Lab Report (SuSe 2025)
Visualization and Medical Image Analysis

Visualization of neuronal activations under class-specific perturbations

Sinem Bilge Güler
M.Sc. Student
s03sgule@uni-bonn.de

Annika Mikliss
Ph.D. Student, Supervisor
mikliss@informatik.uni-bonn.de

University of Bonn, Department of Computer Science

Abstract

Deep learning (DL) models have achieved state-of-the-art performance in medical image segmentation, but they are often criticized as "black box" systems due to the limited interpretability of their learned strategies. In this work, we systematically investigate how class-specific image transformations influence both the output predictions and the internal representations of a DL segmentation model. First, we generate controlled class-specific transformations by applying brightness adjustments. Second, we analyze how these perturbations affect the segmentation results, with a focus on class-dependent variations. Finally, we explore and visualize the importance of individual feature maps to identify which internal activations are most sensitive to transformations. Our results provide insights into the robustness and interpretability of segmentation models, highlighting pathways to improve their reliability in clinical applications.

1. Introduction

Cardiac magnetic resonance imaging (CMR) plays a central role in the diagnosis and monitoring of cardiovascular diseases [SMBB*20]. In particular, the segmentation of the left ventricle (LV), right ventricle (RV), and myocardium (MYO) is critical for quantitative analysis, such as evaluating ejection fraction, ventricular volume, and myocardial thickness [PD11]. Therefore, accurate delineation of these structures is essential for clinical decision making and treatment planning.

Deep learning (DL) methods have demonstrated state-of-the-art performance in cardiac image segmentation [RFB15, IJK*21]. However, despite their high accuracy, such models are often criticized for their lack of robustness and interpretability [MSM18, FBI*19]. Small changes in input data, such as brightness variations, can lead to significant differences in output, raising concerns about their reliability in clinical practice [ZWY*20]. Understanding how these models respond to perturbations and how internal feature representations change is therefore a key step toward building more trustworthy and clinically usable systems.

In this study, we focus on analyzing how class-specific brightness perturbations influence both the outputs and the internal feature representations of a deep learning segmentation model. We worked with a dataset of 10 3D CMR images, from which 100 representative 2D slices were extracted. The model's predictions for the LV, RV and MYO were examined under both original and brightness-adjusted conditions to contextualize changes in the segmentation outcomes. To probe the internal behavior, we placed hooks on normalized convolutional layers with varying channel dimensions and quantified perturbation-induced changes in feature maps using mean absolute deviation (MAD), mean squared error (MSE), and structural similarity index (SSIM). Concretely, our work addresses three tasks: (1) systematically generating class-specific transformations, (2) investigating how the output predictions change under these perturbations, and (3) analyzing and visualizing which feature maps are most strongly affected. Together, these steps address the central question of how class-specific transformations influence the activations inside a deep learning segmentation model.

2. Model Architecture

U-Net, introduced by Ronneberger et al. [RFB15], marked a paradigm shift in biomedical image segmentation by enabling accurate training with limited annotated datasets through an encoder–decoder structure with symmetric skip connections. The encoder path consists of repeated convolution and pooling operations, which progressively reduce the spatial resolution while increasing the feature depth. This process captures high-level semantic context. In contrast, the decoder path performs up-convolutions to gradually restore spatial resolution. Skip connections between encoder and decoder layers ensure that spatial information lost during down-sampling is directly reused during reconstruction, thereby improving localization accuracy. While U-Net forms the architectural backbone of our study, we extend their line of work by probing robustness: rather than focusing only on output Dice scores, we investigate how intermediate feature maps respond to brightness adjustments.

3. Related Works

Beyond the U-Net backbone itself, a variety of approaches have been explored to address robustness, classical segmentation challenges, and domain generalization in medical imaging. Deep learning models for medical image segmentation are often criticized as “black boxes,” since it is unclear which strategies they adopt to solve a task. To address this limitation, Ankenbrand et al. [ASH*21] introduced a sensitivity-analysis framework, `misas`, that systematically perturbs medical images through transformations such as brightness and contrast adjustments etc. Their study primarily focused on robustness by quantifying performance drops (e.g., Dice scores) and visualizing sensitivity to such alterations in two cardiac MRI case studies. In our work, we adopt a similar systematic approach but extend it in two key directions. First, instead of applying only global perturbations, we generate class-specific transformations (e.g., selectively modifying the left ventricle or right ventricle). Second, we go beyond output evaluation and additionally investigate feature maps. This enables us not only to assess segmentation robustness, but also to provide insights into which layers and channels of the network are most affected by input changes.

Cardiac MR segmentation itself is challenging due to noise, partial volume effects, and high inter-patient variability. To address this, Petitjean and Dacher [PD11] conducted a comprehensive review of segmentation methods in short-axis cardiac MR images. They summarized classical approaches, including deformable models, region growing, atlas-based techniques, and related methods, and discussed their performance and limitations under different imaging conditions, such as apical/basal slices or the presence of papillary muscles. While these methods achieved reasonable accuracy in controlled settings, they frequently failed under high variability—particularly for the right ventricle—limiting their clinical applicability. Compared to our work, this review provides a baseline understanding of traditional challenges in cardiac segmentation, which we revisit from the perspective of understanding the effects of brightness adjustments on segmentation.

Isensee et al. [IJK*21] introduced nnU-Net, a self-adapting framework that automatically configures preprocessing, network architecture, training schedule, and postprocessing steps for any given dataset. Demonstrating state-of-the-art performance across 19 international competitions covering 49 segmentation tasks, nnU-Net established itself as a standardized, out-of-the-box baseline that removed the need for manual pipeline design. While their contribution lies in making U-Net-based models adaptive and dataset-agnostic at the system level, our work complements this direction by analyzing how such networks behave internally when subjected to systematic perturbations. In particular, whereas nnU-Net improves robustness through automated configuration of the overall segmentation pipeline, we focus on interpreting the impact of input transformations on the internal activations, examining feature maps at the layer level.

A critical issue for deploying segmentation models in practice is their lack of generalization across domains, such as differences in scanner hardware, acquisition protocols, or patient populations. Zhang et al. [ZWY*20] addressed this by proposing the Deep Stacked Transformation (DST, also referred to as BigAug) framework, which simulates domain shifts during training through extensive stacked augmentations of image quality, appearance, and spatial configuration. Their experiments across MRI and ultrasound datasets demonstrated that DST substantially reduced performance degradation on unseen domains—for example, limiting Dice score drops to about 11%, compared to 25% with CycleGAN-based domain adaptation and 39% with standard augmentation. In contrast, our study does not directly train models for cross-domain robustness but instead analyzes how domain-like perturbations, such as brightness adjustments, propagate within the network. By identifying which channels are most sensitive to such transformations, our approach complements DST by providing interpretability into the sources of model vulnerability.

4. Methods

4.1. Dataset and Preprocessing

Our model distinguishes four classes: class 0 refers to the background of the image, class 1 refers to the right ventricle (RV), class 2 refers to the myocardium (MYO), and class 3 refers to the left ventricle (LV). In Figure 1, a ground-truth segmentation of the LV, RV, and MYO can be seen.

As a dataset we worked with 10 representative 3D cardiac MR images. From each 3D volume, 10 2D slices were extracted, resulting in 100 slices in total. For simplicity and to ensure accurate results, we dismissed the slices that did not produce relevant masks for LV, RV and MYO segments, which left us respectively with 98, 79 and 98 slices.

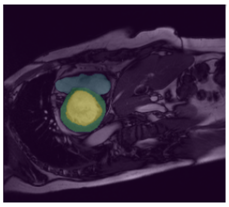
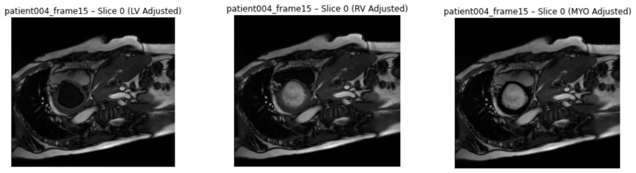


Figure 1: "Ground truth segmentation of a cardiac MR slice showing the left ventricle (yellow), right ventricle (blue), and myocardium (green)"

4.2. Class-Specific Transformations

To systematically probe robustness, we applied targeted brightness perturbations to individual anatomical classes. Masks were defined for the LV, RV and MYO regions. Specifically, pixel intensities within the LV mask were multiplied by a scaling factor of 0.1, while all other structures were left unchanged. The same procedure was applied to the RV and MYO masks.

This selective manipulation enabled us to study class-dependent sensitivity, rather than assessing only global intensity shifts. Both brightness adjustment cases were performed separately, as illustrated in Figure 2.


patient004_frame15 - Slice 0 (LV Adjusted)
patient004_frame15 - Slice 0 (RV Adjusted)
patient004_frame15 - Slice 0 (MYO Adjusted)

(a) LV Brightness Adjusted

(b) RV Brightness Adjusted

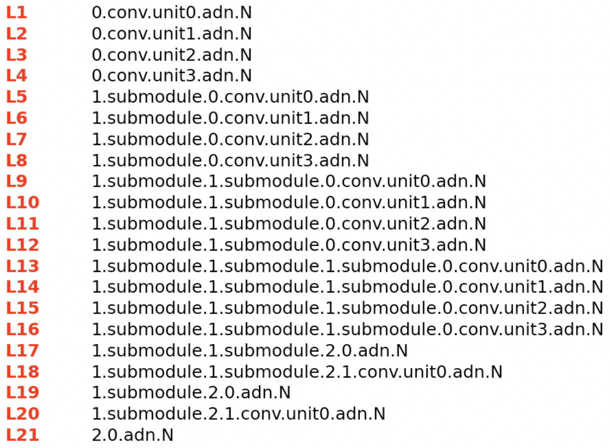
(c) MYO Brightness Adjusted

Figure 2: "Targeted brightness adjustments for the ventricles: (a) left ventricle, (b) right ventricle, (c) myocardium"

4.3. Hooking Strategy and Feature Maps

We employed a pre-trained U-Net variant (`UNetHeart_8_4_4`) that was originally trained on cardiac MR data. The network takes single-channel 2D MR slices as input. The first convolutional layer produces 8 feature maps, and the number of channels doubles at each encoder stage, resulting in four levels of down sampling depth. Following the typical encoder–decoder structure of U-Net, the encoder progressively reduces spatial resolution to capture high-level features, while the decoder restores spatial detail through upsampling and skip connections. In total, the network comprises 21 convolutional layers, distributed across the encoder, bottleneck, and decoder paths, and outputs pixel-level predictions for the LV, RV, and MYO.

For feature map analysis, we put hooks on normalization layers (e.g. `0.conv.unit0.adn.N`) associated with 21 convolutional layers which can be seen in Figure 3. These hooks provides access to intermediate feature maps during forward passes. Unlike raw convolutional outputs, which can vary widely in magnitude across channels, normalization layers rescale activations into a common numerical range. This ensures that comparisons across feature maps remain unbiased, as no single channel dominates the analysis due to scale differences. Working in the normalization stage, we enable a systematic and interpretable evaluation of how class-specific transformations affect internal representations.



L1	0.conv.unit0.adn.N
L2	0.conv.unit1.adn.N
L3	0.conv.unit2.adn.N
L4	0.conv.unit3.adn.N
L5	1.submodule.0.conv.unit0.adn.N
L6	1.submodule.0.conv.unit1.adn.N
L7	1.submodule.0.conv.unit2.adn.N
L8	1.submodule.0.conv.unit3.adn.N
L9	1.submodule.1.submodule.0.conv.unit0.adn.N
L10	1.submodule.1.submodule.0.conv.unit1.adn.N
L11	1.submodule.1.submodule.0.conv.unit2.adn.N
L12	1.submodule.1.submodule.0.conv.unit3.adn.N
L13	1.submodule.1.submodule.1.submodule.0.conv.unit0.adn.N
L14	1.submodule.1.submodule.1.submodule.0.conv.unit1.adn.N
L15	1.submodule.1.submodule.1.submodule.0.conv.unit2.adn.N
L16	1.submodule.1.submodule.1.submodule.0.conv.unit3.adn.N
L17	1.submodule.1.submodule.2.0.adn.N
L18	1.submodule.1.submodule.2.1.conv.unit0.adn.N
L19	1.submodule.2.0.adn.N
L20	1.submodule.2.1.conv.unit0.adn.N
L21	2.0.adn.N

Figure 3: “Normalized layers of the Network with their IDs”

4.4. Evaluation Metrics

We evaluated the effect of class-specific transformations on both prediction outputs and internal activations. This combined setup allows us to assess not only the performance drop caused by class-specific changes but also the internal dynamics of how the model adapts to perturbed inputs.

- **Prediction Output Metric (Dice Similarity Coefficient)** Dice Similarity Coefficient (DSC) [ZWB*04] were calculated on the resulting segmentation of the LV, and RV before and after brightness adjustment to quantify robustness in segmentation. Particularly, the Dice score is computed by comparing the model's predicted segmentation mask with the ground-truth mask.

  $$\text{Dice}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

  where A denotes the predicted segmentation mask and B denotes the ground truth mask. The Dice coefficient ranges from 0 (no overlap) to 1 (perfect overlap).

  .

- **Internal Activation Metrics (MAD, MSE, SSIM):** We measured differences between original and brightness adjusted feature maps using Mean Absolute Difference (MAD), Mean Squared Error (MSE), and Structural Similarity Index (SSIM) the formulas can be seen in the Table 1. These metrics quantify pixel-wise deviation, squared intensity differences, and structural similarity, and were evaluated independently for the LV, RV and MYO regions.

Table 1: Formulas for internal activation metrics. Here, X denotes the original feature map, Y the brightness-adjusted feature map, and N the total number of pixels.

Metric	**Equation**
Mean Absolute Difference (MAD)	$\text{MAD}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \lvert X_i - Y_i \rvert$
Mean Squared Error (MSE)	$\text{MSE}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} (X_i - Y_i)^2$
Structural Similarity (SSIM)	$\text{SSIM}(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}$

4.5. Case Categorization

To systematically analyze how brightness adjustments affect LV segmentation, we categorized model's predictions into distinct cases: Empty Prediction, Only RV, Only LV, Error, Every Class, LV → MYO misclassification, Extreme LV → MYO misclassification, and LV + MYO.

These categories reflect the typical patterns observed in our data and are motivated by explicit rules in our classification function. Empty Prediction captures cases where the network produced fewer than 35 foreground pixels in total (LV, MYO, or RV), effectively failing to segment any structure. Only LV was included because LV is the perturbed class, and we expected the model might still focus on it exclusively; this occurs when LV pixels are predicted while both MYO and RV are absent. Only RV accounts for the possibility that the model confuses LV with RV under brightness changes; here, LV is nearly absent (less than 10 pixels) together with MYO (less than or equal to 20 pixels), while RV dominates (more than 100 pixels).

For misclassification cases, we distinguish between LV → MYO and Extreme LV → MYO. The restrictions of having most of MYO pixels and all LV pixels in the LV ground truth are applied. If at least 10% of the ground-truth LV pixels are mislabeled as myocardium, the slice is labeled as LV → MYO misclassification. If more than 50% of the ground-truth LV pixels are replaced by myocardium (extreme_frac = 0.5), it is labeled as Extreme LV → MYO misclassification which is represented in the Figure 4.

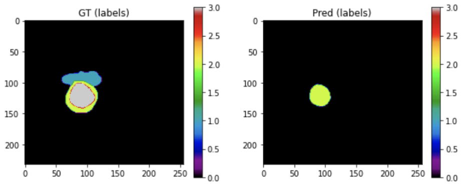


Figure 4: "Extreme LV → MYO misclassification for (a) LV Brightness Adjusted"

The LV + MYO mixed case reflects slices where both LV and MYO predictions occur inside the LV ground truth mask, with at least 70 % of predicted MYO pixels contained inside the LV ground truth this is denoted as LV→MYO_frac. Our study focuses on Extreme LV → MYO misclassification and LV + MYO cases as they contain the most slices compared to other cases.

The "Every Class" case marks situations where LV, RV, and MYO all appear together inside the ground truth LV region and are sufficiently contained. Finally, the Error label acts as a fallback whenever predictions do not satisfy the defined thresholds (e.g., structures extend outside the ground truth of LV or containment criteria are not met).

To systematically analyze how brightness adjustments affect RV segmentation, we categorized predictions into distinct cases: Empty Prediction, Only RV, RV + Others, Error, Every Class, RV → MYO misclassification, Extreme RV → MYO misclassification, and Extreme RV → LV misclassification. These categories reflect the typical patterns observed in our data and are motivated by explicit rules in our classification function.

An Empty Prediction was defined as a case where the total number of predicted pixels was less than 35, indicating that no meaningful segmentation was produced. An Only RV case occurred when RV was predicted exclusively, with no LV or MYO present. The category Every Class was assigned when at least 10 RV pixels were predicted and at least 30% of them lay inside the RV ground truth, but LV or MYO pixels also appeared in the same slice.

Misclassification cases were determined by examining how ground-truth RV pixels were reassigned. When at least 4% of ground-truth RV pixels were misclassified as MYO, the case was categorized as RV→MYO Misclassification. If half or more of the RV pixels were misclassified as MYO, the case was categorized as Extreme RV→MYO Misclassification. Likewise, when at least 50% of ground-truth RV pixels were mislabeled as LV, the slice was categorized as Extreme RV→LV Misclassification.

We renamed Every Class as RV + Others, since all predictions for RV also contained LV and MYO. RV→MYO Misclassification, Error and RV + Others are the cases that contained slices where Error has a large set of slices compared to other two. RV→MYO Misclassification and RV + Others has close number of slices.

In the case of MYO, a case-based categorization was not carried out, as all MYO brightness-adjusted (c) predictions resulted in comparable outputs where all three classes were clearly present see Figure 5.

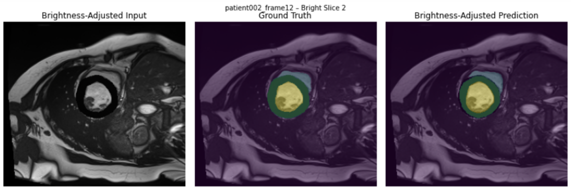

Brightness-Adjusted Input
patient002_frame12 – Bright Slice 2
Ground Truth
Brightness-Adjusted Prediction

Figure 5: Segmentation for (c) MYO Brightness Adjusted

4.6. Calculation of the Top Feature Maps

To identify the models activations that are sensitive to brightness adjustments, for LV our method was comparing the frequency of occurrence of each channel within the selected layers against the most frequent channels per layer observed across the selected cases. For MYO and RV, our method only involved comparing the frequency of each channel's occurrence within the selected layers across all three classes (LV, RV, and MYO).

To achieve this goal, first, the top feature maps were determined by selecting the 3 highest MAD resulted channels from each layer, for each patient and slice. We chose the Top-3 MAD channels as a trade-off: selecting only the Top-1 is too sensitive to noise, while selecting five or more channels dilutes the analysis by including weaker responses. The Top-3 thus provides a balance between robustness and selectivity. For each layer, we then counted how frequently a channel appeared in these Top-3 MAD sets across all patients and slices. This absolute frequency was used to identify the seven most frequently occurring channels per layer (Top-7). Reporting the Top-7 channels per layer allowed us to focus on a stable subset of channels that repeatedly appear among the Top-3 MAD results across patients and slices. Unless stated otherwise, all reported frequencies are absolute counts (i.e., number of times a channel appeared among the Top-3 MAD channels across patients and slices)

For class-specific cases (e.g., for LV: LV+MYO or Extreme LV→MYO misclassification), we matched the channels that appeared in the Top-3 MAD results with the corresponding case slices for the same patient and layer. This allowed us to link specific cases with the channels in which they occurred. We then computed the Top-7 by frequency again, restricted to the case-specific subsets which showed how frequently each channel occurred per layer.

Finally, we compared the general channel frequency distribution with the case-specific distributions. This comparison helped identify which channels or channel groups were most consistently responsible for class detection, and which layers and channels were most sensitive to brightness adjustments.

In situations where no distinct cases could be derived from the model's predicted class specific segmentation outputs after brightness adjustment (e.g., in the MYO class, where pixels from all three classes are present in each output), we instead compared the general channel frequency distributions across different classes.

5. Evaluation

5.1. Overall Dice Scores of Ventricles

In this section, we compare the Dice scores obtained before and after brightness adjustment for both the LV and RV. This allows us to observe how robust the model is when predicting segmentation after a class-specific transformation. The Dice scores provide a direct measure of the performance drop, highlighting the model's sensitivity to intensity perturbations.

In the Figure 6, the box plot clearly highlights the dramatic drop in segmentation performance after brightness adjustment for LV. Although the original images achieved a strong average Dice score of 0.94, the perturbed inputs fell to just 0.05 on average, corresponding to a mean decrease of about negative 0.88. This shift shows that the model, which performs reliably under normal conditions, is almost completely unable to segment the left ventricle once brightness is reduced. The spread of the Bright-Adjusted distribution also suggests that the failures are not isolated but systematic across slices, confirming that LV segmentation is highly sensitive to intensity perturbations.

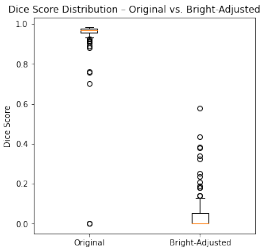


Figure 6: "Dice Score Distribution - Original vs. Bright-Adjusted (LV)

For the RV, in the Figure 7, the original images achieved a mean Dice score of 0.73, showing generally good segmentation with some variability across slices. After brightness adjustment, however, the mean Dice dropped sharply to 0.12, a decrease of about minus 0.61. The boxplot shows that most adjusted slices clustered near zero Dice, with only a few rare outliers approaching higher values. This pattern indicates that RV segmentation is strongly disrupted by brightness changes, though not as completely destroyed as the LV case. Still, the results highlight that RV features are also highly sensitive to intensity perturbations, and the robustness gap between Original and Bright-Adjusted conditions remains striking.

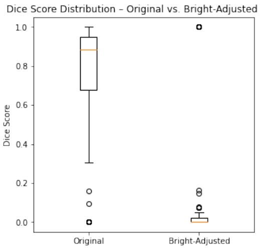


Figure 7: Dice Score Distribution - Original vs. Bright-Adjusted (RV)

For the myocardium, the original images achieved a mean Dice score of approximately 0.86, indicating good segmentation performance. After brightness adjustment, the mean Dice score decreased slightly to 0.84, a reduction of about 0.02. As shown in Figure 8, the boxplots reveal that the majority of slices clustered closely around the upper quartiles in both conditions, with only a few outliers. This suggests that the segmentation of the myocardium was generally robust to brightness changes. Still, the results highlight that MYO features are also moderately sensitive to brightness, evidenced by the small mean decline and occasional outliers, suggesting that the effect is localized to specific slices/channels rather than global.

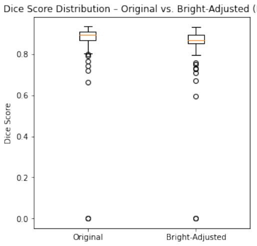


Figure 8: Dice Score Distribution - Original vs. Bright-Adjusted (MYO)

5.2. Case Evaluation

- **(a) LV Brightness Adjusted - Cases: Extreme LV to MYO misclassification and LV + MYO**
  In Figure 9, we illustrate an example of the effect of brightness adjustment on model predictions. The ground truth mask (middle) clearly separates the left ventricle (LV), right ventricle (RV), and myocardium (MYO). However, the corresponding prediction (right) demonstrates a collapse of the MYO region into the LV class. This misclassification occurs because the modification of the brightness alters the global intensity distribution of the input slice, thereby reducing the contrast between adjacent cardiac structures. For our model, it is assumed that it has learned to exploit local intensity differences and spatial boundaries; therefore, its reliability declines when these cues are diminished. In particular, often the MYO is darker than the LV, and brightness shifts can push their distributions closer together. As a result, the network fails to preserve the boundary between LV and MYO, favoring a simplified segmentation where the MYO is absorbed into the LV.

  This observation explains not only the Extreme MYO → LV misclassification case where the MYO class is entirely missing from the prediction as seen in the Figure 9, but also the more general trend of misclassification under brightness perturbations. These effects collectively contribute to the reduced Dice scores in the MYO class and highlight the sensitivity of the model to intensity-based perturbations.

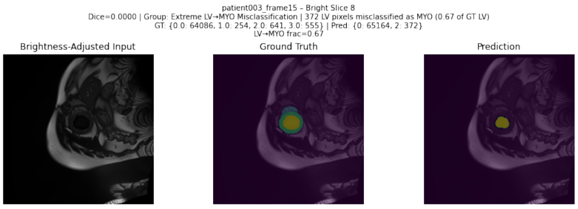

patient003_frame15 – Bright Slice 8
Dice=0.0000 | Group: Extreme LV→MYO Misclassification | 372 LV pixels misclassified as MYO (0.67 of GT LV)
GT: {0.0: 64086, 1.0: 254, 2.0: 641, 3.0: 555} | Pred: {0: 65164, 2: 372}
LV→MYO frac=0.67
Brightness-Adjusted Input
Ground Truth
Prediction

Figure 9: “LV”, “Extreme LV→MYO Misclassification”

In less severe cases, partial overlap or boundary erosion between MYO and LV can be observed. This situation is referred to as LV + MYO, as it resembles a mixture of the two classes, illustrated in Figure 10. Here, the MYO extends into the LV ground-truth region while retaining its class identity. Unlike misclassification cases, the MYO remains distinguishable and does not lose awareness of the LV class.

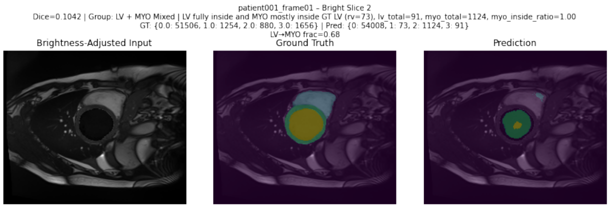

patient001_frame01 – Bright Slice 2
Dice=0.1042 | Group: LV + MYO Mixed | LV fully inside and MYO mostly inside GT LV (rv=73), lv_total=91, myo_total=1124, myo_inside_ratio=1.00
GT: {0.0: 51506, 1.0: 1254, 2.0: 880, 3.0: 1656} | Pred: {0: 54008, 1: 73, 2: 1124, 3: 91}
LV→MYO frac=0.68
Brightness-Adjusted Input
Ground Truth
Prediction

Figure 10: “LV”, “LV + MYO”

- **(b) RV Brightness Adjusted - Cases: RV+Others, RV→MYO misclassification and Error**
For the RV segmentation, in the RV+Others category we observed that RV pixel intensities within the RV region of the prediction were very small, while almost all pixel intensities were preserved for LV and MYO. When checking whether MYO and LV appeared inside the RV region, which might indicate a misclassification, it was observed that the pixel intensities of those classes were not in the RV region. The reason for this checking was the fact that the amount of pixels appeared inside of the RV region was not sufficient.

Two main reasons were identified for the occurrence of an Error case. First, some slices either lacked a large RV region or did not contain RV at all. Second, in some slices there was no or small misclassification of RV→MYO and the RV was not predicted within its anatomical ground truth region but instead appeared outside, sometimes overlapping with MYO. This failure can be attributed to several factors. The RV in cardiac MR images often has low contrast and poorly defined boundaries, making it prone to confusion with background or adjacent structures [WZX*19]. Class imbalance further favors larger structures such as LV and MYO, leading to weaker feature learning for RV [ZWY*20]. Annotation variability and the inherently thin RV boundaries introduce additional uncertainty [ASH*21]. Finally, in slices where the RV is very small or nearly absent, the model may suppress RV predictions entirely to minimize loss, effectively causing it to disappear [FBI*19]. Such cases were therefore classified as Error, as they do not conform to any predefined category. In Figure 11, we see that class 1 (RV) is represented by only 5 pixels, located outside of the RV region.

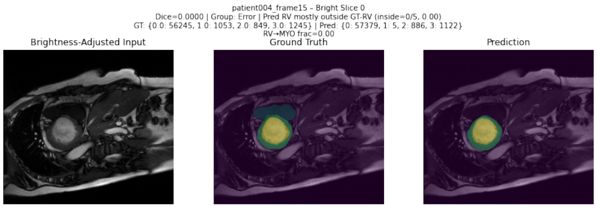

patient004_frame15 – Bright Slice 0
Dice=0.0000 | Group: Error | Pred RV mostly outside GT-RV (inside=0/5, 0.00)
GT: {0.0: 56245, 1.0: 1053, 2.0: 849, 3.0: 1245} | Pred: {0: 57379, 1: 5, 2: 886, 3: 1122}
RV→MYO frac=0.00
Brightness-Adjusted Input
Ground Truth
Prediction

Figure 11: “RV”, “Error”

Finally, the case RV→MYO Misclassification is expected because the myocardium generally appears darker than the surrounding regions. When brightness adjustments reduce the contrast, RV pixels become less distinguishable and are more likely to be reassigned to MYO. This didn’t occur as severe as it happened in LV misclassifications. The distribution of RV→MYO_frac values ranged from 0.00 to 0.18, with a mean of 0.014 and a median of 0.00. In total, 89 out of 100 slices (89%) showed fractions below 0.04, while only 11% exceeded this value. Based on this distribution, we set the threshold at 0.04 to exclude boundary noise and highlight slices with genuine RV→MYO misclassification. However, the amount of pixels observed in RV region in this case was too small.

5.3. Overall Segmentation Results/Dice Scores of Cases

In analyzing segmentation results of the LV brightness adjusted (a), we focused on the two most frequent cases, since they accounted for the largest portion of the LV ground truth. The most frequent outcome was the LV + MYO Mixed case (43 slices), with a mean Dice score of only 0.14, indicating that the model consistently failed to separate the two structures once intensity contrast was reduced. The second outcome, Extreme LV → MYO Misclassification (23 slices), performed even worse, with a mean Dice score of 0.0, reflecting a complete failure to recognize the LV when it was systematically misclassified as myocardium. As shown in Figure 6, such cases substantially contribute to the very low mean Dice score of 0.05 observed for brightness-adjusted images. The absence of “good” categories such as Only LV or Every Class suggests that brightness adjustment does not merely degrade performance slightly, but instead drives the network decisively into misclassification modes. Taken together, these results highlight a clear vulnerability: LV–MYO boundary features are highly sensitive to brightness changes, and robustness in this region cannot be assumed under intensity perturbations.

Under brightness perturbation, right ventricle segmentation remained highly unstable. The majority of slices (58) fell into the generic Error category, with a mean Dice score of only 0.002, indicating virtually no overlap with ground truth. A smaller portion of slices (12) were categorized as RV→MYO Misclassification, reflecting cases where ground-truth RV pixels were reassigned to the myocardium; here the mean Dice was 0.015, again demonstrating very limited segmentation quality. Finally, 9 slices appeared as RV+Others, but even in this case the mean Dice reached only 0.073, showing that correct RV predictions were minimal. No stable RV cases were observed overall, suggesting that brightness changes severely disrupt the network's ability to capture RV structure, far more than expected from mild intensity variation.

Considering the insufficient RV pixels in the brightness-adjusted segmentation and the lack of clear cases, we decided to combine these three cases under the "Error" case. In this situation, we cannot compare channel-occurrence frequencies in the selected layers with the per-layer most-frequent channels across the selected cases. Therefore, as stated in the subsection 4.6 we treat RV and MYO differently.

5.4. Brightness Adjustment Analysis for RV (Extra)

To systematically assess how brightness scaling influences RV segmentation, we quantified pixel intensities within the right ventricle (RV) ground truth masks under two conditions: scale = 0.1 (10% brightness preserved) and scale = 0.2 (20% brightness preserved). In total, 52,383 RV pixels were analyzed across all patients and slices. As can be seen from the Table 2, at scale = 0.1, RV intensities were markedly compressed, with a mean of 21.0 ± 8.4 (SD). In contrast, at scale = 0.2, intensities shifted upward, with a mean of 42.1 ± 16.9. These results confirm that excessive darkening (10% brightness) pushes RV intensities dangerously close to background noise, making the RV less distinguishable for the model. By preserving slightly more brightness (20%), the separation between RV and background intensities improves substantially, aligning with our observation that the model achieved higher RV segmentation accuracy under this condition.

Table 2: Mean pixel intensities (± standard deviation) of RV ground truth regions under different brightness scales. All values are aggregated across patients and slices ($n = 52{,}383$ pixels).

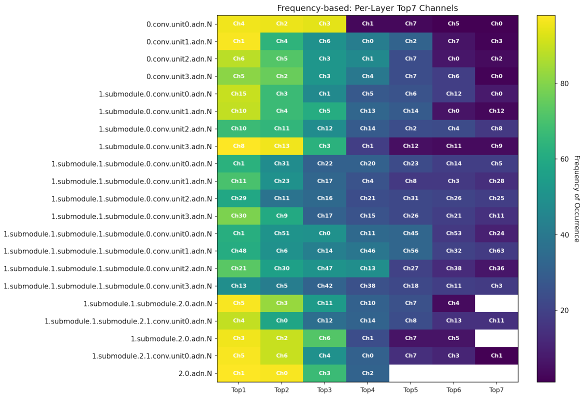
Scale	Mean	SD	Pixels
0.1 (10% brightness)	21.0	8.4	52,383
0.2 (20% brightness)	42.1	16.9	52,383

To further investigate the effect of intensity alterations on segmentation performance, we tested the model across a range of scaling factors. The results revealed that Dice scores varied noticeably with increasing perturbation strength, indicating that the model's sensitivity is not uniform but depends strongly on the magnitude of class-specific intensity changes.

5.5. Evaluation of the Top Channel Maps

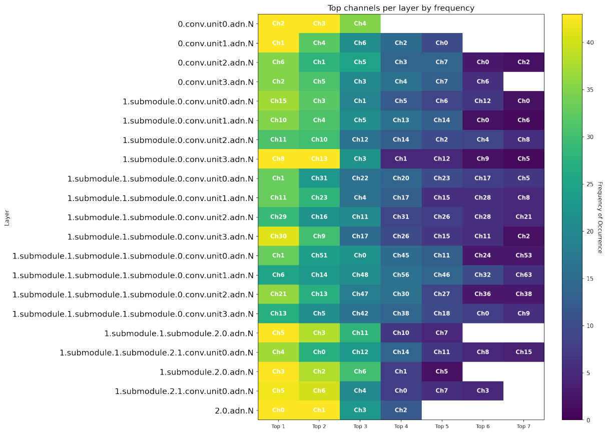
After identifying the top channel maps in subsection 4.6, we now examine how these selected feature maps behave across layers. Here, Top-k refers to the set of the k-th most frequently occurring channels in a given layer, while Top$k$ channel denotes the k-th ranked channel by frequency within that layer. The goal of this evaluation is to determine whether the network consistently relies on a small, stable subset of channels or whether certain layers exhibit variability and sensitivity to brightness adjustments. For clarity, we refer to layers as L1–L21 (see Figure 3 for mapping).

- **LV-Specific**
  The Figure 12 shows the heatmap of the Top-7 channels and their frequencies per layer for the LV class. Channel IDs that belong to Top-7 category are shown in the heatmap, and color coded frequencies represent absolute counts of how often each channel appeared among the Top-3 MAD results (brighter colors = higher counts) see subsection 4.6 for details of the computation.



	Top1	Top2	Top3	Top4	Top5	Top6	Top7
0.conv.unit0.adn.N	Ch4	Ch2	Ch3	Ch1	Ch7	Ch5	Ch0
0.conv.unit1.adn.N	Ch1	Ch4	Ch6	Ch0	Ch2	Ch7	Ch3
0.conv.unit2.adn.N	Ch6	Ch5	Ch3	Ch1	Ch7	Ch0	Ch2
0.conv.unit3.adn.N	Ch5	Ch2	Ch3	Ch4	Ch7	Ch6	Ch0
1.submodule.0.conv.unit0.adn.N	Ch15	Ch3	Ch1	Ch5	Ch6	Ch12	Ch0
1.submodule.0.conv.unit1.adn.N	Ch10	Ch4	Ch5	Ch13	Ch14	Ch0	Ch12
1.submodule.0.conv.unit2.adn.N	Ch10	Ch11	Ch12	Ch14	Ch2	Ch4	Ch8
1.submodule.0.conv.unit3.adn.N	Ch8	Ch13	Ch3	Ch1	Ch12	Ch11	Ch9
1.submodule.1.submodule.0.conv.unit0.adn.N	Ch1	Ch31	Ch22	Ch20	Ch23	Ch14	Ch5
1.submodule.1.submodule.0.conv.unit1.adn.N	Ch11	Ch23	Ch17	Ch4	Ch8	Ch3	Ch28
1.submodule.1.submodule.0.conv.unit2.adn.N	Ch29	Ch11	Ch16	Ch21	Ch31	Ch26	Ch25
1.submodule.1.submodule.0.conv.unit3.adn.N	Ch30	Ch9	Ch17	Ch15	Ch26	Ch21	Ch11
1.submodule.1.submodule.1.submodule.0.conv.unit0.adn.N	Ch1	Ch51	Ch0	Ch11	Ch45	Ch53	Ch24
1.submodule.1.submodule.1.submodule.0.conv.unit1.adn.N	Ch48	Ch6	Ch14	Ch46	Ch56	Ch32	Ch63
1.submodule.1.submodule.1.submodule.0.conv.unit2.adn.N	Ch21	Ch30	Ch47	Ch13	Ch27	Ch38	Ch36
1.submodule.1.submodule.1.submodule.0.conv.unit3.adn.N	Ch13	Ch5	Ch42	Ch38	Ch18	Ch11	Ch3
1.submodule.1.submodule.2.0.adn.N	Ch5	Ch3	Ch11	Ch10	Ch7	Ch4	
1.submodule.1.submodule.2.1.conv.unit0.adn.N	Ch4	Ch0	Ch12	Ch14	Ch8	Ch13	Ch11
1.submodule.2.0.adn.N	Ch3	Ch2	Ch6	Ch1	Ch7	Ch5	
1.submodule.2.1.conv.unit0.adn.N	Ch5	Ch6	Ch4	Ch0	Ch7	Ch3	Ch1
2.0.adn.N	Ch1	Ch0	Ch3	Ch2			

Figure 12: "Per-layer feature map frequency analysis for LV"



	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7
0.conv.unit0.adn.N	Ch2	Ch3	Ch4				
0.conv.unit1.adn.N	Ch1	Ch4	Ch6	Ch2	Ch0		
0.conv.unit2.adn.N	Ch6	Ch1	Ch5	Ch3	Ch7	Ch0	Ch2
0.conv.unit3.adn.N	Ch2	Ch5	Ch3	Ch4	Ch7	Ch6	
1.submodule.0.conv.unit0.adn.N	Ch15	Ch3	Ch1	Ch5	Ch6	Ch12	Ch0
1.submodule.0.conv.unit1.adn.N	Ch10	Ch4	Ch5	Ch13	Ch14	Ch0	Ch6
1.submodule.0.conv.unit2.adn.N	Ch11	Ch10	Ch12	Ch14	Ch2	Ch4	Ch8
1.submodule.0.conv.unit3.adn.N	Ch8	Ch13	Ch3	Ch1	Ch12	Ch9	Ch5
1.submodule.1.submodule.0.conv.unit0.adn.N	Ch1	Ch31	Ch22	Ch20	Ch23	Ch17	Ch5
1.submodule.1.submodule.0.conv.unit1.adn.N	Ch11	Ch23	Ch4	Ch17	Ch15	Ch28	Ch8
1.submodule.1.submodule.0.conv.unit2.adn.N	Ch29	Ch16	Ch11	Ch31	Ch26	Ch28	Ch21
1.submodule.1.submodule.0.conv.unit3.adn.N	Ch30	Ch9	Ch17	Ch26	Ch15	Ch11	Ch2
1.submodule.1.submodule.1.submodule.0.conv.unit0.adn.N	Ch1	Ch51	Ch0	Ch45	Ch11	Ch24	Ch53
1.submodule.1.submodule.1.submodule.0.conv.unit1.adn.N	Ch6	Ch14	Ch48	Ch56	Ch46	Ch32	Ch63
1.submodule.1.submodule.1.submodule.0.conv.unit2.adn.N	Ch21	Ch13	Ch47	Ch30	Ch27	Ch36	Ch38
1.submodule.1.submodule.1.submodule.0.conv.unit3.adn.N	Ch13	Ch5	Ch42	Ch38	Ch18	Ch0	Ch9
1.submodule.1.submodule.2.0.adn.N	Ch5	Ch3	Ch11	Ch10	Ch7		
1.submodule.1.submodule.2.1.conv.unit0.adn.N	Ch4	Ch0	Ch12	Ch14	Ch11	Ch8	Ch15
1.submodule.2.0.adn.N	Ch3	Ch2	Ch6	Ch1	Ch5		
1.submodule.2.1.conv.unit0.adn.N	Ch5	Ch6	Ch4	Ch0	Ch7	Ch3	
2.0.adn.N	Ch0	Ch1	Ch3	Ch2			

Figure 13: "Frequency distribution of channels associated with LV + MYO case

The Figure 13 shows the heatmap of the Top-7 channels and their frequencies per layer for the LV class-specific LV+MYO case. Channel IDs that belong to Top-7 category are shown in the heatmap, and color coded frequencies represent absolute counts of how often each channel appeared among the Top-3 MAD results matched with slices classified as LV+MYO (brighter colors

= higher counts) see subsection 4.6 for details of the computation.

The Figure 14 shows the heatmap of the Top-7 channels and their frequencies per layer for the LV class-specific Extreme LV -> MYO misclassification case. Channel IDs that belong to Top-7 category are shown in the heatmap, and color coded frequencies represent absolute counts of how often each channel appeared among the Top-3 MAD results matched with slices classified as Extreme LV -> MYO misclassification (brighter colors = higher counts) see subsection 4.6 for details of the computation.

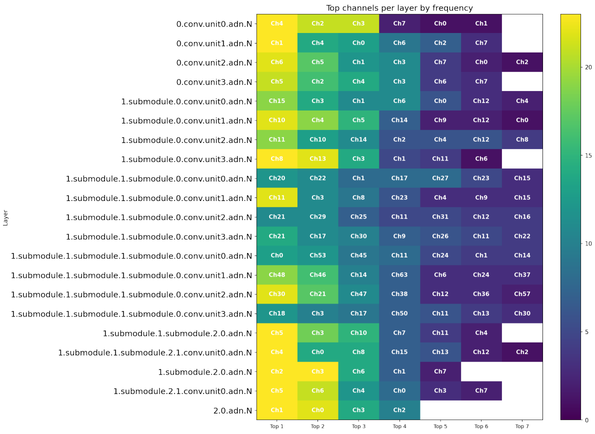


Figure 14: “Frequency distribution of top channels associated with Extreme LV–MYO misclassification case”

In Figure 12 and Figure 13, Figure 14 shows the same channel IDs consistently appear among the most frequently activated in the first six layers (L1–L6). This indicates a stable reliance on a small set of low-index feature maps during the early stages of LV representation. Similarly, in the deeper layers (L17-L21), the Top*1* and Top*2* channels also remain identical across all heatmaps. This repetition suggests that, even as the network progresses in depth, it continues to draw upon the same small set of feature maps, rather than expanding to new channel combinations.

Moreover, in Figure 12, layer L16 displays lower frequencies among its top channels compared to other layers. Moreover, in Figure 13, the same layer shows identical values for its Top-5 channels compared to Figure 12. When the corresponding feature maps are examined alongside the cases with dice scores, it is noted that the MAD values are notably high and the absolute differences clearly visible. On the other hand, the Dice scores that is computed between the ground truth mask and the predicted segmentation mask of the LV brightness adjusted image, is notably lower than the original Dice scores (see DSC).

To examine L16, firstly the Top*1* channel is investigated. Figure 15 shows Ch13 with MAD = 0.5630, the second highest MAD observed for this layer in patient001_frame01, slice 2. Moreover, the highest MAD value occurs at Ch42 and the third highest MAD value occurs at Ch05 (which are also among the Top-3 channels). This example coincides with a Dice score of 0.1042 and is labeled as an LV+MYO mixed case see the Figure 10. Moreover, for the patient002_frame12, slice 2 under the same layer, notes Ch42 as it is the highest MAD valued channel and Ch13 as it is second highest MAD valued channel. For patient003_frame01, slice 2 again the Ch13 with highest MAD value, Ch42 as the second highest and Ch05 as the third highest appears. This suggests that Ch13, Ch42, and Ch05 in L16 are activations with LV-specific brightness sensitivity. By contrast, the RV feature-map frequency analysis Figure 17 shows different Top-3 channels for L16. Taken together, these observations suggest that **L16–Ch13**, **L16–Ch42**, and **L16–Ch05** are sensitive to LV detection, consistent with class-specific channel subspace in this layer.

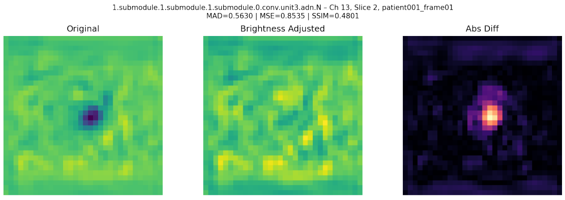

1.submodule.1.submodule.1.submodule.0.conv.unit3.adn.N – Ch 13, Slice 2, patient001_frame01
MAD=0.5630 | MSE=0.8535 | SSIM=0.4801
Original
Brightness Adjusted
Abs Diff

Figure 15: “Patient001_frame01, slice 2, Ch13”

Furthermore, in Figure 14, L11 has lower frequent top values compared to other layers. We also observed that Top*4* in Figure 12 which is Ch21 is positioned as Top*1* in Figure 14. When we further examined Ch21, we noticed that it appears mostly in the slice 8 of the patients (patient004_frame15, patient004_frame01, patient002_frame12, patient002_frame01, patient001_frame12, patient001_frame01) as the highest MAD valued channel. Also Ch16 is the Top*3* value of the Figure 12 and Top*7* value of the Figure 14. This appears as the highest MAD-valued feature in slice 8 of patient003_frame01 and patient003_frame15 see Figure 16, both corresponding to the case of Extreme LV→MYO misclassification in the Figure 9.

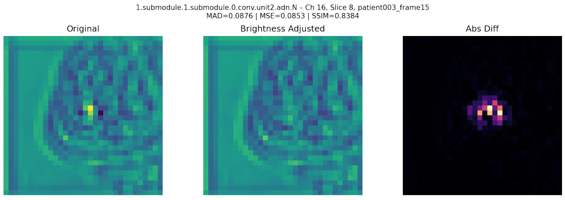

1.submodule.1.submodule.0.conv.unit2.adn.N – Ch 16, Slice 8, patient003_frame15
MAD=0.0876 | MSE=0.0853 | SSIM=0.8384
Original
Brightness Adjusted
Abs Diff

Figure 16: “patient003_frame15, L11, slice 8, Ch16”

Therefore, it can be said that Ch21 and Ch16 in particular are strongly slice-dependent and consistently peaks at slice 8, highlighting its sensitivity to LV→MYO boundary ambiguities. In the case Extreme LV→MYO misclassification, Dice typically equals to 0, we report MYO→LV_frac and LV pixel deficit instead of Dice see subsection 5.3. In both of the Ch21 and Ch16 situation we see over 0.5 frac value and this also indicates that slice 8 is particularly brightness-sensitive.

Ch21 is a multi-class involved feature map, contributing both to LV and RV analysis and becoming dominant in misclassification as it can also be seen in the L11 as Top*5*, with similar low frequency, of the Figure 17. Moreover, this channel appears as the Top*2* for Figure 18 with relatively higher frequency. This indicates shared representation across classes, which can increase the risk of class confusion. On the other hand, **L11-Ch16** remains LV-specific and only spikes in the Extreme LV→MYO misclassification case, which points to a purely LV-sensitive channel that fails under perturbation. The reappearance of the same channel IDs in both heatmaps supports that Extreme LV→MYO misclassification is not caused by entirely new features, but by a shift in reliance on existing channels (notably boundary-sensitive Ch21 and LV-specific Ch16).

In Figure 12, for L8, we observe drop in channel frequency after Top-2 channels. Ch3, Ch8, and Ch13 are present across all slices of patient001_frame01. In patient001_frame12, Ch3 appears in slices 1–7 and 9, while Ch8 and Ch13 appear in slices 1–9. For patient002_frame01, Ch8 and Ch13 is present in all slices (0–9) meanwhile Ch3 is only present in slices 1, 6 and 8. Across these cases, the MAD of Ch3 is consistently lower than that of Ch8 and Ch13. As this drop involves Ch3 in Figure 12 and aligns with the cases shown in Figure 13 and Figure 14, we compared Ch3 with the first two channels. The MAD values for this layer are very small (0.01–0.13), and there is no substantial gap between Ch3 and Ch8/Ch13. The same channel IDs also appear among the Top-7 MAD entries of the RV heatmap Figure 17; therefore, we cannot identify class specificity or a clear effect of brightness adjustment at L8.

- **RV-Specific**
  The Figure 17 shows the heatmap of the Top-7 channels and their frequencies per layer for the RV class. Channel IDs that belong to Top-7 category are shown in the heatmap, and color coded frequencies represent absolute counts of how often each channel appeared among the Top-3 MAD results (brighter colors = higher counts) see subsection 4.6 for details of the computation.

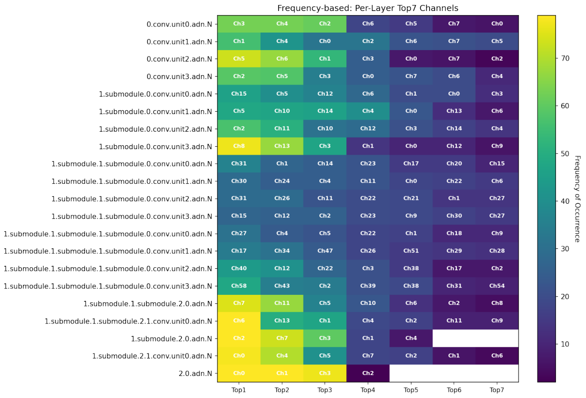


Figure 17: “Per-layer feature map frequency analysis for RV”

In he Figure 17, the same channel IDs frequently activated in the first six layers (L1–L6). This indicates a stable reliance on a small set of low-index feature maps during the early stages of RV representation.

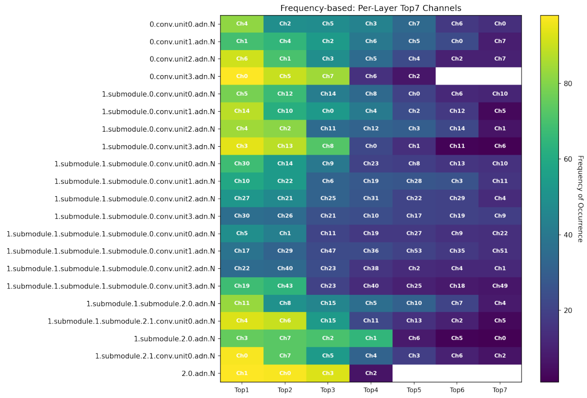
Firstly, in Figure 17, L17 shows Ch7 and Ch11 as its most frequent channels and there is a sudden drop in the frequencies starting after Top-3 channels. Patient001_frame01 has Ch7 in all of its slices expect slice 8, and Ch11 in slices 1-4 and 6-8. Patient001_frame12 has both of the channels in every slices and for each slices Ch7 is ranked with the highest MAD values. For patient002_frame01, we also see Ch7 and Ch11 in every slices of each patients. Bases on these gatherings, we can say that Ch7 and Ch11 are brightness sensitive. However, Ch7 and Ch11 appears also in the Figure 12 and in the Figure 18 Ch11 also has a high frequency so we can say that those channels are not RV class dependent.

In L20, Ch4 appears as Top*2* meanwhile this channel has lower frequency in the Figure 12 and Figure 18. Ch4 appears in the slices 1-7 of patient001_frame01 and in all slices of patient001_frame12. In first 6 slices of patient002_frame01 either ranked as highest MAD valued channel or second highest MAD valued channel and the same situation persists for patient002_frame12. For patient003_frame01 it appears in slices 1-7 as the second highest MAD valued channel. Moreover, for patient003_frame15 it appears to be the second highest MAD valued channel for all slices. It also appears in the other slices of the other patients. We can say that Ch04 has RV related brightness sensitivity. Since this channel appear in the Figure 12 as Top*3* and Figure 18, we cant say that it is RV class dependent. Based on our results, unfortunately, class information of RV is not possible to extract with this model.

Both Figure 12 and Figure 17 although they are for the different classes, they look similar. When we look at the Figure 12, this might not just show channels that are important for LV class but RV class too as LV influenced the segmentation of RV class as well. The reason is that when we adjusted the LV class brightness and this didn’t affect the only segmentation of LV but also of the RV and MYO. So the segmentation of RV dependent on the other classes. If we look closer at the RV, we can design a transformation that changes the segmentation of the other classes but not RV but this is not within the scope of this study.

- **MYO-Specific**
  The Figure 18 shows the heatmap of the Top-7 channels and their frequencies per layer for the MYO class. Channel IDs that belong to Top-7 category are shown in the heatmap, and color coded frequencies represent absolute counts of how often each channel appeared among the Top-3 MAD results (brighter colors = higher counts) see subsection 4.6 for details of the computation.

In L7, the Figure 18 identifies Ch4 as Top*1* and Ch2 as Top*2*, and their slice-wise presence is pervasive across patients: for patient001_frame12, Ch4 appears on slices 1–9 while Ch2 appears on 1, 3, 6–9; for patient001_frame01, Ch4 appears on 2–7 and Ch2 on 5–9; for patient002_frame01, Ch4 appears on 0–3, 5–7 and Ch2 all slices except 2; for patient002_frame12, both Ch4 and Ch2 appear on 0–9 (all slices); and for patient003_frame01, Ch4 appears on 0–8 and Ch2 on 0–9. This broad prevalence in MYO contrasts with LV at the same layer, where the top channels are Ch10/Ch11 and occur less frequently Figure 12.



	Top1	Top2	Top3	Top4	Top5	Top6	Top7
0.conv.unit0.adn.N	Ch4	Ch2	Ch5	Ch3	Ch7	Ch6	Ch0
0.conv.unit1.adn.N	Ch1	Ch4	Ch2	Ch6	Ch5	Ch0	Ch7
0.conv.unit2.adn.N	Ch6	Ch1	Ch3	Ch5	Ch4	Ch2	Ch7
0.conv.unit3.adn.N	Ch0	Ch5	Ch7	Ch6	Ch2		
1.submodule.0.conv.unit0.adn.N	Ch5	Ch12	Ch14	Ch8	Ch0	Ch6	Ch10
1.submodule.0.conv.unit1.adn.N	Ch14	Ch10	Ch0	Ch4	Ch2	Ch12	Ch5
1.submodule.0.conv.unit2.adn.N	Ch4	Ch2	Ch11	Ch12	Ch3	Ch14	Ch1
1.submodule.0.conv.unit3.adn.N	Ch3	Ch13	Ch8	Ch0	Ch1	Ch11	Ch6
1.submodule.1.submodule.0.conv.unit0.adn.N	Ch30	Ch14	Ch9	Ch23	Ch8	Ch13	Ch10
1.submodule.1.submodule.0.conv.unit1.adn.N	Ch10	Ch22	Ch6	Ch19	Ch28	Ch3	Ch11
1.submodule.1.submodule.0.conv.unit2.adn.N	Ch27	Ch21	Ch25	Ch31	Ch22	Ch29	Ch4
1.submodule.1.submodule.0.conv.unit3.adn.N	Ch30	Ch26	Ch21	Ch10	Ch17	Ch19	Ch9
1.submodule.1.submodule.1.submodule.0.conv.unit0.adn.N	Ch5	Ch1	Ch11	Ch19	Ch27	Ch9	Ch22
1.submodule.1.submodule.1.submodule.0.conv.unit1.adn.N	Ch17	Ch29	Ch47	Ch36	Ch53	Ch35	Ch51
1.submodule.1.submodule.1.submodule.0.conv.unit2.adn.N	Ch22	Ch40	Ch23	Ch38	Ch2	Ch4	Ch1
1.submodule.1.submodule.1.submodule.0.conv.unit3.adn.N	Ch19	Ch43	Ch23	Ch40	Ch25	Ch18	Ch49
1.submodule.1.submodule.2.0.adn.N	Ch11	Ch8	Ch15	Ch5	Ch10	Ch7	Ch4
1.submodule.1.submodule.2.1.conv.unit0.adn.N	Ch4	Ch6	Ch15	Ch11	Ch13	Ch2	Ch5
1.submodule.2.0.adn.N	Ch3	Ch7	Ch2	Ch1	Ch6	Ch5	Ch0
1.submodule.2.1.conv.unit0.adn.N	Ch0	Ch7	Ch5	Ch4	Ch3	Ch6	Ch2
2.0.adn.N	Ch1	Ch0	Ch3	Ch2			

Figure 18: “Per-layer feature map frequency analysis for MYO”

Moreover, visual inspection of representative cases, patient002_frame12, slice 2 for both Ch4 and Ch2 shows clear absolute difference along myocardial boundaries indicating brightness sensitivity see Figure 19 and Figure 20. Additionally, the MAD scores for Ch4 are generally higher than for Ch2. This indicates that Ch4 matches the MYO ground truth better and contributes more to the segmentation and it is more MYO-dependent than Ch2. At the same time, the dominance of Ch4/Ch2 is concentrated in MYO rather than LV, evidencing MYO class dependence. However, when we look at the L7 of Figure 17 we see that Ch2 is also present as Top2 but with a slightly lower frequency. Altogether, these observations support that in L7 the channels Ch4 and Ch2 are brightness-sensitive but only **L7-Ch04** is MYO class-dependent.

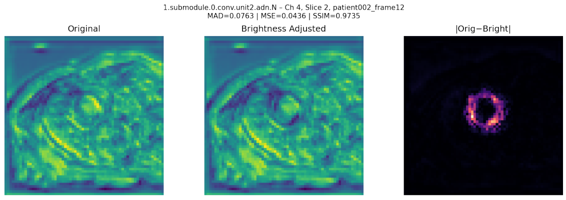

1.submodule.0.conv.unit2.adn.N – Ch 4, Slice 2, patient002_frame12
MAD=0.0763 | MSE=0.0436 | SSIM=0.9735
Original
Brightness Adjusted
|Orig−Bright|

Figure 19: “patient002frame12 ch04 slice02”

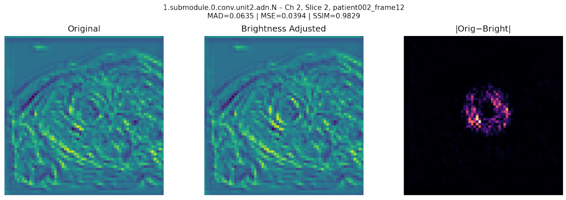

1.submodule.0.conv.unit2.adn.N – Ch 2, Slice 2, patient002_frame12
MAD=0.0635 | MSE=0.0394 | SSIM=0.9829
Original
Brightness Adjusted
|Orig−Bright|

Figure 20: “patient002frame12 ch02 slice02”

6. Conclusion

In this work, we investigated how class-specific perturbations affect both the output and internal representations of a U-Net–based segmentation model for cardiac MR images. By selectively altering the brightness of individual classes, we demonstrated that such targeted changes can lead to substantial performance drops, particularly misclassifications of the myocardium as the left ventricle. Beyond output-level evaluation, our feature map analysis revealed that specific layers and channels are disproportionately sensitive to these perturbations, offering insights into the internal dynamics of the network. For LV class specific channels we gathered **L16–Ch13, L16–Ch42, L16–Ch0, L11-Ch16** which are also brightness adjustment sensitive. For MYO class specific channels, we gathered **L7-Ch04** which is also brightness adjustment sensitive. Moreover, the class information of RV is not possible to extract with this model, we only gathered brightness sensitive channels. Together, these findings depicts how class-specific transformations influence the activations inside a deep learning segmentation model.

7. Further Improvements

The experimental results demonstrate that our current segmentation framework is particularly vulnerable to intensity perturbations, such as brightness adjustments, which often result in partial or complete misclassification of the myocardium (MYO) as the left ventricle (LV). This sensitivity reflects the model’s reliance on local intensity contrasts to separate adjacent cardiac structures. To mitigate these weaknesses and improve robustness, several directions for further development can be identified.

First, expanding the data augmentation strategy to include a richer set of photometric transformations would likely enhance the model’s invariance to intensity shifts. While geometric augmentations are already widely applied, augmentations such as brightness, contrast, and gamma adjustments have been shown to improve generalization to variable acquisition conditions [ZZX*19]. Moreover, approaches that learn transformation distributions for augmentation rather than relying on hand-crafted perturbations can produce more diverse and realistic training samples, thereby reducing the risk of domain shift [ZZX*19]. Second, loss function design offers a powerful mechanism to address class-specific weaknesses. Since LV misclassification is disproportionately high, rebalancing strategies such as class-weighted cross-entropy or focal loss can encourage the network to better preserve underrepresented classes. Furthermore, boundary-aware loss formulations, such as the Boundary Dice or Boundary IoU, have been shown to improve the delineation of structure interfaces by penalizing deviations specifically along region boundaries [KBD*19, WLX*23].

References

[ASH*21] Ankenbrand M. J., Shainberg L., Hock M., Lohr D., Schreiber L. M.: Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac mri. *BMC Medical Imaging 21*, 1 (2021), 27. doi:10.1186/s12880-021-00551-1. 2, 6

[FBI*19] Finlayson S. G., Bowers J. D., Ito J., Zittrain J., Beam A. L., Kohane I. S.: Adversarial attacks on medical machine learning. *Science 363*, 6433 (Mar. 2019), 1287–1289. doi:10.1126/science.aaw4399. 1, 6

[IJK*21] Isensee F., Jaeger P. F., Kohl S. A. A., Petersen J., Maier-Hein K. H.: nnu-net: a self-adapting framework for u-net-based medical image segmentation. *Nature Methods 18*, 2 (Feb. 2021), 203–211. doi:10.1038/s41592-020-01008-z. 1, 2

[KBD*19] KERVADEC H., BOUCHTIBA J., DESROSIERS C., ET AL.: Boundary loss for highly unbalanced segmentation. In *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)* (2019), pp. 285–296. URL: http://arxiv.org/abs/1812.07032. 10

[MSM18] MONTAVON G., SAMEK W., MÜLLER K.-R.: Methods for interpreting and understanding deep neural networks. *Digital Signal Processing 73* (Feb. 2018), 1–15. doi:10.1016/j.dsp.2017.10.011. 1

[PD11] PETITJEAN C., DACHER J.-N.: A review of segmentation methods in short axis cardiac mr images. *Medical Image Analysis 15*, 2 (2011), 169–184. doi:10.1016/j.media.2010.12.004. 1, 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI* (2015), vol. 9351, pp. 234–241. doi:10.1007/978-3-319-24574-4_28. 1

[SMBB*20] SCHULZ-MENGER J., BLUEMKE D. A., BREMERICH J., ET AL.: Standardized image interpretation and post-processing in cardiovascular magnetic resonance: Society for cardiovascular magnetic resonance (scmr) board of trustees task force on standardized post-processing. *Journal of Cardiovascular Magnetic Resonance 22*, 1 (Feb. 2020), 19. doi:10.1186/s12968-020-00610-6. 1

[WLX*23] WANG H., LIU Y., XU Z., ET AL.: Boundary iou: Improving semantic segmentation evaluation with boundary-based metrics. *IEEE Transactions on Image Processing 32* (2023), 2580–2591. doi:10.1109/TIP.2023.3248009. 10

[WZX*19] WANG Y., ZHANG Y., XUAN W., KAO E., CAO P., TIAN B., ORDOVAS K., SALONER D., LIU J.: Fully automatic segmentation of 4d mri for cardiac functional measurements. *Medical Physics 46*, 1 (Jan. 2019), 180–189. doi:10.1002/mp.13245. 6

[ZWB*04] ZOU K. H., WARFIELD S. K., BHARATHA A., TEMPANY C. M. C., KAUS M. R., HAKER S. J., WELLS W. M., JOLESZ F. A., KIKINIS R.: Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology 11*, 2 (Feb. 2004), 178–189. doi:10.1016/S1076-6332(03)00671-8. 3

[ZWY*20] ZHANG L., WANG X., YANG D., SANFORD T., HARMON S., TURKBEY B., WOOD B. J., ROTH H., MYRONENKO A., XU D., XU Z.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging 39*, 7 (2020), 2531–2540. doi:10.1109/TMI.2020.2973595. 1, 2, 6

[ZZX*19] ZHAO Z., ZHENG J., XU S., ET AL.: Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 8543–8553. doi:10.1109/CVPR.2019.00874. 10