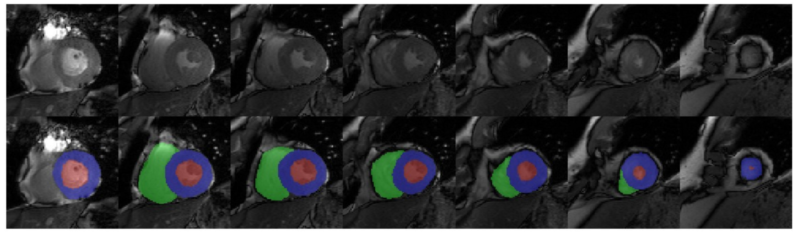
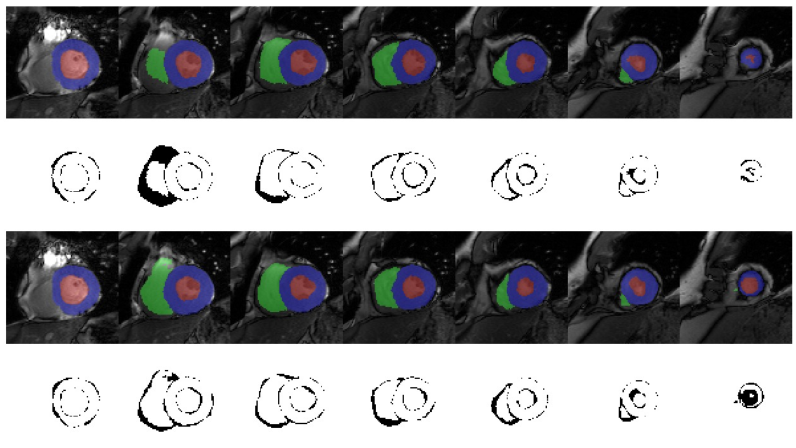
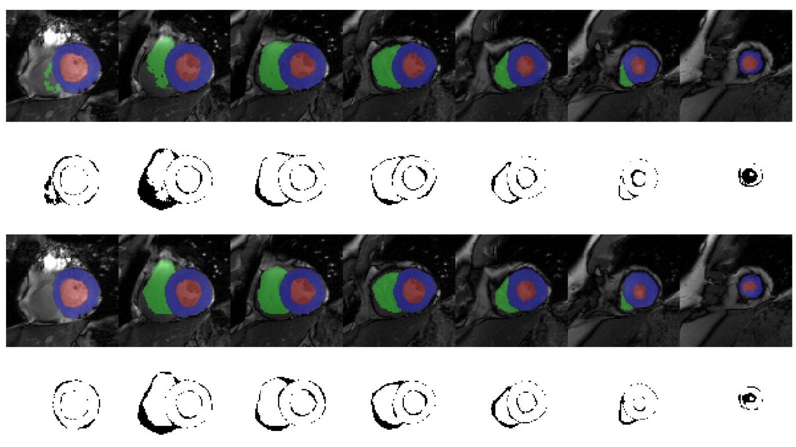


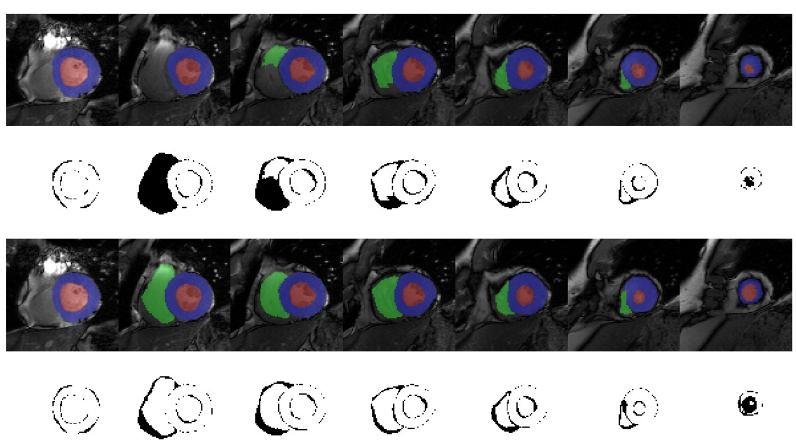
(a) Input volume (top row) with overlaid ground-truth segmentation (bottom row).



(b) Segmentation results for SegNet (top two rows) and DMR-SegNet (bottom two rows).

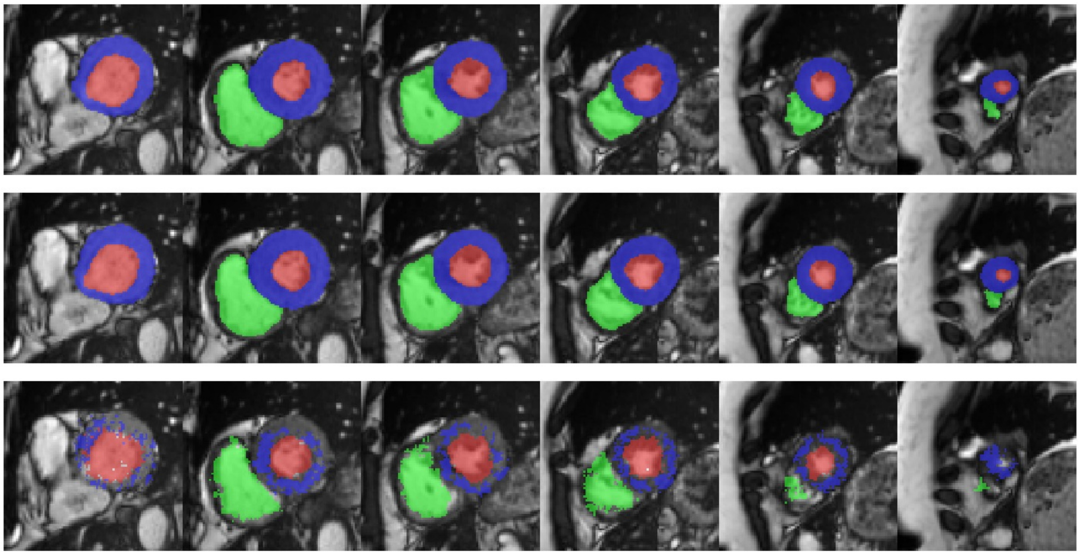


(c) Segmentation results for USegNet (top two rows) and DMR-USegNet (bottom two rows)

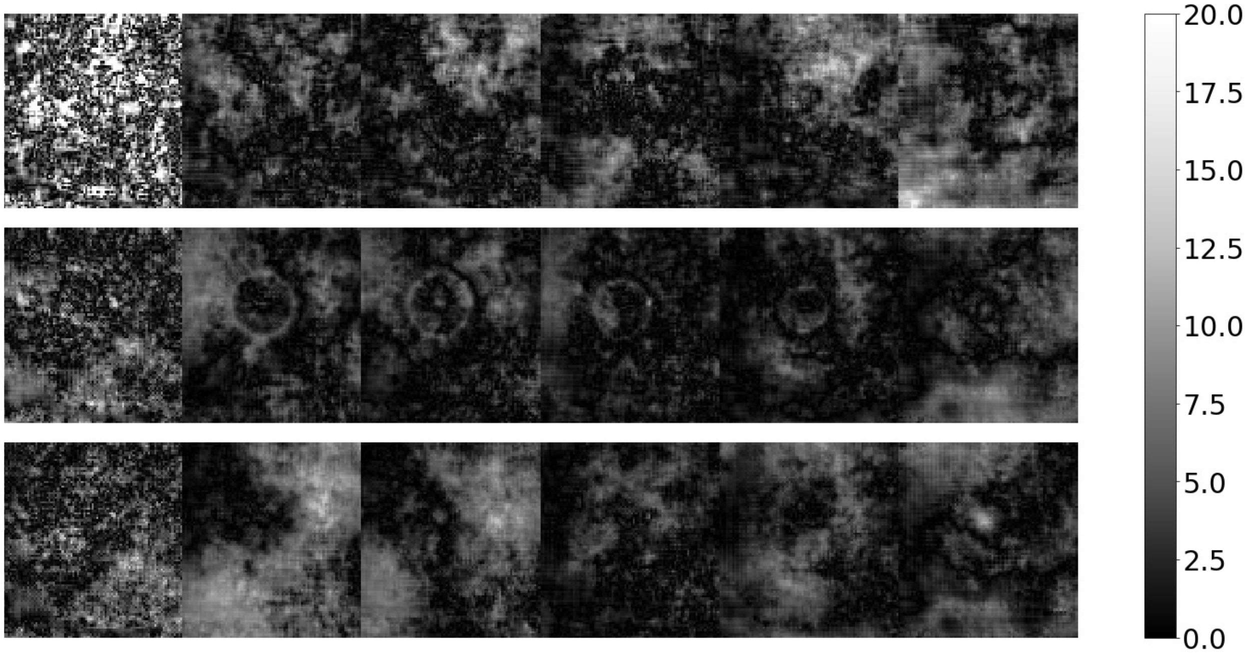


(d) Segmentation results for UNet (top two rows) and DMR-UNet (bottom two rows)

Figure S1: Ground-truth and automatic segmentation obtained from all trained models for a test patient. In each sub-figure, the segmentation obtained from the baseline and regularized model are overlaid onto the volume and shown in first and third rows, respectively; corresponding disagreement (in black) between the obtained segmentations and the ground-truth is shown in second and fourth rows, respectively.

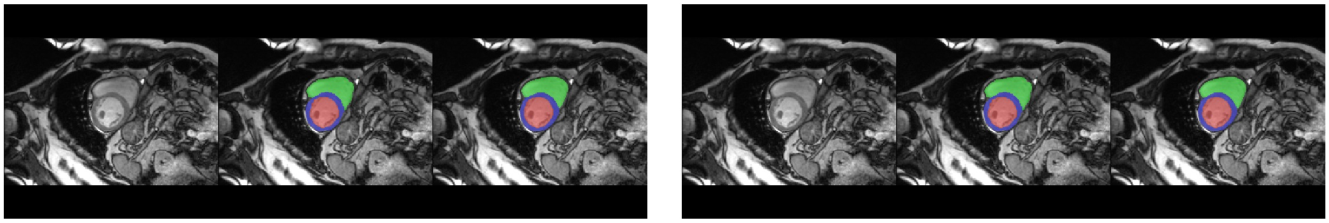


(a) Input volume with: (top row) ground-truth segmentation overlaid, (middle row) segmentation obtained from the DMR-UNet model, and (bottom row) segmentation obtained after thresholding the predicted distance map at zero levelset.

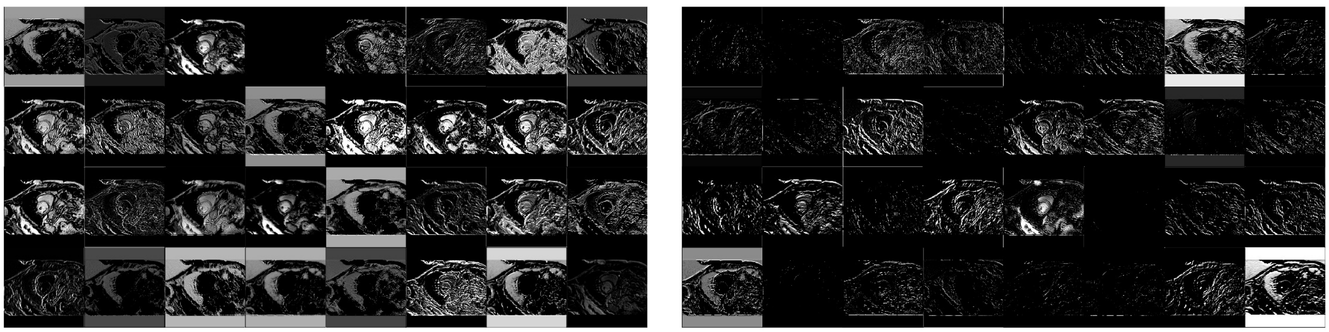


(b) Absolute difference between the ground-truth and predicted distance maps. First, second, and third row show the error in RV, LV myocardium, and LV bloodpool, respectively.

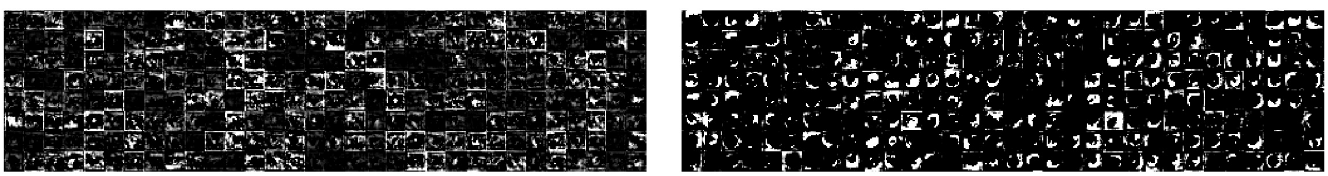
Figure S2: Visualization of (a) the segmentation obtained by thresholding the predicted distance map and (b) absolute error between the ground-truth and predicted distance maps for all chambers. Shown is only a cropped region around the heart, the error in predicted distance map is higher for the regions farther from the heart.



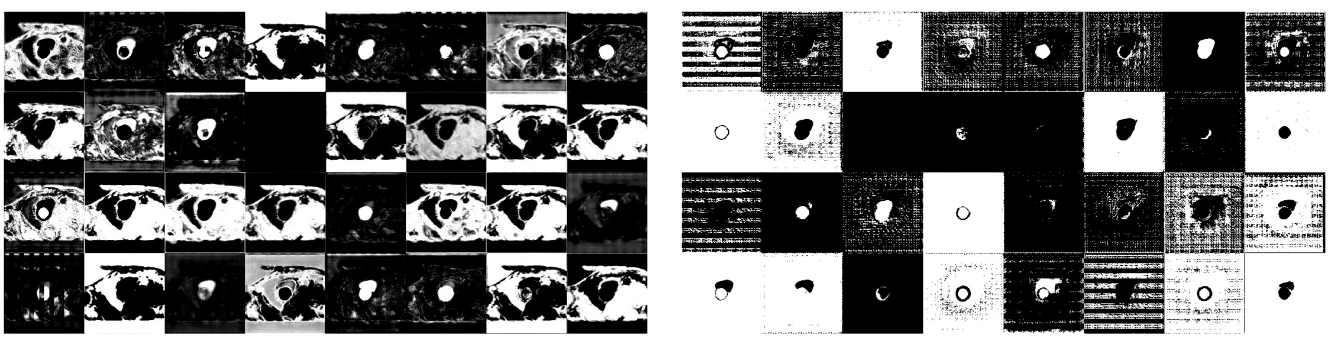
(a) From left to right: input image, ground-truth, and automatic segmentation overlay.



(b) 32 feature maps before first max-pooling operation.

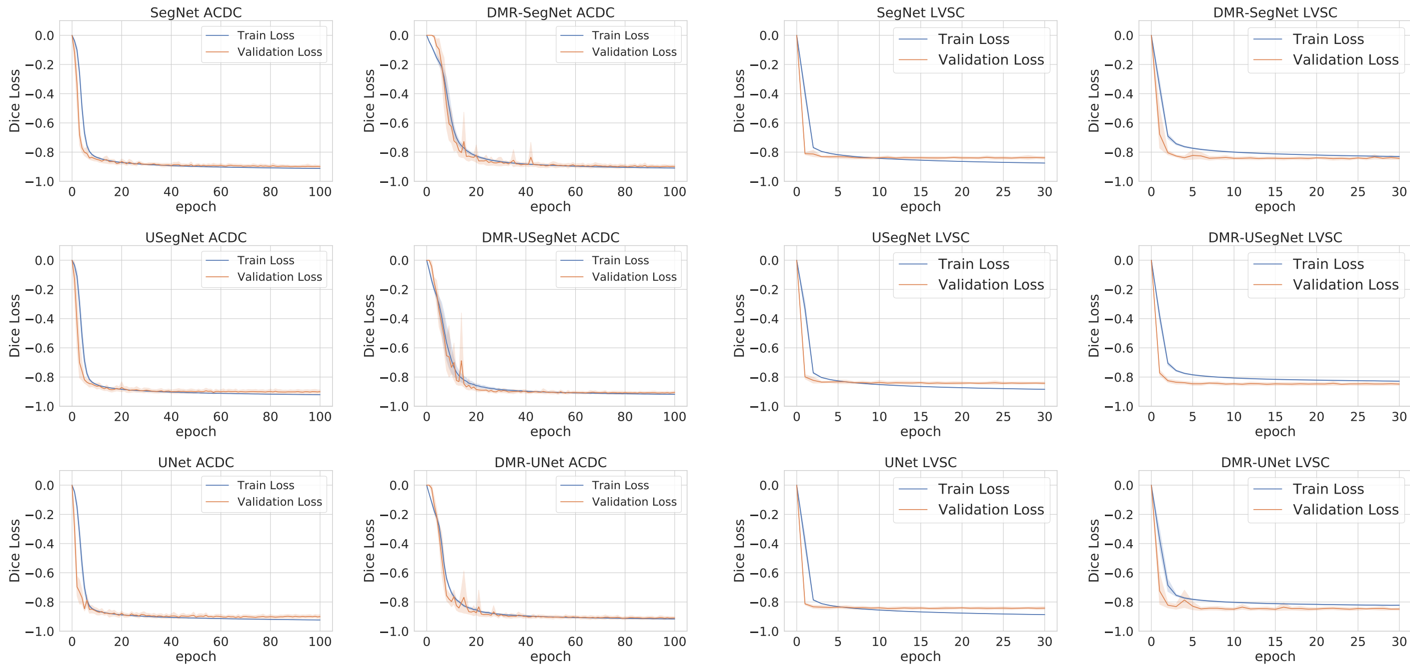


(c) 256 feature maps from the bottle-neck layer.

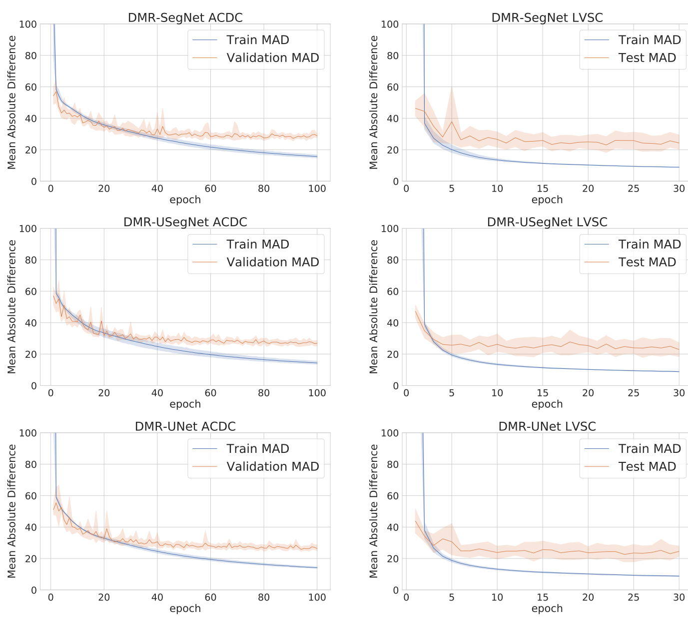


(d) 32 feature maps before the final 1×1 convolution.

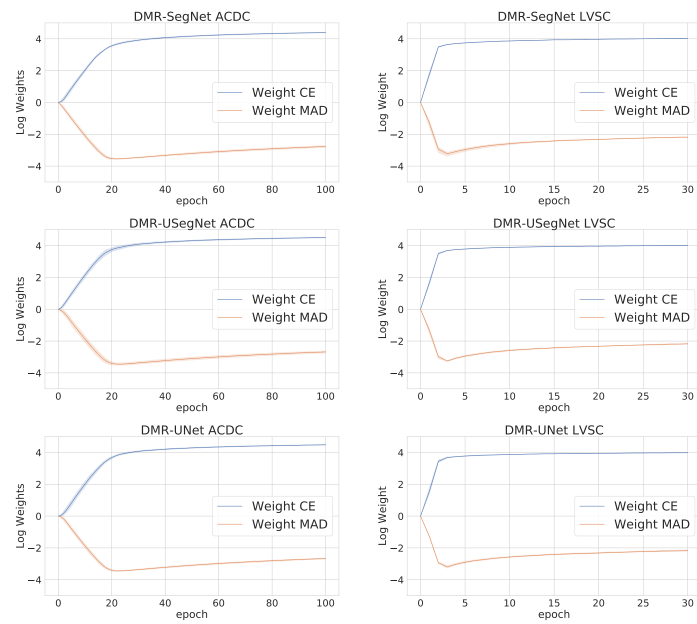
Figure S3: Feature maps visualized for the UNet (left column) and DMR-UNet (right column) model. We can observe the UNet model preserves the intensity information and propagates it throughout the network, hence, is more sensitive to the dataset-specific intensity distribution. On the other hand, the DMR-UNet model focuses more on the edges and other discriminative features, producing sparse feature maps, while ignoring dataset-specific intensity distribution. However, the results obtained for intra-dataset segmentation (shown here for ACDC dataset) is similar for both models, whereas, there is a significant improvement in cross-dataset segmentation after distance map regularization.



(a) Training and validation Dice loss for segmentation task. ACDC (left two columns) and LVSC (right two columns).

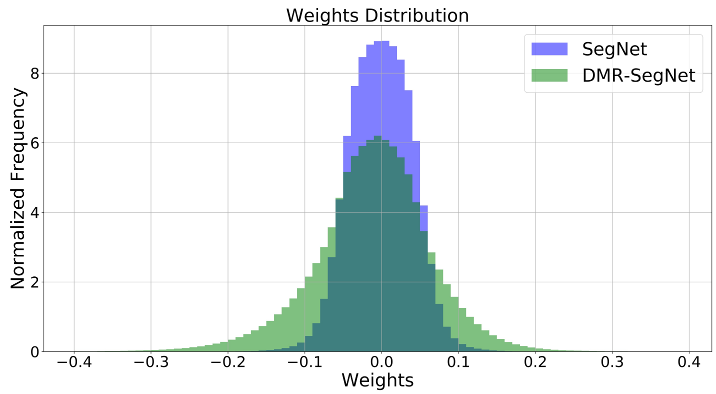


(b) Training and validation mean absolute difference error for distance map regression task. ACDC (left) and LVSC (right).

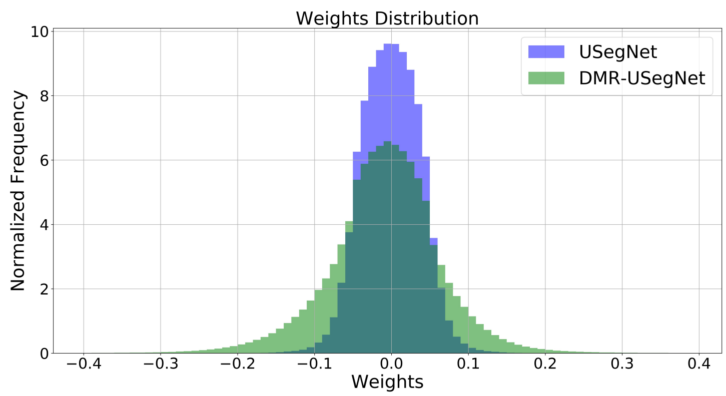


(c) Log Weights learned for cross-entropy and mean absolute difference losses. ACDC (left) and LVSC (right).

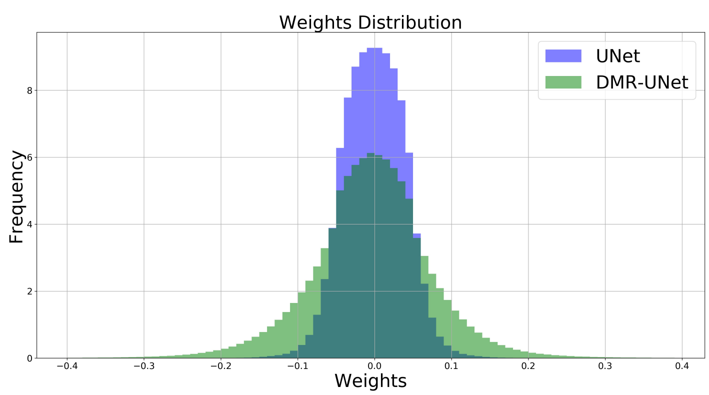
Figure S4: Mean and 95% bootstrap confidence interval for training and validation losses (a and b), and the learned weights for cross-entropy and mean absolute difference losses (c), on ACDC and LVSC dataset across five-fold cross-validation. Since the cross-entropy loss is harder to interpret, we plot the corresponding dice loss computed during training and validation. We can observe lower difference between the training and validation dice loss for the distance map regularized models, demonstrating their ability to prevent overfitting.



(a) Weights distribution for SegNet and DMR-SegNet models.



(a) Weights distribution for USegNet and DMR-USegNet models.



(a) Weights distribution for UNet and DMR-UNet models.

Figure S5: Weights distribution before and after distance map regularization for models trained across five-fold cross-validation. We can observe the number of non-zero weights increases after the distance map regularization, hence, better utilizing the network capacity.