MozzaVID: Mozzarella Volumetric Image Dataset

Supplementary Material

1. Train/validation/test splitting strategy

The splitting strategy for the mozzaVID dataset is tailored for the specific requirements of both classification targets, resulting in two separate approaches.

The coarse target is designed to explore high-level structural differences induced by the cheese recipe parameters. However, it is assumed that volumes from the same scan, or the same sample, are physically dependent due to the proximity of the imaged structures. Thus, the splitting was done on a sample level to ensure independence between volumes in each split. Since there are only 6 samples per scan, the train/val/test split was done in ratios of approximately 67%, 16%, and 16%, respectively.

In contrast, models trained on the 150 classes of the fine target are expected to exploit and evaluate the degree of the aforementioned proximity-based structural dependence. Because of that, the splits for fine target were simply performed on individual volume level, with a standard ratio of 70%, 20%, and 10% for train, validation, and test, respectively.

2. Data visualization

2.1. Scans

In Fig. 2, we introduce a set of example scan slices that illustrate the structural variability across the cheese samples. Subsequently, in Fig. 4c, we use slices from all scans to explore the representation learned by one of the models. To provide a clean and comprehensive overview of the variability, the same slices are arranged in an ordered grid in Fig. 8.

To further supplement the overview of the scanned samples, Fig. 9 showcases example slices from different fine-grained classes. These are organized into sets of six samples originating from the same cheese type, emphasizing their structural similarity and consequent increase in the complexity of the problem. However, it is important to note that a single 2D slice may not fully capture the sample's structural characteristics, which are likely to be more effectively discerned through a full 3D representation.

2.2. Metadata

In Sec. 3.1, we provide an overview of the metadata, including experimental design parameters, which are later visualized in a reduced form in Fig. 6.

The following PCA reduction was applied to three parameters: rotor speed, temperature, and additive type. The additive type is a categorical variable with three categories (None, CaCL2, and Citric Acid), to which a unique number is assigned. Each parameter is then normalized to maintain the confidentiality of the exact recipe. The resulting values are presented in Fig. 5, normalized to a zero mean and standard deviation of one.

3. Experiments

3.1. Top-k accuracy

In Tab. 3, top-1 accuracy is provided as a primary evaluation of model performance. To provide further context, top-k accuracy scores are also listed in Tab. 4 (top-2 for coarse granularity and top-5 for fine granularity).

For both coarse and fine models, there is a significant increase in accuracy compared to Tab. 3, with Large 3D models approaching near-perfect performance. This indicates that the models can easily choose a set of the most probable classes, but may struggle to discriminate between a few most similar neighbors.

Importantly, the examination of confusion matrices for both granularities shows no systematic trends or recurring pairs of misclassified classes, suggesting that the space of investigated structures is sufficiently broad and relatively uniformly distributed.

3.2. Learning rate fine tuning

In Sec. 4, we outline the experimental design, including the investigated models and ablation studies. The training setup across all models is standardized to ensure a fair comparison of the models. However, certain hyperparameters should be fine-tuned to provide the most representative result. In this study, we focus on fine-tuning the learning rate, as it has the most significant impact on the consistency of model performance. Given the training and convergence constraints, fine-tuning was performed only for 2D models and 3D-fine models, while the remaining models used a default learning rate of 10^{-4}.

Fine-tuning was conducted using the *Weights & Biases* parameter sweep over a log-uniform distribution in the range $lr \in < 10^{-2}, 10^{-6} >$, using the Bayesian optimization. Each model was tested with 30 hyperparameter variants, provided this could be achieved in 1 day of training. Otherwise, the number of tested variants was lowered to 15. The best configuration was selected based on the highest validation accuracy. The final learning rates for all models are listed in Tab. 5.

Although not all models underwent fine-tuning, those that did were also the ones most likely to benefit from it. The performance of the tested models tended to be more

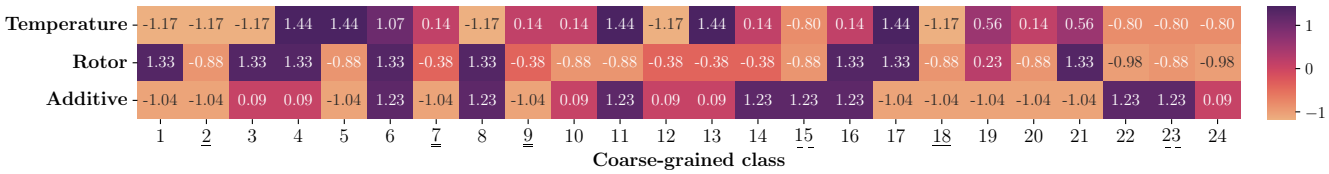


Figure 5. Overview of the variation in the normalized experimental design parameters of the first 24 cheese types (coarse-grained classes). Data for class 25 (Cagliata) is not available. Underlines highlight three pairs of cheese produced with the same set of parameters.

Table 4. Top-k accuracy of the trained models. Top-2 accuracy is reported for the coarse granularity and top-5 for the fine granularity, based on the total class count in both targets.

Granularity	Coarse, top-2						Fine, top-5			
Split	Small		Base		Large		Base		Large	
Dimensionality	**2D**	**3D**	**2D**	3D	**2D**	3D	**2D**	3D	**2D**	3D
ResNet50	0.567	0.876	0.883	0.991	0.832	0.997	0.853	0.991	0.983	0.999
MobileNetV2	0.660	0.901	0.888	0.951	0.910	0.972	0.783	0.996	0.985	0.989
ConvNeXt-S	0.516	0.714	0.787	0.735	0.787	0.908	0.867	0.946	0.914	0.996
ViT-B/16	0.423	0.516	0.602	0.915	0.594	0.866	0.566	0.964	0.888	0.997
Swin-S	0.557	0.814	0.800	0.972	0.787	0.958	0.889	0.993	0.988	0.999
Average	0.545	0.764	0.792	0.920	0.782	0.940	0.792	0.978	0.952	0.996

Table 5. Learning rate of the trained models on all investigated setups after fine-tuning. Base and Large 3D models were assigned a default learning rate due to the limited resources and slow convergence.

Granularity	Coarse						Fine			
Split	Small		Base		Large		Base		Large	
Dimensionality	**2D**	**3D**	**2D**	3D	**2D**	3D	**2D**	3D	**2D**	3D
ResNet50	9.9×10^{-4}	7.3×10^{-5}	1.1×10^{-3}	10^{-4}	2.8×10^{-3}	10^{-4}	9.0×10^{-4}	10^{-4}	3.3×10^{-4}	10^{-4}
MobileNetV2	2.1×10^{-3}	5.8×10^{-4}	1.6×10^{-3}	10^{-4}	3.3×10^{-3}	10^{-4}	1.9×10^{-3}	10^{-4}	6.0×10^{-4}	10^{-4}
ConvNeXt-S	3.4×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	10^{-4}	1.6×10^{-3}	10^{-4}	1.4×10^{-3}	10^{-4}	1.0×10^{-3}	10^{-4}
ViT-B/16	1.8×10^{-5}	6.0×10^{-5}	4.1×10^{-5}	10^{-4}	1.6×10^{-5}	10^{-4}	3.4×10^{-6}	10^{-4}	5.0×10^{-5}	10^{-4}
Swin-S	1.8×10^{-4}	5.4×10^{-5}	1.2×10^{-4}	10^{-4}	2.9×10^{-5}	10^{-4}	2.4×10^{-6}	10^{-4}	8.8×10^{-5}	10^{-4}

volatile and sensitive when trained on the smaller dataset variants. In the case of 2D-Small models, small learning rate changes often resulted in a 20 to 30 percentage point difference in accuracy, while for 2D-Large models, this difference was limited to approximately 1 to 5 percentage points. This behaviour suggests that the performance of 3D models, even without fine-tuning, remains representative and close to optimal. Furthermore, the strong performance of non-fine-tuned 3D models only strengthens the reported impact of the 3D representation and all conclusions drawn from it.

3.3. PCA visualization

In Sec. 4.3, we introduce the outline of the learned representation analysis experiment. There, a PCA-based compression of the sample experimental design parameter space is described, together with an introduction of a colormap that is later used for the interpretation of the UMAP representation in Fig. 4b. The visualization of the PCA space, together with the colormap overlay, can be found in Fig. 6.

3.4. Rotation ablation study

In Sec. 6.1, we discuss structure orientation as a potential source of bias, noting that while sample preparation and data augmentations largely mitigate this issue, it is not systematically addressed in the raw data. To evaluate the impact of volume orientation on the accuracy of the investigated models, we conducted an additional ablation study using a modified set of transforms.

This study focused on the ResNet50 model, given its consistent performance, and utilized the Base dataset as a central, representative example. Both 2D and 3D models were trained for coarse-grained and fine-grained tasks. Apart from the data augmentation, the training setup was exactly the same as in the original experiments.

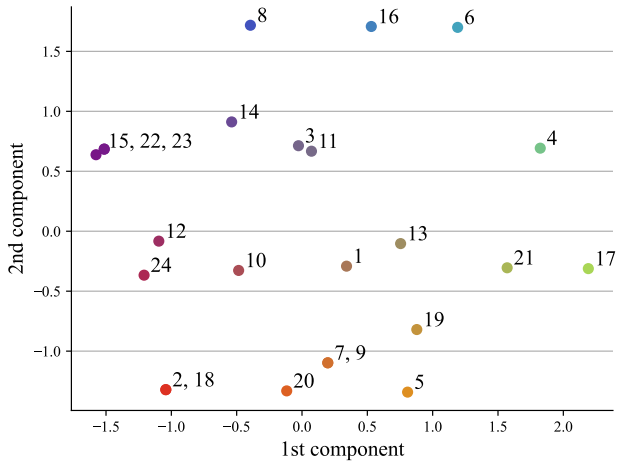


Figure 6. PCA of the experimental design parameters used to define the mozzarella cheese variations. Class 25 (Cagliata cheese) is omitted – it was prepared separately, and its exact recipe is not available. Coloring of the points serves as a basis for visual analysis of structural similarities learned by the models (Fig. 4b).

Table 6. Results of rotation ablation study using the ResNet50 model and the Base dataset instance. Reference results copied from Tab. 3.

Granularity	Coarse		Fine	
Dimensionality	2D	3D	2D	3D
Reference	0.741	0.957	0.563	0.683
With rotations	0.747	0.945	0.498	0.643

The modified set of transforms included the following elements:

1. Normalization.
2. Random 90° rotation ($p = 0.5$).
3. Random flipping in X and Y axis ($p = 0.5$).
4. Random rotation in a −30° to 30° range ($p = 0.5$).

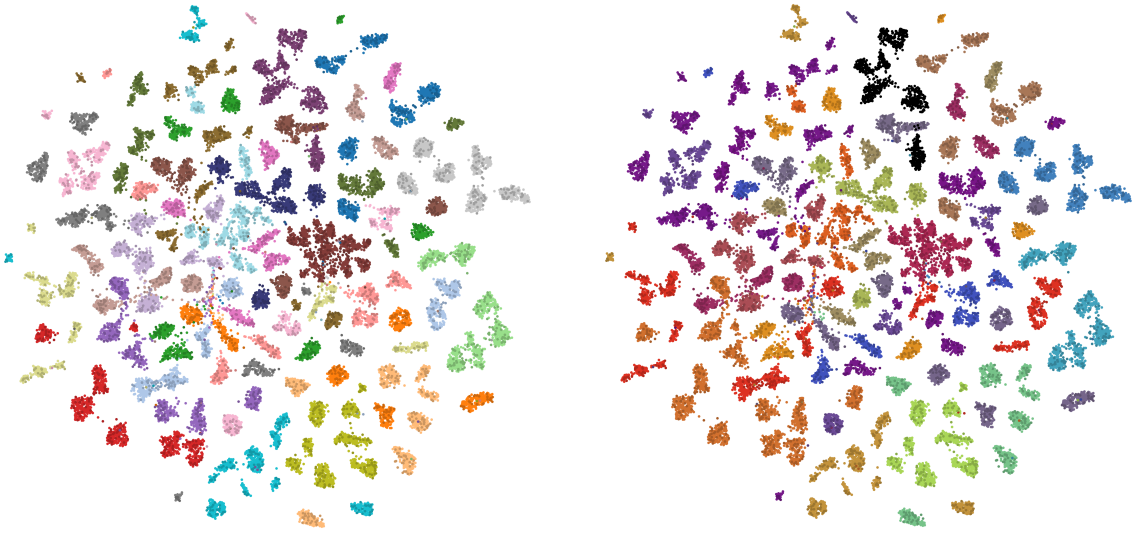
Elements 1 and 3 were retained from the original pipeline. The combined set of transforms covers most of the possible structure orientations while minimizing information loss at the most extreme angles.

The results of the study (Tab. 6) suggest that volume orientation does not exhibit a clear or consistent impact on training accuracy. The coarse-grained models seem to perform similarly with the new setup. The impact on fine-grained models is more negative compared to coarse-grained models, which may be due to fine-grained models relying more heavily on structure orientation, as each sample (fine-grained class) is cut in a single direction, with variation introduced only through data augmentation. However, this effect is not significant enough to undermine the validity of the dataset or the presented results.

3.5. Fine-grained model UMAP

A similar experiment to the one described in Sec. 4.3 and visualized in Fig. 4 was conducted using the best-performing fine-grained model. The resulting UMAPs are shown in Fig. 7, though the slice-based visualization is omitted here for readability. The colormaps remain consistent with those used in the coarse-grained analysis.

Despite a different target, in most cases the network still groups samples from the same coarse-grained class in close proximity (Fig. 7a). This behavior indicates that structural similarities within these samples significantly influence the embedding space constructed by the model. As shown in Fig. 7b, the alignment with PCA space appears similar or even stronger compared to the coarse-grained model. The UMAP reveals four distinct zones with purple samples in the top-left, orange/red in the bottom-left, green in the bottom-right, and blue on the right. While some exceptions are present, these can often be explained by structural properties that deviate from the PCA parameters, as discussed in Sec. 5.2. This layout suggests that the fine-grained model captures a more nuanced representation of the structural variability, which may result from the need to detect subtler differences between samples and the absence of regularizing constraints imposed by the 25 coarse-grained classes.



(a) Clusters colored by coarse-grained class.

(b) Clusters colored by experimental design PCA color space. Class 25 is colored black.

Figure 7. UMAP generated from second-to-last layer feature representations of the best-performing model in the fine-grained classification task (ResNet50 trained on the Large dataset). Reduction parameters: `n_neighbors=30, min_dist=0.8`.

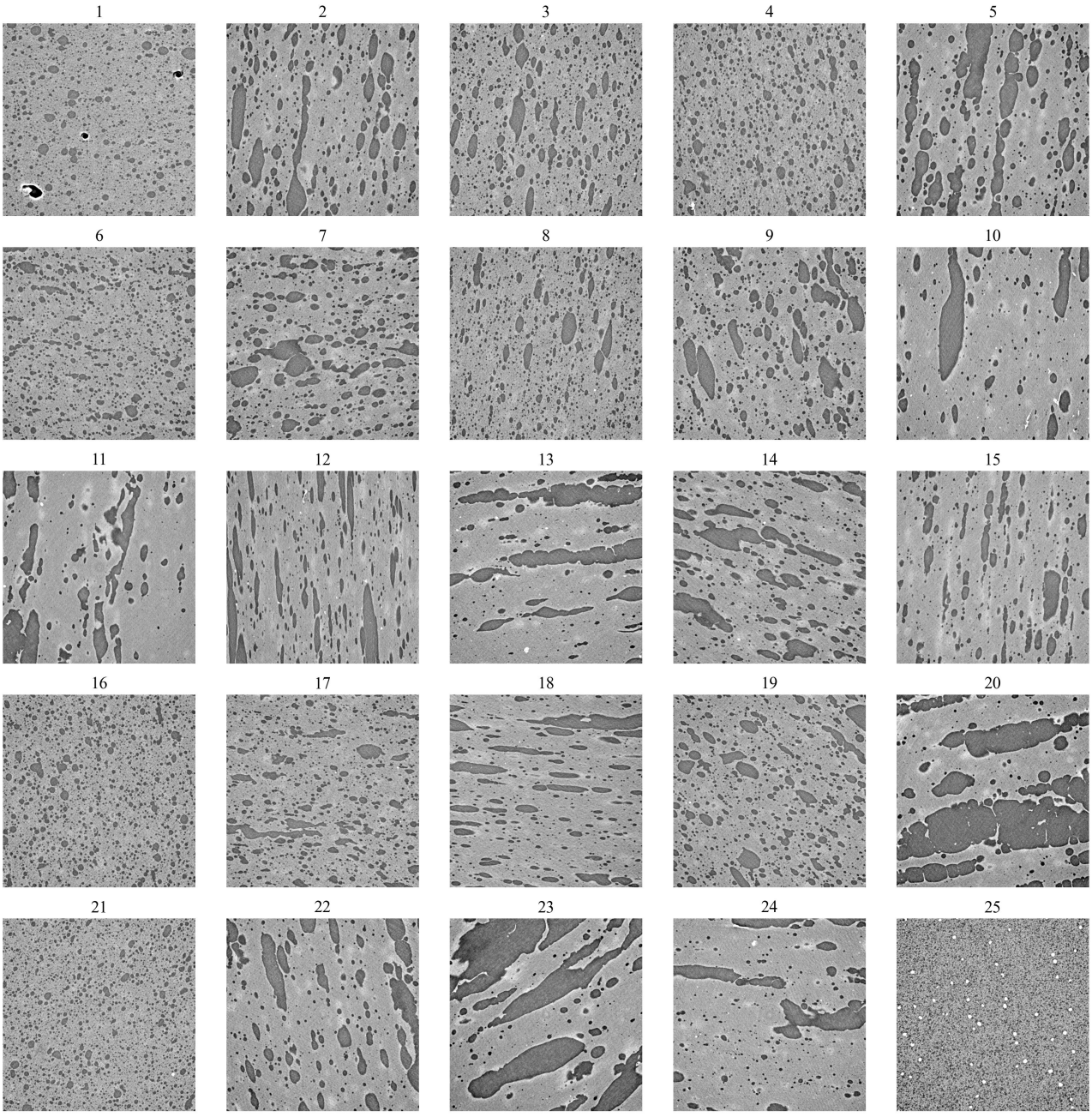


Figure 8. Overview of slices from each cheese type, forming the 25 coarse-grained classes.

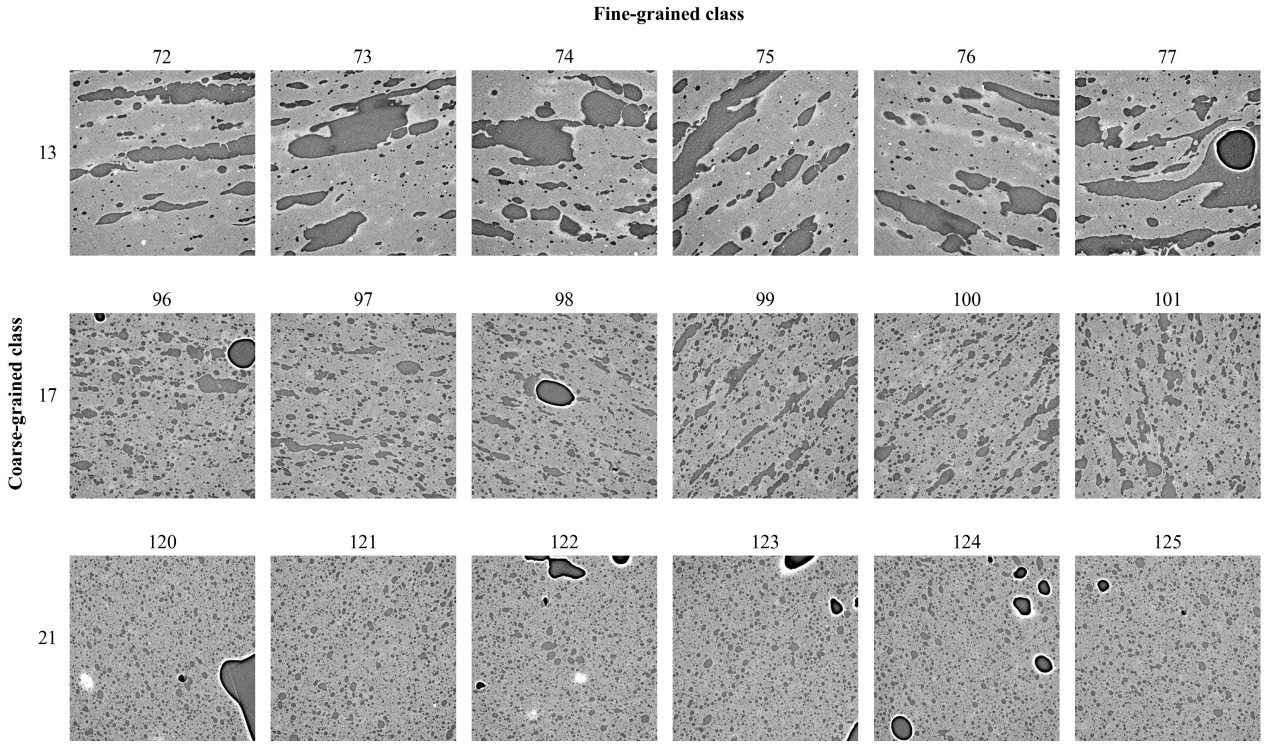


Figure 9. Example slices from the fine-grained classes. Each row represents a set of six samples from one cheese type (coarse-grained class), forming six consecutive fine-grained classes.