

Latent Representation of Microstructures using Variational Autoencoders with Spatial Statistics-Space Loss

Andy Cai
 University of Toronto
 Toronto, ON, Canada
 andy.cai@mail.utoronto.ca

Sayed Sajad Hashemi
 Georgia Institute of Technology
 Toronto, ON, Canada
 sajjads.hashemi@gmail.com

Noah Paulson
 Argonne National Laboratory
 Lemont, IL, USA
 npaulson@anl.gov

Michael Guerzhoy
 University of Toronto
 Toronto, ON, Canada
 guerzhoy@cs.toronto.edu

Abstract

We propose the use of Cross-Entropy on 2-point spatial statistics (CESS) as a reconstruction loss term for a variational autoencoder, creating a small latent-space representation of microstructures from which microstructures can be reconstructed. A prospective application of small reversible microstructure representations is more efficient optimization for materials properties in latent space. In Materials Science, 2-point spatial statistics have been shown to be good representations of microstructure properties, and have many desirable invariances (translation, phase label, inversion). To our knowledge, we are the first to demonstrate a system that successfully creates latent representations of realistic simulations of microstructures by using an error term that minimizes the distance between input and reconstruction in spatial statistics space. We also demonstrate some promising preliminary qualitative results that show improved quality of reconstructions using CESS loss. We demonstrate compression from 224x224 binary microstructure images to 14x14 latent representations.

1. Introduction

1.1. Microstructures

The structure of a material strongly influences its properties; this is known as the structure-property linkage. Small-scale structure, also known as microstructure, is a key component of this. For example, a key microstructural feature is grain size, which is the size of individual crystal regions within a material. Grain size is strongly correlated with mechanical properties such as yield strength, due to the Hall-Petch relationship, which links smaller grains to increased strength.

[8]. Understanding microstructures is thus critical for material design.

A common representation of microstructure is through a microscope image, roughly at the μm scale (see Fig. 1 for an example); however, this representation is difficult to use in computational material design, as it requires additional further processing before useful data can be extracted.

A more tractable representation is the probability distribution of local state at a certain location, sometimes also referred to as 1-point spatial statistics [6]. One common local state used is phase, which represents the "type" of material present at a location; this can be distinguished through chemical or structural properties. For the rest of this paper, we will be using the term microstructure to refer to this representation.

This data is often derived from the pixel-wise labeling of microstructure images; instead of a true probability distribution, each pixel will be an observation of a specific phase (i.e. $m(h, \mathbf{x}) = 1$). Microstructures are thus essentially a series of segmentation masks for each phase.

In this paper, we work primarily with 2-phase microstructures, which can be represented as a single binary image, with each phase represented as a pixel value of 0 or 1. However, the ideas proposed in the paper are easily extensible to microstructures with more phases.

1.2. 2-Point Spatial Statistics

Representations of microstructure are often localized to small regions, so the probabilistic properties that can be generalized to the rest of the material are often more important than specific details. One common probabilistic measure is the 2-point spatial statistics; intuitively, it describes how likely it is to find two specific phases separated by a

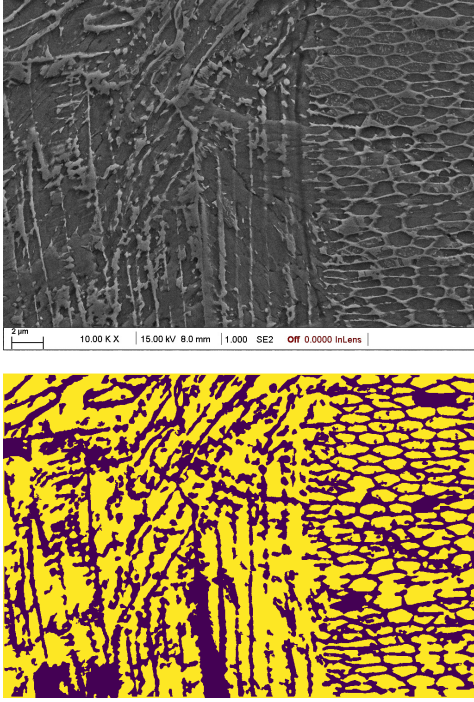


Figure 1. Examples of different representations of microstructure. Top: Microscope image of steel. Bottom: Probability distribution of presence of Austenite (a phase of iron); yellow represents a probability of 1, purple represents a probability of 0.

certain distance and direction within a microstructure. It can be stated formally as follows:

$$\begin{aligned} f(h, h' | \mathbf{r}) &= \frac{1}{\text{Vol}(\Omega_x)} (m(h, \mathbf{x}) \otimes m(h', -\mathbf{x}))(-\mathbf{r}) \\ &= \frac{1}{\text{Vol}(\Omega_x)} \int_{\Omega_x} m(h, \mathbf{x}) m(h', \mathbf{x} + \mathbf{r}) d\mathbf{x} \quad (1) \end{aligned}$$

, where \mathbf{r} is the offset vector, h and h' are the respective phases, $m(h, \mathbf{x})$ is the probability of state h at location \mathbf{x} , and Ω_x is the space of valid locations.

The computation of 2-point spatial statistics can be done efficiently through the Fast Fourier Transform using the Convolution Theorem:

$$m(h, \mathbf{x}) \otimes m(h', -\mathbf{x}) = \mathcal{F}^{-1} \{ \mathcal{F}\{m(h, \mathbf{x})\} \mathcal{F}\{m(h', -\mathbf{x})\} \} \quad (2)$$

Paulson et al. [15] have shown that 2-point spatial statistics effectively represent material properties while simultaneously being highly compressible through classical techniques like Principal Component Analysis (PCA). However, modern machine learning techniques are likely capable of extracting more compressibility.

2-point spatial statistics also have some inherent invariances, like periodic translation invariance and inversion invariance.

Our work is motivated the insight in Materials Science that 2-point spatial statistics are a good representation of microstructures [15]. We therefore work on obtaining a latent representation that can produce a reconstruction that is close to the encoded microstructure in spatial statistics space, which would indicate the representation is good.

1.3. Main Contributions

We create a compressed, reversible latent space that explicitly preserves statistical information, which could be used as a computationally efficient proxy for microstructures.

We created this latent space by training a Variational Autoencoder (VAE) with a reconstruction loss that explicitly minimized the distance between the 2-point spatial statistics of the input and the reconstruction.

Our paper builds on prior work by Hashemi et al. [9], who introduced the use of spatial statistics loss for microstructure modeling, but failed to show its effective convergence on a realistic dataset. We propose a new formulation of this loss, the Cross Entropy on Spatial Statistics (CESS), along with several variants that address training difficulties. We demonstrate successful model convergence on a realistic synthetic microstructure dataset, addressing some of the previous limitations.

Qualitative results also indicate that the quality of reconstructions trained with spatial statistics loss is slightly better compared to models trained without spatial statistics loss.

1.3.1. Potential application

The main potential application in materials design is search for materials with favorable properties in latent space, while being able to “go back” to pixel space both in order to compute properties of materials that correspond to particular latent representations and to generate microstructure from a good latent code that was found.

2. Related Work

The problem of representing microstructures in a lower-dimensional space has been addressed by recent research, but not generally with an explicit focus on preserving probabilistic properties like spatial statistics. Most related to our work is the research done by Sardeshmukh et al. [19], which utilizes a VAE trained with “style loss” to mitigate the issues with Mean Squared Error (MSE) on microstructure images. They define style loss as follows:

$$\mathcal{L}_{style}(x, \hat{x}) = \sum_{l=0}^L w \left[\frac{1}{4C_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - \hat{G}_{ij}^l) \right] \quad (3)$$

, where G^l is the Gram matrix of the feature map from layer l created by passing the image to the VGG-19 model.

Existing research has shown that feature maps of CNNs, especially from the first layer, resemble Fourier filter banks [3], potentially implying that style loss has some structural similarity to spatial statistics loss, in the sense that both are computed with outputs of Fourier-like filters as inputs.

Bostanabad et al. [2] demonstrated the reconstruction of microstructures with similar spatial statistics, but they do not create a compressed representation. Existing research by Paulson et al. [15] on the compression of 2-point statistics used PCA, but that work did not demonstrate reconstruction of the microstructure from the latent space.

3. Methodology

3.1. Variational Autoencoders (VAEs)

The Variational Autoencoder (VAE) [10] is an architecture that performs unsupervised learning of latent representations of data.

We can define the properties of a latent space z with a prior, $p(z)$, and some parameterized relation to x , $p_\theta(x|z)$. This allows us to easily compute the joint distribution $p(x, z) = p(z)p_\theta(x|z)$; however, we need $p(x)$ to compute the likelihood of the data. Marginalization is computationally infeasible, and applying Bayes' rule would require knowledge of the ground-truth encoder, $p(z|x)$, so instead a lower-bound approximation of $\log p(x)$ is derived:

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x) \| p(z))$$

This is known as the **Evidence Lower Bound (ELBO)**. The term on the left is commonly referred to as the reconstruction term; it intuitively represents how well the data is preserved after being passed through the encoder and decoder.

3.2. Spatial Statistics Space Loss

Typically, the assumption is that the decoder distribution $p_\theta(x|z)$ is Gaussian, and thus that the reconstruction loss is the Mean Squared Error (MSE).

$$p_\theta(x|z) = N \exp\left(-\frac{(x - F_\theta(z))^2}{2\sigma^2}\right) \quad (4)$$

$$-\log p_\theta(x|z) = \frac{(x - F_\theta(z))^2}{2\sigma^2} + C \quad (5)$$

$$= L_{\text{MSE}}(x, F_\theta(z)) + C \quad (6)$$

However, this has been shown to be a poor assumption for images [20] and microstructures [19], as physically similar microstructures (i.e. differing by a small shift) can be distant in terms of MSE.

To address this, we sought a loss function that preserves 2-point spatial statistics, as it would preserve important structure-property information, result in a compress-

ible latent space, and possess desirable invariance properties. We initially investigated two loss formulations based on spatial statistics: Mean Squared Error on Spatial Statistics (MSESS) (7) and Cross Entropy on Spatial Statistics (CESS) (8).

$$L_{\text{MSESS}}(f_1, f_2) = \frac{1}{\text{Vol}(\Omega_r)} \sum_{\mathbf{r}} \sum_{h, h'} (f_1(h, h'|\mathbf{r}) - f_2(h, h'|\mathbf{r}))^2 \quad (7)$$

$$L_{\text{CESS}}(f_1, f_2) = \frac{1}{\text{Vol}(\Omega_r)} \sum_{\mathbf{r}} \sum_{h, h'} f_1(h, h'|\mathbf{r}) \log f_2(h, h'|\mathbf{r}) \quad (8)$$

An auxiliary "clamp" loss which forced pixels towards 0 or 1 was also experimented with. This was motivated by previous work from Fullwood et al. [6], which showed that microstructures could generally be reconstructed from their spatial statistics if a binary constraint was enforced. It was defined as the negative log likelihood of a Gaussian mixture with means at 0 and 1:

$$L\{x\} = - \sum_{i,j} \log \left(\exp \frac{-x_{ij}^2}{2\sigma} + \exp \frac{-(x_{ij} - 1)^2}{2\sigma} \right) \quad (9)$$

A rough test of the effectiveness of these losses was done by using them to reconstruct simple synthetic microstructures through direct gradient descent (i.e. the pixel values of an image were optimized to match the spatial statistics). The target microstructures were generated by adding 5 circles of radius 20px at random locations to an image of size 224x224; these will be referred to as synthetic circle microstructures for the rest of this paper.

Reconstructions using loss functions without a component of clamp loss are unreasonable and often fail to look similar to their original microstructures (see Fig. 2).

Reconstructions with clamp loss were more reasonable and similar to the original microstructure, with the cross entropy loss often reconstructing the original microstructure with an invariant transformation (see Fig. 3).

However, initial training runs using CESS and clamp loss led to poor results (see Fig. 4), producing unrealistic microstructures and often failing to converge. A hypothetical explanation for this can be found in Supplementary Material 1.4.

2 main variants of CESS were proposed to solve this problem: a linear combination of MSE on microstructures and CESS (10), and CESS on only the center NxN crop of spatial statistics (11).

$$L(m_1, m_2, f_1, f_2) = \alpha L_{\text{MSE}}(m_1, m_2) + L_{\text{CESS}}(f_1, f_2) \quad (10)$$

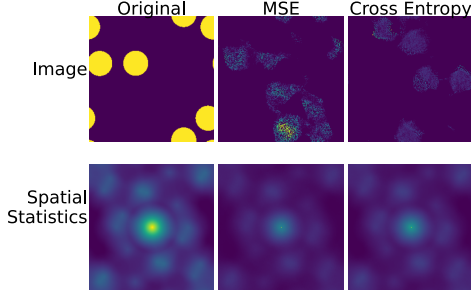


Figure 2. Initial reconstructions of microstructure from spatial statistics using various loss functions.

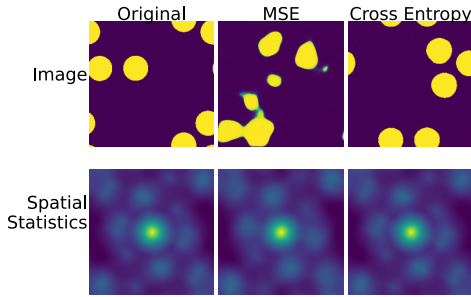


Figure 3. Reconstruction of microstructure from spatial statistics using various loss functions and a combination of clamp loss.

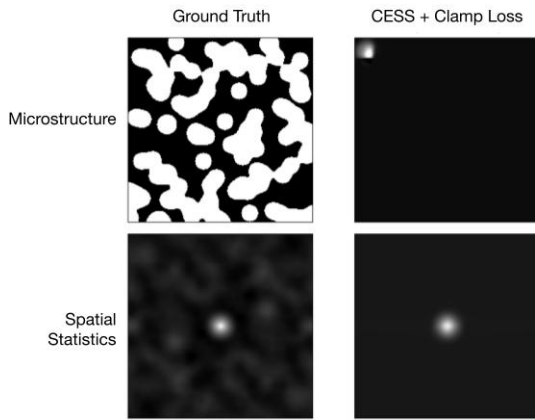


Figure 4. Results from training with only CESS and clamp loss.

$$L(f_1, f_2) = \sum_{\mathbf{r} \in [-\frac{N}{2}, \frac{N}{2}]} \sum_{h, h'} f_1(h, h' | \mathbf{r}) \log f_2(h, h' | \mathbf{r}) \quad (11)$$

3.3. Model Architecture

Our model architecture (see Fig. 5) consists of repeating blocks of skip-connected convolutions, each followed by downsampling/upsampling for the encoder and decoder respectively. The number of skip-connected convolutions is

a hyperparameter, and will be referred to as depth later in the paper. The number of total downsampling/upsampling blocks is fixed at 4 to create a bottleneck of 14x14, from an initial image size of 224x224. The architecture was heavily inspired by TAESD [1], a VAE typically used to improve the results of the image generation from stable diffusion.

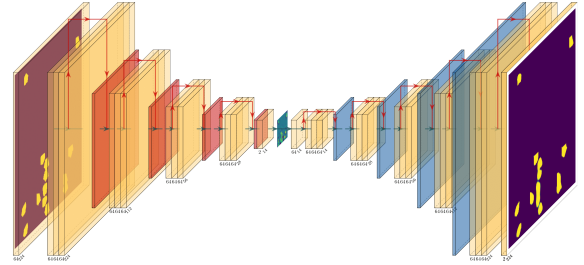


Figure 5. Architecture diagram for model with depth 1. Yellow layers are convolutions, red layers are pooling, and blue layers are upsampling. A ReLU activation is assumed to follow every convolution layer.

3.4. Model Training

5 main experiments were run:

- Model trained with MSE
- Model trained with linear combination of MSE and CESS
- Model trained with MSE, then finetuned with CESS
- Model trained with MSE, then finetuned with CESS on a 50x50 center crop
- Model trained with Style Loss

For all models, a Bayesian hyperparameter sweep was conducted to optimize for their respective losses on the validation set, with 25 trials trained to 5 epochs. All models were then trained for 30 epochs, using the RAdam [13] optimizer and a Cosine Warm Restart learning rate scheduler [14] with a period of 10 epochs. The specific optimal hyperparameters for each model can be found in Supplementary Material 1.1.

3.5. Dataset

The microstructures used for the experiments were sourced from the MICRO2D dataset [16], a set of 87 379 synthetic 2-phase microstructures represented as 256x256 binary images. These are categorized into 10 general classes, examples of which are visualized in Fig. 6. One of the explicit goals of the dataset is to cover a diverse set of 2-point spatial statistics, which is ideal for testing our model's effectiveness. The images were finally resized to 224x224 so that experiments utilizing VGG-based style loss could be consistent.

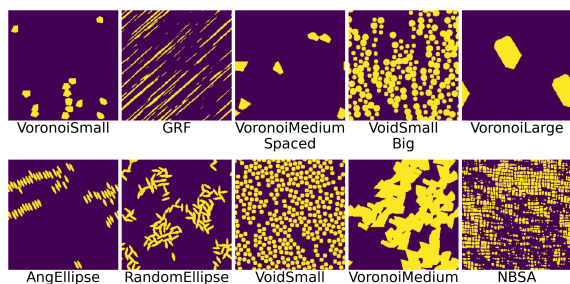


Figure 6. Sample image for each of the 10 classes in the MICRO2D dataset.

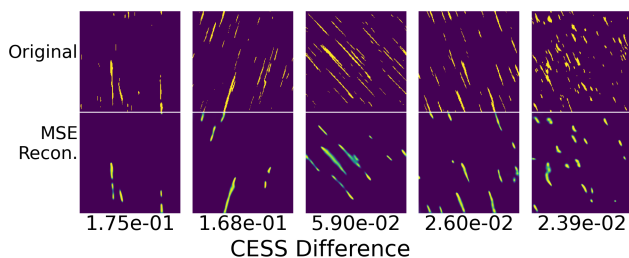


Figure 7. Outlier samples and reconstructions for model trained with MSE loss. CESS score of MSE reconstruction is labeled below.

4. Results

Our goal is to obtain latent representations that can generate plausible-looking generated outputs that are close to the inputs in spatial statistics space. This is motivated by the intuition from Materials Science that microstructures that are similar in spatial statistics space are structurally similar. We assess our system both qualitatively and quantitatively.

4.1. Quantitative Results

Our main goal was to determine whether the model was able to converge using this loss by analyzing the CESS losses on the test set (visualized in Fig. 9). As seen in the figure, the median CESS on the test set was slightly smaller for models trained with CESS compared to models trained without. Mean test set CESS was larger than the median for all models, and also showed a larger difference between models. This suggested the existence of outliers, and that the model trained on MSE was more heavily impacted by them. Some of these outliers are visualized in Fig. 7.

The existence of these outliers and the fact that the loss distribution resembled a log-normal prompted us to look at log CESS (visualized in Fig. 10). The resulting histogram contains a small mode with lower loss, which we hypothesize means that there exist a class of microstructures that are especially easy to optimize for.

Qualitative examination of outliers with high loss

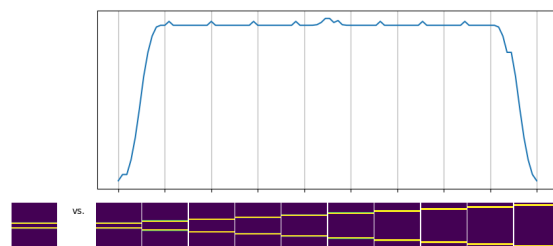


Figure 8. Example of the sensitivity to small changes present in "line-like" spatial statistics. CESS is computed between the image on the far-left and the images on the x-axis, which modify the spacing of the lines.

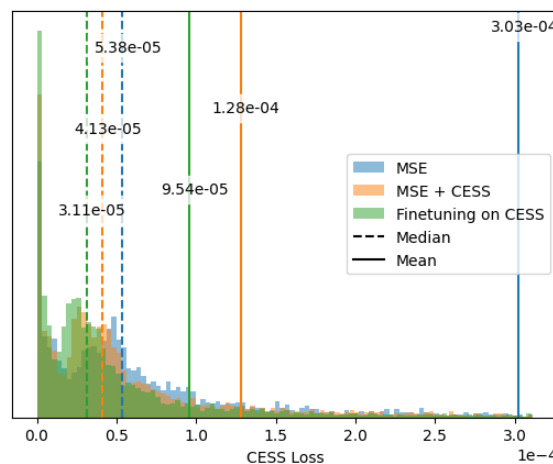


Figure 9. Histogram of CESS loss for various models on test set, along with lines representing their associated medians. Largest quintile is cutoff to make visualization clearer.

showed that they are mainly sparse and line-like; these are likely hard to deal with as the 2-point spatial statistics will also be sparse, resulting in high sensitivity to small changes. As an illuminating example, if two 1-pixel horizontal lines are spaced 100 pixels apart, then the CESS to an image with lines spaced 101 pixels apart will be equivalent to an image with lines spaced 200 pixels apart (see Fig. 8). We hypothesize that the model trained with MSE is especially impacted as it seems to ignore some of the smaller features in its reconstruction.

To further examine convergence with less influence from outliers, the CESS loss was compared pairwise for each model on each sample in the test set. The percentage of samples where Model A (y-axis) had a lower loss than Model B (x-axis) is shown in Fig. 11. A value close to 0.5 would indicate close to equivalent performance, whereas a value close to 1 would indicate Model A was better than Model B.

Models using CESS are seen to have much better performance when compared to models trained with MSE and

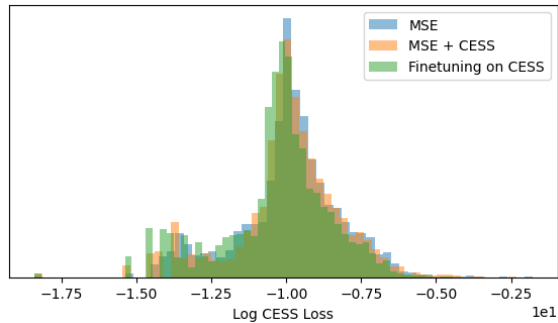


Figure 10. Histogram of Log CESS loss for various models on test set.

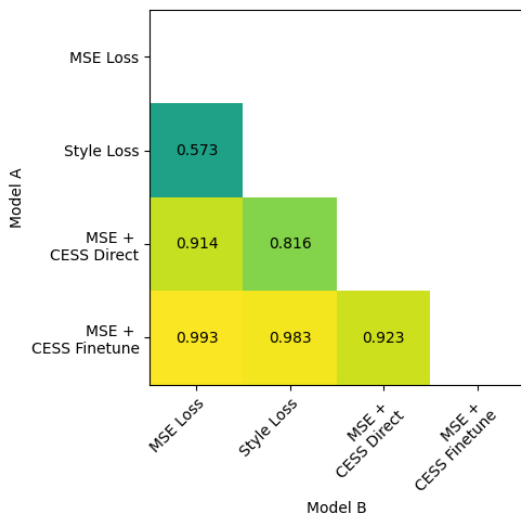


Figure 11. Pairwise comparison of models on the sample level. The number in each cell represents the percentage of test set samples where Model A had lower CESS than Model B.

Style Loss; this can be seen by the high values in the bottom left 2x2 grid. The model finetuned on CESS also significantly outperformed the model directly trained with a linear combination, as seen by the high value (0.923) in the 4th row, 3rd column.

4.2. Qualitative Results

Some qualitative analysis was also done on reconstructions of random samples from each category. An example from the AngEllipse class can be seen in Fig. 12; the rest can be seen in Supplementary Material 1.2. In general, the model trained purely with MSE seems to preserve fine structure the worst, producing blurry reconstructions and often combining grains together into blurry blobs. Models trained with CESS seem to better preserve fine detail like grain structure and size, with the best model being the one finetuned with CESS with a 50x50 crop. The model trained with style loss

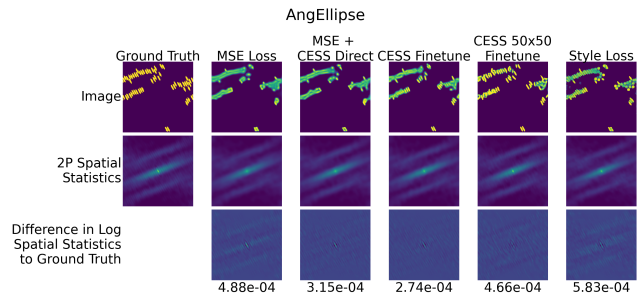


Figure 12. Reconstructions from models of a sample from the AngEllipse class. The first row shows the original microstructure, the second row shows the 2-point autocorrelation for phase 1, and the third row shows the difference in log spatial statistics to ground truth.

is also effective at preserving fine structure qualitatively, although 2-point spatial statistics of its reconstructions are sometimes further from ground truth than reconstructions from the MSE model, suggesting some tradeoff in accuracy in spatial statistics. More qualitative analysis for some reconstructions from each class can be found in Supplementary Material 1.3.

5. Conclusion

We have demonstrated the successful use of CESS, a loss which explicitly focuses on 2-point spatial statistics, as a reconstruction loss in a VAE. We have also demonstrated some preliminary positive qualitative results, including reconstructions that seem to preserve fine detail better than a model trained with only MSE. Future work could include quantifying the effectiveness of the latent space through its use in a property prediction model, or modifying the loss to address current shortcomings, like poor convergence on sparse microstructures.

We demonstrate compression from 224x224 binary microstructure images to 14x14 latent representations. Future uses of the latent space could be as a proxy for microstructures in current material design applications, for instance in property prediction, or in a Latent Diffusion [17] like model, where generation is done in the latent space.

References

- [1] O. B. Bohan, “madebyollin/taesd,” Jan. 2025, original-date: 2023-04-16T22:55:55Z. [Online]. Available: <https://github.com/madebyollin/taesd> 4
- [2] R. Bostanabad, A. T. Bui, W. Xie, D. W. Apley, and W. Chen, “Stochastic microstructure characterization and reconstruction via supervised learning,” *Acta Materialia*, vol. 103, pp. 89–102, Jan. 2016. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1359645415007259> 3

- [3] R. Chowers and Y. Weiss, “What do cnns learn in the first layer and why? a linear systems perspective,” 2023. [Online]. Available: <https://arxiv.org/abs/2206.02454> 3
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255, iSSN: 1063-6919. [Online]. Available: <https://ieeexplore.ieee.org/document/5206848>
- [5] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trounev, and G. Peyré, “Interpolating between Optimal Transport and MMD using Sinkhorn Divergences.”
- [6] D. T. Fullwood, S. R. Niezgoda, and S. R. Kalidindi, “Microstructure reconstructions from 2-point statistics using phase-recovery algorithms,” *Acta Materialia*, vol. 56, no. 5, pp. 942–948, Mar. 2008. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1359645407007458> 1, 3
- [7] L. A. Gatys, A. S. Ecker, and M. Bethge, “A Neural Algorithm of Artistic Style,” Aug. 2015. [Online]. Available: <https://arxiv.org/abs/1508.06576v2>
- [8] E. O. Hall, “The deformation and ageing of mild steel: Iii discussion of results,” *Proceedings of the Physical Society. Section B*, vol. 64, no. 9, p. 747, sep 1951. [Online]. Available: <https://dx.doi.org/10.1088/0370-1301/64/9/303> 1
- [9] S. S. Hashemi, M. Guerzhoy, and N. H. Paulson, “Toward Learning Latent-Variable Representations of Microstructures by Optimizing in Spatial Statistics Space,” Feb. 2024, arXiv:2402.11103 [cs]. [Online]. Available: <http://arxiv.org/abs/2402.11103> 2
- [10] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013. [Online]. Available: <https://arxiv.org/abs/1312.6114> 3
- [11] W. Lenz, “Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern,” *Z. Phys.*, vol. 21, pp. 613–615, 1920. [Online]. Available: <https://cds.cern.ch/record/460663>
- [12] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the Loss Landscape of Neural Nets.”
- [13] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” 2021. [Online]. Available: <https://arxiv.org/abs/1908.03265> 4
- [14] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016. 4
- [15] N. H. Paulson, M. W. Priddy, D. L. McDowell, and S. R. Kalidindi, “Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics,” *Acta Materialia*, vol. 129, pp. 428–438, May 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S135964541730188X> 2, 3
- [16] A. E. Robertson, A. P. Generale, C. Kelly, M. O. Buzzy, and S. R. Kalidindi, “MICRO2D: A Large, Statistically Diverse, Heterogeneous Microstructure Dataset,” *Integrating Materials and Manufacturing Innovation*, vol. 13, no. 1, pp. 120–154, Mar. 2024. [Online]. Available: <https://doi.org/10.1007/s40192-023-00340-4>
- [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. 6
- [18] A. Sardeshmukh, S. Reddy, B. P. Gautham, and P. Bhattacharyya, “Material Microstructure Design Using VAE-Regression with Multimodal Prior.” Feb. 2024, arXiv:2402.17806 [cond-mat, stat]. [Online]. Available: <http://arxiv.org/abs/2402.17806>
- [19] A. Sardeshmukh, S. Reddy, B. Gautham, and P. Bhattacharyya, “TextureVAE: Learning Interpretable Representations of Material Microstructures Using Variational Autoencoders.” in *AAAI Spring Symposium: MLPS*, 2021. 2, 3
- [20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” *CoRR*, vol. abs/1801.03924, 2018. [Online]. Available: <http://arxiv.org/abs/1801.03924> 3