Cross-Modality Distillation: A Case for CGANs

{Siddharth Roheda*, Benjamin S Riggan†, Hamid Krim*, and Liyi Dai‡}
*North Carolina State University, Raleigh, NC † U.S. Army Research Laboratory, Adelphi, MD
‡Army Research Office, RTP, Raliegh, NC

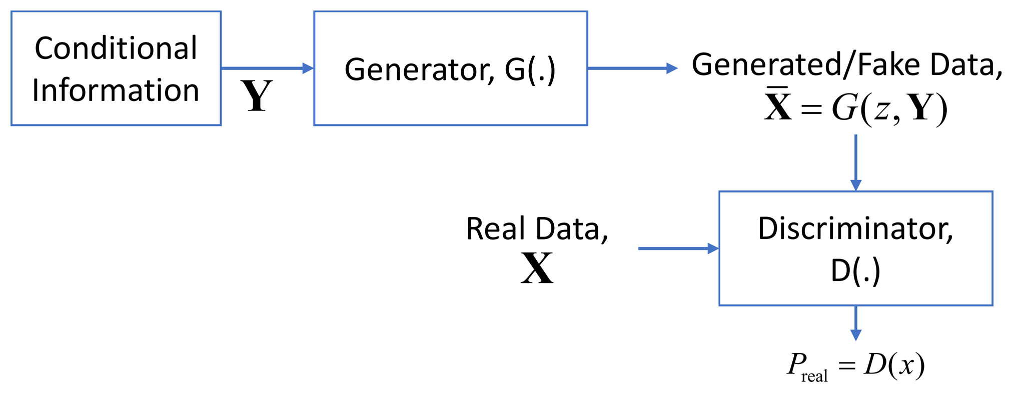
Objectives

Our goal is to safeguard detection/recognition accuracy of a sensor network when some sensors in the network are noisy, missing, or damaged by,

1. Exploiting prior knowledge about the relationship between sets of modalities.
2. Distilling knowledge from the missing modalities into the model trained for dealing with available modalities.

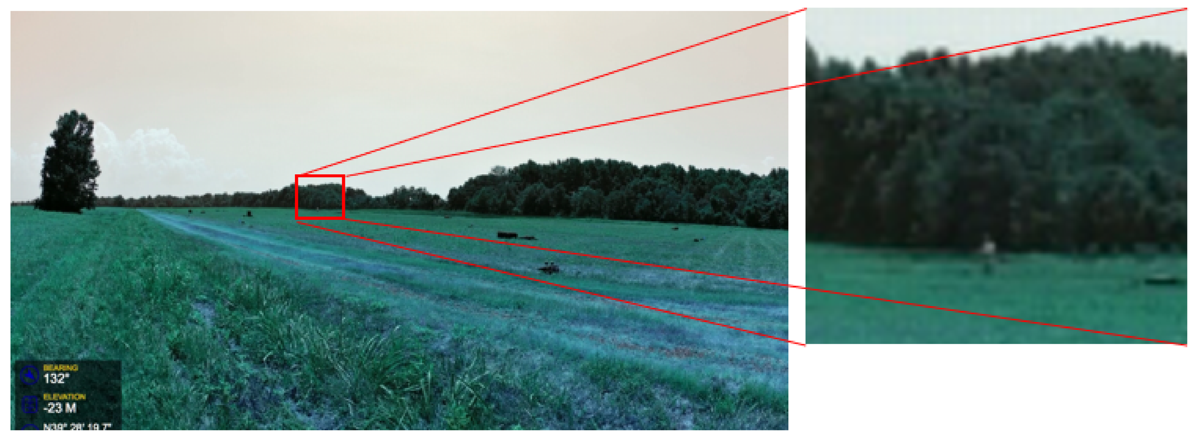
Introduction

The concept of a conditional Generative network can be used, where the input to the generator is the available modality, while the output is the targetted feature space.

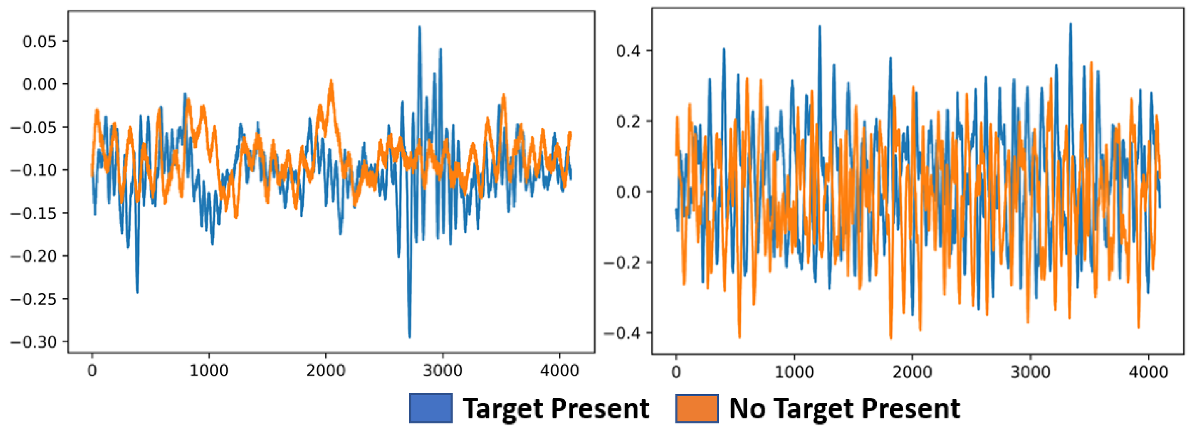


Dataset

We use pre-collected data from a network of seismic sensors, acoustic sensors, and video cameras deployed in a field:



A sample video frame with a small human target

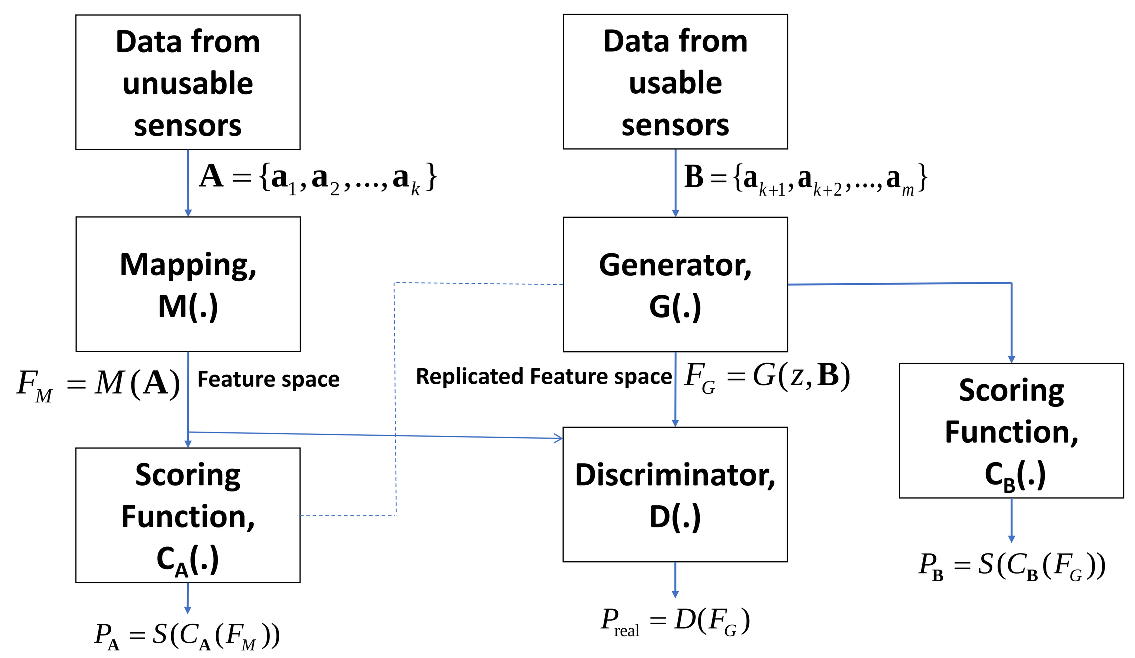


Samples from seismic (left) and acoustic (right) sensors

Given these sensor observations, the objective is to detect humans in the field.

Proposed Model

We extend the idea of Conditional Generative Adversarial Networks to generate features with highly discriminitive information, and modify the original CGAN objective function to ensure this:



$$\min_{G(.)}\max_{D(.)} J(D,G) = V(D,G) + \alpha C_{A_{Loss}}(F_G, U) + \beta L^2(F_M, F_G) \quad (1)$$

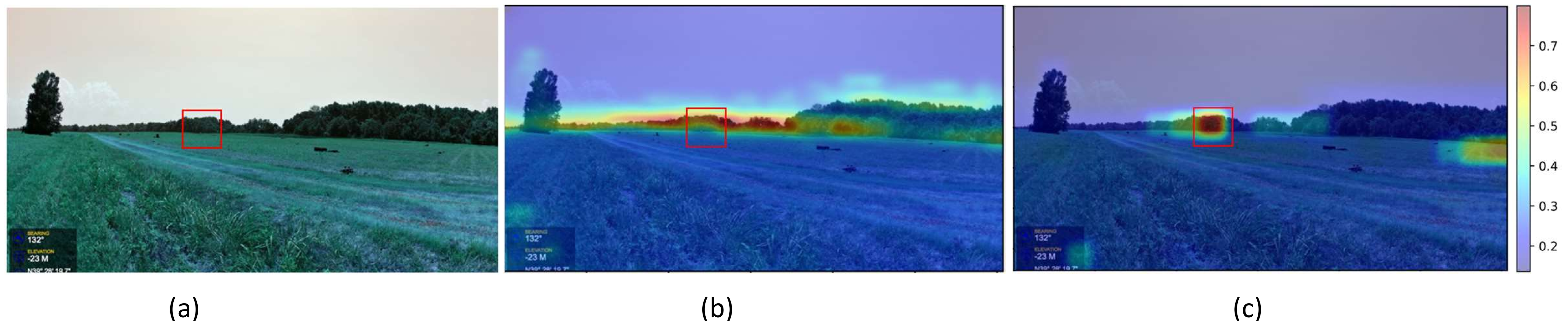
Where,

$$V(D,G) = \mathbb{E}_{B \sim p_{data}(B), M(A) \sim p_{data}(M(A))}[log D(B, M(A))] + \mathbb{E}_{B \sim p_{data}(B), z \sim p_z(z)}[log(1 - D(B, G(z, B)))] \quad (2)$$

$$C_{Loss}(X, L) = \sum_i -l_i log S(C(x_i)) \quad (3)$$

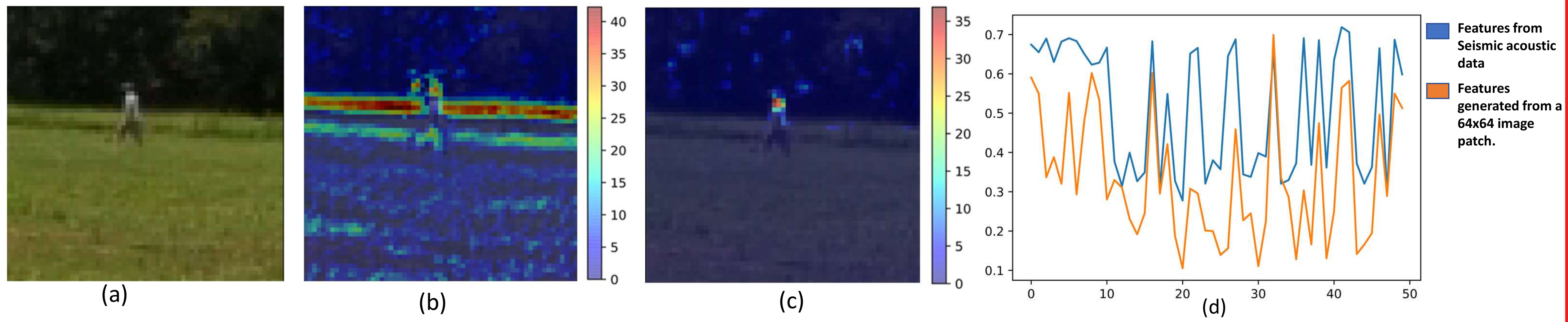
$$L^2(F_M, F_G) = \sum (f_{M_i} - f_{G_i})^2 \quad (4)$$

Results



(a): Original video frame, (b): Detection Probability when network was trained using just the video frames, (c): Detection Probability when network was trained using proposed Cross-Modal distillation. Red box marks true location of target

Detection result on a frame captured from a different angle shows robustness of the proposed algorithm: (a): Original video frame, (b): Detection Probability when network was trained using proposed Cross-Modal distillation. Red box marks true location of target



(a): Positive Patch, (b): Output from convolutional layer trained using video frames only, (c): Output from convolutional layer trained using proposed Cross-Modal distillation, (d): Comparison between real and corresponding generated features

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Conclusion

We proposed a CGAN based technique for cross-modal distillation that:

- Can generate representative features from unavailable modalities,
- Takes into account the pre-trained classifiers from unavailble modalities,
- Safeguards Detection accuracy when sensors with high discriminitive information are damaged.

Contact Information

Email sroheda@ncsu.edu
Phone +1 (984) 244 4685