Table 4. **Known and Novel Classes selected from the MedMNIST Benchmark**

Datasets	**Blood**	**Path**	**Derma**	**OCT**	**Tissue**	**OrganA,C,S**
Known Classes	$1-5,7$	$1-5,7$	$1,3-6$	$1,2,4$	$1-2,4-5,7-8$	$1,5-11$
Novel Classes	$6,8$	$6,8,9$	$2,7$	3	$3,6$	$2,3,4$

Table 5. **Ablation Study on the Choice of Inlier/Outlier Specification on the ISIC2019 Benchmark.** We report the average AUROC (%) scores across modality shifts and semantic novelty detection.

Pix. In	**Lat. In**	**Pix. In + Pix. Out**	**Pix. In + Lat. Out**	**Lat In. + Pix. Out**
71.3	72.2	77.3	61.5	$\mathbf{91.8}$

Supplementary Material - Exploring Inlier and Outlier Specification for Improved Medical OOD Detection

A. Dataset Descriptions

MedMNIST Benchmark. (i) Blood MNIST consists of $17,092$ human blood cell images collected from healthy individuals corresponding to 8 different classes; (ii) Path MNIST is a histology image dataset of colorectal cancer with $107,180$ samples of non-overlapping, hematoxylin and eosin stained image patches from 9 different classes; (iii) Derma MNIST is a skin lesion dataset curated from the HAM1000 (Tschandl *et al.*) database. It contains a total of $10,015$ images across 7 cancer types; (iv) Oct MNIST contains $109,309$ optical coherence tomography (OCT) retinal images corresponding to 4 diseases; (v) Tissue MNIST is a kidney cortex image dataset curated from the Broad Bioimage Benchmark Collection with $236,386$ images from 8 classes; (vi) (vii) (viii) Organ(A,C,S) MNIST are images of abdominal CT collected from the Axial, Coronal and Sagittal planes of 3D CT images from the Liver-tumor segmentation benchmark. The datasets contain $58,850$, $23,660$ and $25,221$ images across 11 classes respectively. For each of MedMNIST datasets, we consider the validation splits from all remaining datasets for evaluating modality shift detection performance. On the other hand, we use a subset of classes held-out during training to evaluate the novel class detection.

ISIC 2019 and NCT Datasets. For these two benchmarks, the following datasets were used to evaluate OOD detection performance. For each OOD dataset, we highlight if its a modality shift (M) or a semantic shift/novel class (S) along with the corresponding ID dataset (ISIC or NCT) with which the model was trained:- (i) Camelyon-17 (WILDS) [1](M:ISIC, S: NCT) is a histopathology dataset of tumor and non-tumor breast cells with approximately 450K images curated from five different medical centers. We randomly sample 3000 examples from the dataset for OOD detection; (ii) Knee (M: ISIC, M: NCT) Osteoarthritis severity grading dataset contains X-ray images of knee joints with examples corresponding to arthritis progression. We used 825 examples chosen randomly from the dataset for evaluation; (iii) CXR (M: ISIC, M: NCT)[*] is a chest X-ray dataset curated from the MIMIC-CXR database containing $1,083$ samples corresponding to disease states namely normal, pneumonia and congestive heart failure and (iv) Retina (M: ISIC, M: NCT) is a set of 1500 randomly chosen retinal images with different disease progressions from the Diabetic Retinopathy detection benchmark from Kaggle[*]; (v) Clin Skin (S: ISIC) contains 723 images of healthy skin [22]; (vi) Derm-Skin (S: ISIC) consists of 1565 dermoscopy skin images obtained by randomly cropping patches in the ISIC2019 database [22]; (vii) NCT 7K (S: NCT) contains 1350 histopathology images of colorectal adenocarcinoma with no overlap with NCT [22]. In addition, we use 2000 randomly chosen examples from ISIC as a source of modality shift for the detector trained on NCT and vice-versa. Moreover, in both cases, novel classes unseen while training are also used to evaluate detection under semantic shifts.

B. Ablation Study

In addition to the existing baseline methods, we performed an ablation study on the ISIC-2019 benchmark to compare other choices of inlier and outlier specification. As showed in Table 5, we observe that the inclusion of the latent space inliers (Lat. In) alone during the optimization is not sufficient to obtain high quality OOD detectors. This is due to the fact that optimizing to minimize the energy scores for the latent inliers in fact leads to over-generalization and cannot effectively demarcate decision boundaries between inliers and outliers. In practice, such over-generalization can even affect ID accuracy. In order to circumvent this issue, we propose the use of diverse, pixel-space outlier samples. Moreover, as argued in the paper, (i) using pixel-space inliers without any outlier synthesis leads to inferior OOD detection performance, and (ii) pixel-space inliers + pixel-space outliers is consistently better than pixel-space inliers+ Latent-space outliers, further emphasizing the

[*] https://github.com/cxr-eye-gaze/eye-gaze-dataset

[*] https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data

Table 6. **AUPR scores for novel class detection on the MedMNIST benchmark.** We report the AUPR (Input) scores using different approaches with a $40-2$ WideResNet backbone.

In Dist.	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Blood	47.89	25.02	21.98	37.44	35.72	73.32
Path	20.29	12.33	36.23	11.74	26.61	30.35
Derma	79.36	75.18	82.4	57.58	80.69	82.28
OCT	53.56	57.35	63.1	61.39	76.62	79.45
Tissue	55.53	48.37	39.75	55.36	66.96	87.79
OrganA	86.82	42.48	51.72	35.7	68.22	95.51
OrganS	72.61	31.99	51.81	72.56	81.69	89.32
OrganC	73.48	38.95	47.06	68.08	71.79	95.16

Table 7. **Full results for modality shift detection on the BloodMNIST dataset (ID) using the $40-2$ WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Path	88.0/57.4	71.6/38.8	67.9/33.2	75.7/36.2	97.6/82	99.0/96.5
Derma	89.3/73.8	69.6/68.8	70.1/78.2	97.4/97.3	83.6/86.2	98.8/99.1
OCT	96.7/89.3	98.2/95.1	95.8/82.6	100/100.0	95.8/77.3	100.0/99.9
Tissue	98.8/73.4	99.2/98.8	98.4/89.6	100/100.0	98.9/93.6	100/100.0
OrganA	99.4/84.0	95.9/89.3	86.7/71.4	100/100.0	98.2/94.3	100.0/99.9
OrganC	99.3/90.6	96.1/95.0	84.8/81.4	100/100.0	98.2/97.2	100.0/99.9
OrganS	99.5/92.2	95.5/93.8	85.6/81.6	100/100.0	98.6/97.7	100.0/99.9

Table 8. **Full results for modality shift detection on the PathMNIST dataset (ID) using the $40-2$ WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Blood	92.1/94.4	99.3/99.7	79.4/92.8	89.7/97.7	64.9/91.0	95.3/99.2
Derma	72.8/95.7	74.3/94.7	39.8/85.1	91.0/98.2	43.6/87.4	98.1/99.7
OCT	91.8/95.9	69.2/70.3	86.6/84.3	98.1/97.8	79.7/81.8	99.7/99.6
Tissue	73.4/96.4	67.7/66.0	72.0/70.8	95.8/93.0	71.5/63.6	100.0/99.9
OrganA	76.7/83.4	72.9/75.9	72/66.7	99.6/99.7	52.7/60.7	100.0/100.0
OrganC	76.0/92.7	78.0/89.9	71.5/85.4	99.4/99.8	56.7/83.4	99.8/99.9
OrganS	75.1/90.9	81.4/91.4	75.4/86.5	99.4/99.8	58.4/83.4	99.8/99.9

effectiveness of diverse pixel-space outliers. Finally, synthesizing inliers and outliers from the latent space is not feasible, since both approaches sample from the tails of the class-conditioned distributions.

Table 9. **Full results for modality shift detection on the DermaMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Blood	87.4/86.9	77.0/77.6	90.0/88.0	85.2/71.1	80.7/80.1	91.8/89.9
Path	82.2/66.1	72.7/44.5	82.7/57.9	90.4/46.2	92.5/72.0	89.6/56.1
OCT	79.0/49.8	72.6/49.0	81.4/73.6	100.0/99.5	55.1/35.4	99.5/95.4
Tissue	57.2/43.9	84.2/52.2	87.3/66.7	99.9/96.7	78.3/54.9	99.8/96.4
OrganA	79.2/76.2	49.9/20.7	85.9/73.5	97.3/83.6	85.3/66.5	98.2/90.2
OrganC	78.2/85.5	47.4/33.1	85.0/85.0	96.9/90.8	83.8/76.3	98.5/96.2
OrganS	78.2/85.4	44.8/31.7	85.2/80.7	96.6/90.3	84.0/76.4	99.1/98

Table 10. **Full results for modality shift detection on the OctMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Blood	97.4/94.7	60.5/91.2	75.2/96.1	97.5/99.5	97.0/99.6	100/100
Path	99.1/68.4	49.0/55.0	67.4/81.7	98.3/98.4	96.8/97.6	100/100.0
Derma	96.4/99.3	52.1/89.2	66.9/96.4	99.7/100.0	99.4/99.9	100/100
Tissue	78.0/42.3	40.9/25.4	62.0/64.2	54.8/40.8	90.9/80.8	97.5/95.0
OrganA	97.3/48.4	50.3/67.4	70.1/86.3	99.8/99.8	88.2/92.4	99.9/99.9
OrganC	98.2/72.9	48.2/81.2	66.8/92.6	99.7/99.9	94.1/98.5	100.0/100.0
OrganS	98.3/70.9	49.8/80.6	67.5/92.4	99.8/99.9	94.3/98.6	100.0/100.0

Table 11. **Full results for modality shift detection on the TissueMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Blood	93.9/99.4	68.0/97.8	54.9/95.9	100/100	99.6/100.0	100/100
Path	93.8/98.0	83.7/95.3	64.4/87.1	99.9/100.0	97.7/99.0	100/100
Derma	91.0/99.1	84.3/99.2	74.9/98.5	99.2/100.0	98.9/99.9	100.0/100
OCT	20.9/54.9	46.8/75.2	25.0/64.0	13.7/49.8	11/49.0	77.6/89.5
OrganA	97.7/99.5	75.5/91.8	69.5/90.6	86.1/94.4	63.3/88.7	99.6/99.9
OrganC	97.1/99.6	76.4/97	67.2/95.7	84.3/97.3	62.2/94.9	99.6/100.0
OrganS	97.2/99.6	75.3/96.6	65.3/95.2	84.4/97.2	58.7/94.1	99.6/99.9

C. Details of Experiments in the Main Paper

Dataset Preprocessing. We first split each of the datasets into two categories namely (i) data from classes known during training (*known classes*) and (ii) data from classes unknown while training (*novel classes*) where the latter constitutes OOD data with semantic shifts. The dataset from the former category is split in the ratio of $90:10$ for training and evaluating the predictive models. Table 4 provides the list of known and novel classes for the MedMNIST benchmark. In case of

Table 12. **Full results for modality shift detection on the OrganAMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	**VOS**	**VOS++**	**NDA**	**NDA++**	**Ours**
Blood	99.4/98.2	64.7/79.8	65.7/81.7	32.4/72.7	100.0/100.0	100.0/100.0
Path	99.5/91.4	67.7/49.8	67.3/50.8	83.1/66.3	99.4/98.8	99.2/98.9
Derma	96.4/99.4	76.9/89.5	76.0/91.2	78.5/92.1	99.6/99.9	99.4/99.9
OCT	88.2/98.4	63.9/37.2	92.4/74.8	70.6/47.1	93.2/88.1	99.9/99.8
Tissue	94.4/90.6	95.2/64.3	87.5/45.3	86.3/41.8	88.8/66.2	100/100.0

Table 13. **Full results for modality shift detection on the OrganSMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	**VOS**	**VOS++**	**NDA**	**NDA++**	**Ours**
Blood	92.8/64.2	49.4/66.9	64.7/77.2	91.6/93.8	100.0/100.0	99.9/99.9
Path	97.2/34.1	41.3/23.4	55.5/40.1	90.7/68.5	97.4/90.5	99.0/97.1
Derma	92.5/98.3	34.6/59.2	70.1/76.0	97.9/98.6	100.0/100.0	99.2/99.6
OCT	80.3/99.1	52.8/21.9	46.1/25.0	90.6/80.4	72.6/39.4	92.8/77.4
Tissue	93.6/90.6	79.4/18.6	74.4/18.0	99.3/93.2	94.7/76.2	100.0/99.8

Table 14. **Full results for modality shift detection on the OrganCMNIST dataset (ID) using the** $40-2$ **WideResnet (AUROC/AUPR metrics)**

OOD Data	Methods					
	G-ODIN	**VOS**	**VOS++**	**NDA**	**NDA++**	**Ours**
Blood	96.6/89.8	56.0/63.6	41.0/60	86.6/88.8	100.0/100.0	99.6/99.8
Path	98.8/59.4	53.2/23.1	71.9/36.5	98.1/93.1	99.8/99.2	99.7/99.4
Derma	97.8/95.5	53.2/65.3	68.2/77.2	96.2/97.2	99.7/99.8	98.7/99.3
OCT	96.2/87.3	59.5/19.6	79.0/34.6	92.0/66.5	91.0/65.4	98.2/96.0
Tissue	83.9/66.1	61.0/10.3	62.5/9.8	93.4/48.9	80.6/30.3	99.4/97.4

ISIC2019, we choose BKL, VASC and SCC as novel classes while MUC, BACK and NORM are chosen as novel classes for the NCT(Colorectal) benchmark.

Training Details.

Estimating Class-specific Means and Joint Covariance: We estimate the means and joint covariance via maximum likelihood estimation during training, similar to [7]. We employ K queues each of size 1000 where each queue is filled during every iteration until their pre-specified capacities with the class specific latent embeddings (extracted from the penultimate layer) of the training data. We then adopt an online strategy to update the queues such that they contain much higher quality embeddings of the data as the training progresses. In particular, we enqueue one class-specifc latent embedding to the respective queues while dequeuing one embedding from the same class.

Sampling the Latent Space: In practice, we select samples close to the class specific boundaries based on the n^{th} smallest likelihood ($n = 64$) among N examples ($N = 10,000$) synthesized from the respective Gaussian distributions.

General Hyperparameters: We train the $40-2$ WideResNet and ResNet-50 architectures for 100 and 50 epochs with learning rates of $1e{-}3$ and $1e{-}4$ respectively. We reduce the learning rate by a factor of 0.5 every 10 epochs using the Adam optimizer with a momentum of 0.9 and a weight decay of $5e^{-4}$. We choose a batch size of 128 for datasets from MedMNIST and 64

Table 15. **Evaluation on the ISIC2019 benchmark.** We report AUPR scores obtained with a ResNet-50 model trained on the ISIC2019 dataset. Note, we show results for both novel classes (blue), and modality shifts (red). In each case, the first and second best performing methods are marked in green and orange respectively.

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Novel Classes	53.07	65.3	56.61	49.89	53.02	67.71
Clin Skin	85.94	77.45	91.68	81.68	85.95	92.92
Derm Skin	81.18	84.95	85.04	84.23	89.72	95.56
Wilds	74.42	66.18	54.24	78.47	83.48	99.72
Colorectal	79.5	48.11	85.86	76.79	79.93	98.99
Knee	88.37	58.71	94.08	94.91	98.08	97.83
CXR	87.62	87.39	90.12	88.96	74.77	98.5
Retina	99.08	83.57	92.24	77.86	80.57	99.87

Table 16. **Evaluation on the colorectal cancer benchmark.** We report AUPR scores obtained with a ResNet-50 model trained on the the colorectal cancer dataset [16]. Note, we show results for both novel classes (blue) and modality shift detection (red).

OOD Data	Methods					
	G-ODIN	VOS	VOS++	NDA	NDA++	Ours
Novel Classes	46.55	97.69	93.43	96.48	95.38	99.31
NCT 7K	85.15	96.75	93.19	97.49	94.26	99.54
WILDS	46.54	98.15	96.51	80.29	94.29	98.4
ISIC2019	78.49	88.33	94.55	99.67	84.88	99.96
Knee	98.53	99.59	95.95	99.76	94.88	100.0
CXR	98.30	99.92	94.58	99.98	94.31	99.99
Retina	99.92	97.35	99.05	99.96	91.64	100.0

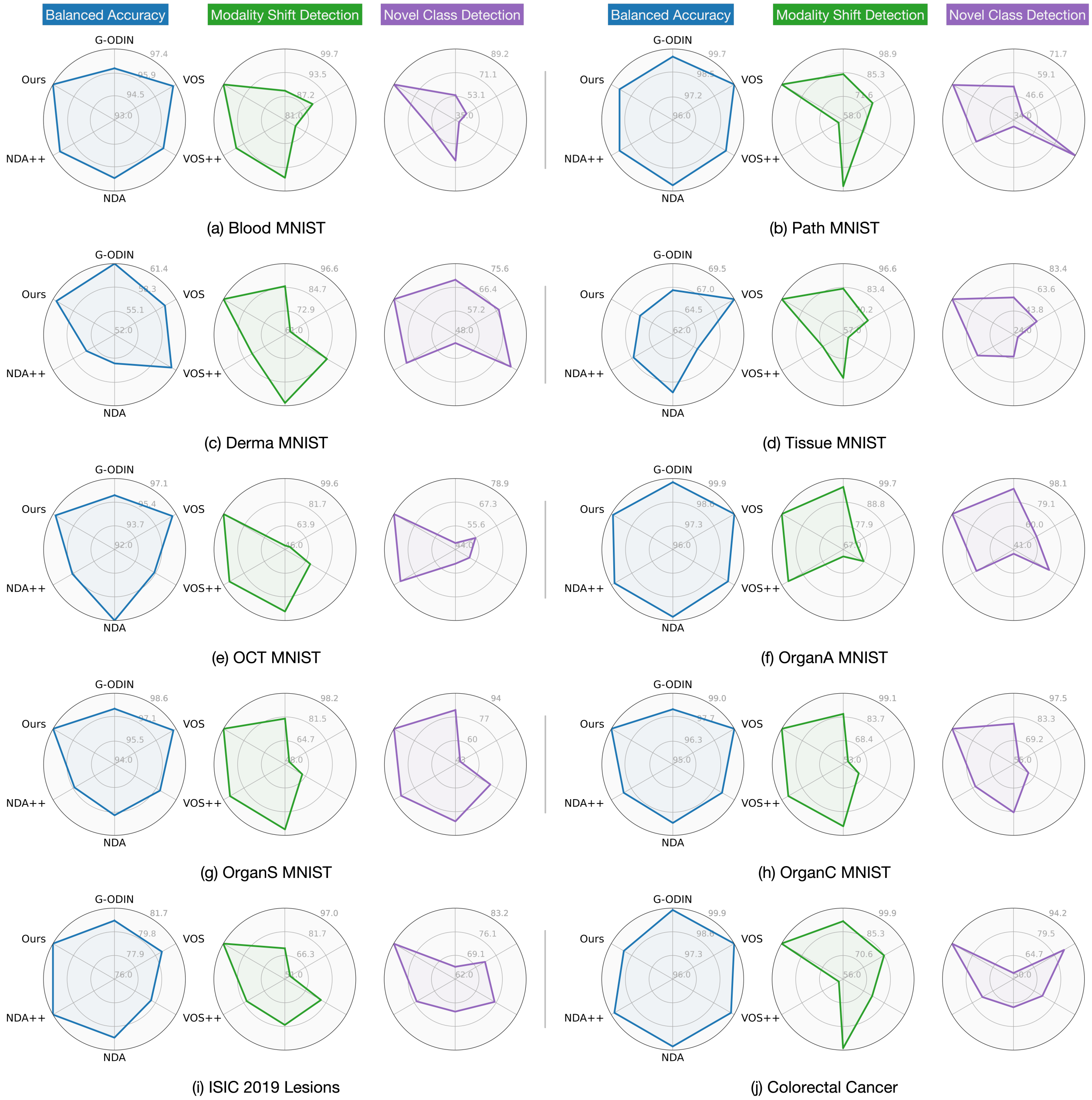
for the full-sized images. For all experiments including the baselines (except G-ODIN), we use a margin $m_{\text{ID}} = -20$ and $m_{\text{OOD}} = -7$ with $\alpha = \beta = 0.1$. We introduce pixel-space synthetic outliers during the beginning of training for our approach and baselines except for the VOS variants where we introduce the outliers at epoch 40 following standard practice.

D. Fine-grained results for MedMNIST, ISIC2019 and Colorectal Cancer Benchmarks

Figure 6 summarizes the performance of different calibration strategies in terms of balanced accuracy (average of sensitivity and specificity) and AUROC scores for modality shift and novel class detection. Further, in Table 6, we provide the AUPRIn scores for novel class detection for each of the MedMNIST datasets. In general, we find that our approach significantly outperforms the baselines except in the case of PathMNIST and DermaMNIST. Tables 7 - 14 provide the AUROC/AUPRIn scores for modality shift detection obtained with each of the MedMNIST datasets. Tables 15 and 16 provide the AUPR scores for both novel class and modality shift detection on the full resolution benchmarks.

E. Additional Histograms of Negative Energy Scores

Figures 7 and 8 (first row) depict the histograms of the negative energy scores where BloodMNIST and DermaMNIST are used as ID, and DermaMNIST and OrganaMNIST are used as modality shifts respectively. The second row corresponds to the histograms associated with the novel class detection in each case. We find that our approach produces well-separated distributions and much higher scores for ID data in all examples.



(a) Blood MNIST

(b) Path MNIST

(c) Derma MNIST

(d) Tissue MNIST

(e) OCT MNIST

(f) OrganA MNIST

(g) OrganS MNIST

(h) OrganC MNIST

(i) ISIC 2019 Lesions

(j) Colorectal Cancer

Figure 6. **Evaluation of OOD detectors calibrated with different inlier/outlier constructions.** The radar plots correspond to models trained on each of the benchmarks and they report the respective balanced test accuracy (%) (left), average AUROC (%) for modality shifts (middle) and novel class (right) detection.

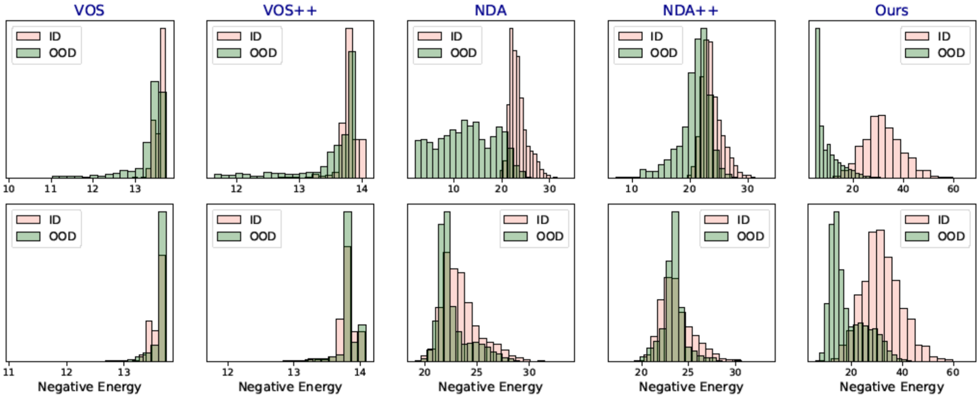


Figure 7. **Histograms of negative energy scores**. We plot the scores obtained using different inlier and outlier specifications. With BloodMNIST as ID, the top row corresponds to modality shifts (OOD: DermaMNIST) and the bottom row shows novel classes.

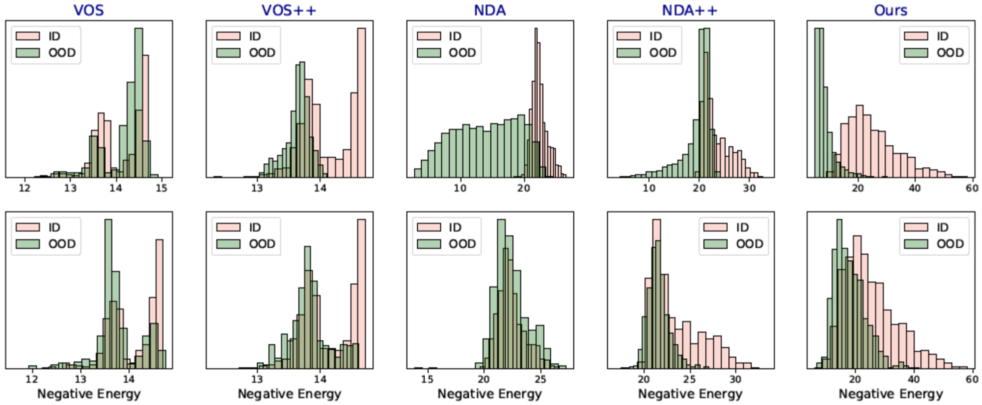


Figure 8. **Histograms of negative energy scores.** We plot the scores obtained using different inlier and outlier specifications. With DermaMNIST as ID, the top row corresponds to modality shift (OOD: OrganAMNIST) and the bottom row shows novel classes.