frames taken during calm weather where the VIL and lighting frames are very sparse, and the IR imagery has very little variance.

Acknowledgements

The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center [25] for providing HPC and consultation resources that have contributed to the research results reported within this paper/report.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. 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. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

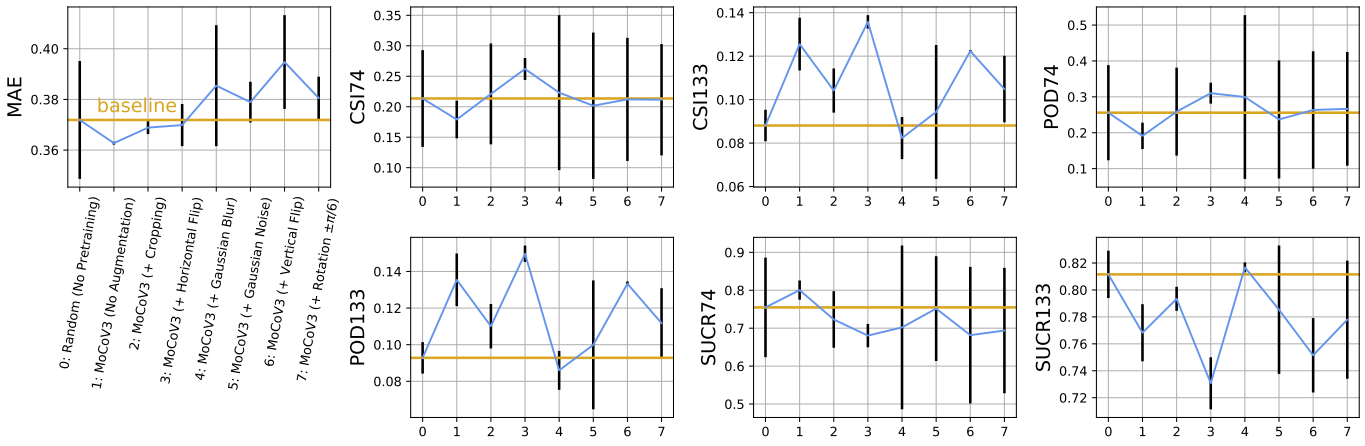

Fig. 8. **Contrastive Learning for SEVIR.** For mean absolute error lower is better. For every other evaluation measure, higher is better.

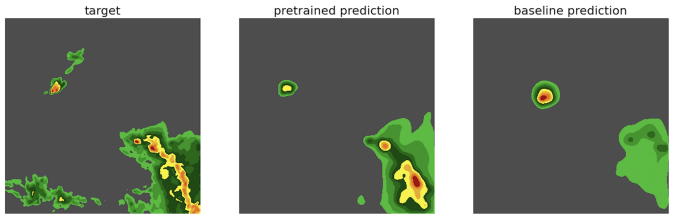

Fig. 9. **Pretrained encoder - generated samples** Pretrained models better identify the sparse high VIL values.

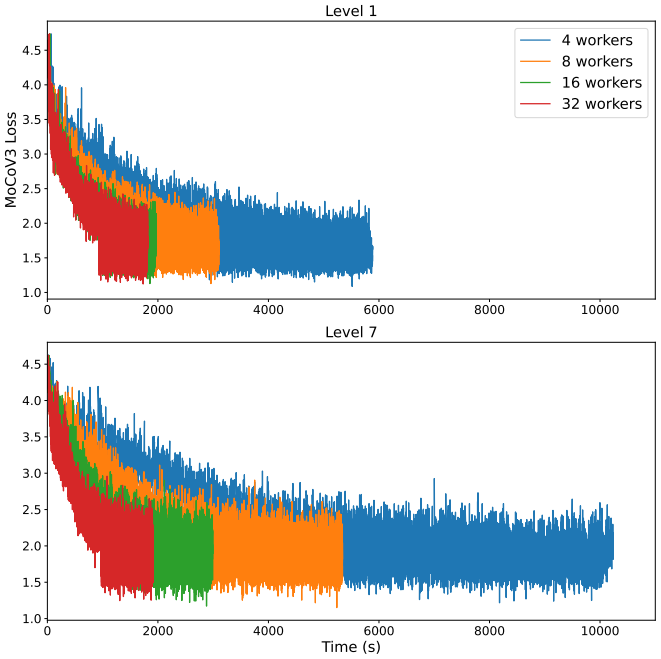

Fig. 10. **Benchmarking the data loaders on MIT SuperCloud.** Level 1 and 7 correspond to the experiments from Figure 9. We show MoCoV3 loss as a function of time on a single 32G V100 GPU and a variable number of CPU workers. The only deviation from the declared hyperparameters is that we train for 5 epochs with 1 epoch of linear warmup of the learning rate.

This material is also based in part upon work supported by the Air Force Office of Scientific Research under the award number FA9550-21-1-0317 and the U. S. Army Research Office through the Institute for Soldier Nanotechnologies at MIT, under Collaborative Agreement Number W911NF-18-2-0048. This work is also supported in part by the the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/).

REFERENCES

[1] Sébastien M R Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. learn2learn: A library for Meta-Learning research. August 2020. URL http://arxiv.org/abs/2008.12284.

[2] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv e-prints*, pages arXiv–2104, 2021.

[5] Louis Clouâtre and Marc Demers. FIGR: few-shot image generation with reptile. *CoRR*, abs/1901.02199, 2019. URL http://arxiv.org/abs/1901.02199.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and

L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[7] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/finn17a.html.

[9] Vijay Gadepally, Gregory Angelides, Andrei Barbu, Andrew Bowne, Laura J. Brattain, Tamara Broderick, Armando Cabrera, Glenn Carl, Ronisha Carter, Miriam Cha, Emilie Cowen, Jesse Cummings, Bill Freeman, James Glass, Sam Goldberg, Mark Hamilton, Thomas Heldt, Kuan Wei Huang, Phillip Isola, Boris Katz, Jamie Koerner, Yen-Chen Lin, David Mayo, Kyle McAlpin, Taylor Perron, Jean Piou, Hrishikesh M. Rao, Hayley Reynolds, Kaira Samuel, Siddharth Samsi, Morgan Schmidt, Leslie Shing, Olga Simek, Brandon Swenson, Vivienne Sze, Jonathan Taylor, Paul Tylkin, Mark Veillette, Matthew L Weiss, Allan Wollaber, Sophia Yuditskaya, and Jeremy Kepner. Developing a series of ai challenges for the united states department of the air force. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7, 2022. doi: 10.1109/HPEC55821.2022.9991948.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[12] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey, 2020.

[13] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[15] B. Lake, R. Salakhutdinov, and J. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332 – 1338, 2015.

[16] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. The omniglot challenge: a 3-year progress report, 2019.

[17] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Alexander Pritzel, Suman Ravuri, Timo Ewalds, Ferran Alet, Zach Eaton-Rosen, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[18] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.3031549. URL http://dx.doi.org/10.1109/ACCESS.2020.3031549.

[19] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[21] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[23] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, Sep 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL https://doi.org/10.1038/s41586-021-03854-z.

[24] Xiaoli Ren, Xiaoyong Li, Kaijun Ren, Junqiang Song, Zichen Xu, Kefeng Deng, and Xiang Wang. Deep learning-based weather prediction: a survey. *Big Data Research*, 23:100178, 2021.

[25] Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, et al. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

[26] Ileana Rugina, Rumen Dangovski, Mark Veillette, Pooya Khorrami, Brian Cheung, Olga Simek, and Marin Soljačić. Meta-learning and self-supervised pretraining for real world image translation. *arXiv preprint arXiv:2112.11929*, 2021.

[27] Timothy J. Schmit, Paul Griffith, Mathew M. Gunshor, Jaime M. Daniels, Steven J. Goodman, and William J. Lebair. A closer look at the abi on the goes-r series. *Bulletin of the American Meteorological Society*, 98(4):681 – 698, 2017. doi: 10.1175/BAMS-D-15-00230.1. URL https://journals.ametsoc.org/view/journals/bams/98/

4/bams-d-15-00230.1.xml.

[28] Martin G. Schultz, C. Betancourt, Bing Gong, Felix Kleinert, Michael Langguth, Lukas Hubert Leufen, Amirpasha Mozaffari, and Scarlet Stadtler. Can deep learning beat numerical weather prediction? *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379, 2021. URL https://api.semanticscholar.org/CorpusID:231919857.

[29] Arvind Sridhar. Meta-GAN for few-shot image generation. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. URL https://openreview.net/forum?id=SE3Gy6E_PWq.

[30] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *CoRR*, abs/1903.03096, 2019. URL http://arxiv.org/abs/1903.03096.

[31] Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir : A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22009–22019. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Paper.pdf.

[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2017.

[33] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation, 2022.

[34] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. *CoRR*, abs/1910.12027, 2019. URL http://arxiv.org/abs/1910.12027.

[35] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training, 2020.

[36] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans, 2020.