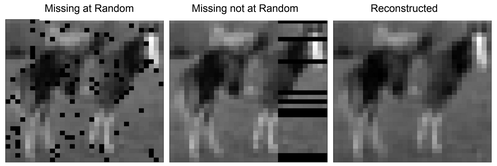
SpatialGAIN: Spatial Generative Adversarial Imputation Networks and its Application to Zero-Shot Pixel Imputation

Justin Young
justiny@stanford.edu

Stanford | ENGINEERING
Computer Science

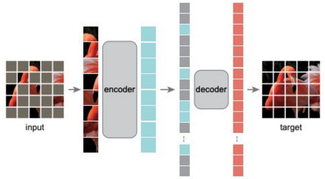
Problem Statement

Pixel Imputation

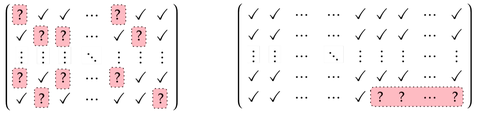


Existing Methods

- Leading approaches are successful but computationally very expensive to train (e.g. Vision Transformers, He et al. [2021])
- Deterministic imputations do not allow for inference (e.g. DAE)



- Zero-shot methods that use no training data are inflexible.



Goals

We propose Spatial Generative Adversarial Networks (SpatialGAIN):

1. Computationally Feasible
2. Yields Sampling Distribution of Imputations

We first prove SpatialGAIN's viability by showing its performance when trained on a class of images from ImageNet (**Experiment 1**). Then, we use SpatialGAIN as a tool in the zero-shot setting (**Experiment 2**):

3. Can be used to help flexibly impute in the zero-shot setting.

Methods

Generative Adversarial Networks (GAIN)

Shown as a improvement to a wide array of discriminative and generative models (Yoon et al. [2018]). Allows for incomplete training data, unlike autoencoder models!

$$\text{Mask } \boldsymbol{M} \in \{0,1\}^d \quad \text{Data } \boldsymbol{X} \in \mathbb{R}^d \quad \text{Observed } \tilde{X}_i = \begin{cases} X_i & \text{if } M_i = 1 \\ * & \text{otherwise} \end{cases}$$

$$\text{Dataset } \mathcal{D} = \{(\tilde{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)})\}_{i=1}^n \quad \text{Learn } P(\boldsymbol{X}|\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}})$$

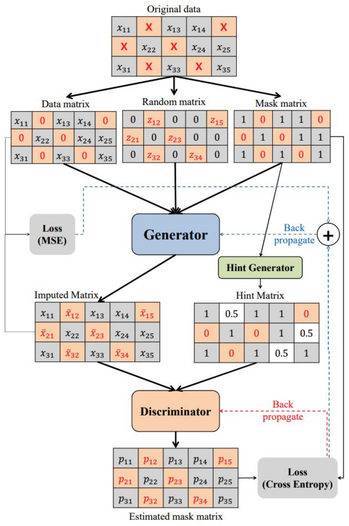
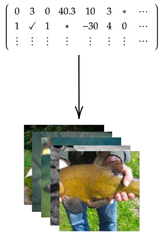
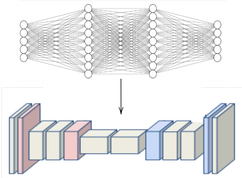


Figure 1 in Yoon et al. [2018]



GAIN to SpatialGAIN: Data



GAIN to SpatialGAIN: D and G

Left: Original GAIN framework
Right: Changes of SpatialGAIN

SpatialGAIN

Same set-up but extends framework to image data by using CNNs, a different masking matrix, and slightly altered hint generator.

Experiment 1

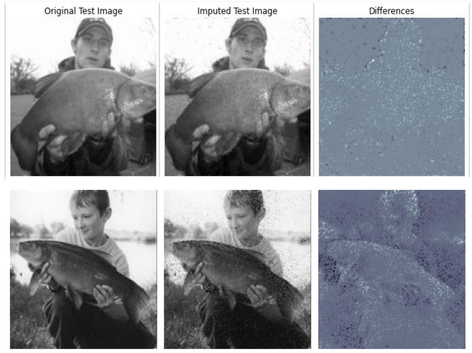
Baseline. We compare GAIN to SpatialGAIN to see whether we reap benefits on image vs. tabular data. This provides a first sanity check. Additional baselines like DAE are left for future work.

Dataset. We use UCI's Spambase dataset and Imagenette, a downsampled version of ImageNet. We perform a 80/20 train/test split on 4601 observations from Spambase (tabular data) and 1350 160x160 grayscale images from Imagenette (downsampled ImageNet).

Methods. In order to use SpatialGAIN on tabular data, we artificially reshape tabular data. In order to use GAIN on image data, we flatten the image into a vector.

	GAIN	SpatialGAIN
Spam	0.051 (0.002)	0.109 (0.022)
Imagenette	0.171 (0.002)	0.076 (0.005)

Table 1. Experiment 1 Results: RMSE on Test Set



Top: Example of SpatialGAIN imputation
Bottom: Example of GAIN imputation

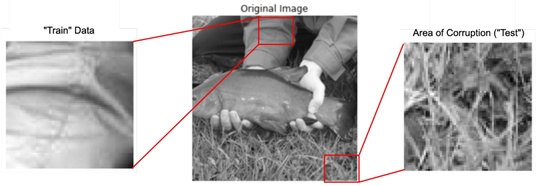
Analysis. Significant degradation of GAIN on image data while that of SpatialGAIN on tabular data is relatively modest.. From a relatively small training set (<1000), this suggests SpatialGAIN is able to learn spatial dependencies and what constitutes a fish and its salient surroundings. Note from the heat map that SpatialGAIN struggles at borders.

Experiment 2

Baseline. In the zero-shot setting, we compare SpatialGAIN to matrix compilation to see whether a deep learning model can work.

Dataset. We fix one randomly chosen 320x320 from Imagenette.

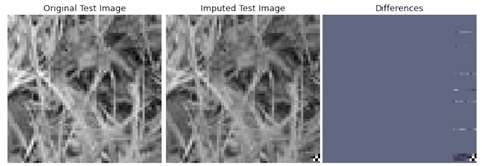
Methods. We consider 20% MCAR and MNAR missingness in a localized region. We construct a training set from surrounding subsets.



	SpatialGAIN	MC	NN
MCAR	0.090 (0.002)	0.054 (n/a)	0.143 (n/a)
MNAR	0.125 (0.013)	0.021 (n/a)	0.151 (n/a)

Table 2. Experiment 2 Results: RMSE on $\boldsymbol{A}$, MCAR and MNAR.

Analysis. We see SpatialGAIN performs poorly relative to matrix completion but still reasonable. We pruned the "train" set based on pixel similarity, helping enforce learning of spatial dependence.



SpatialGAIN, Zero-shot Setting with MNAR Missingness

Future Work

Experiment 1:

- Establishing more baselines, namely with DAE and VAE.
- Understanding MCAR and MNAR missingness with training data.
- Exploring the impact of training data size on performance.

Experiment 2:

- Determining what subsets may see success (low RMSE) and why.
- Exploring different missingness structures and missing percentages.