Visualization of neuronal activations under class-specific perturbations

Prepared by : Sinem Bilge Güler
Supervised by: Annika Mikliss

Tasks

Task 1

Systematically generate class-specific transformations

Task 2

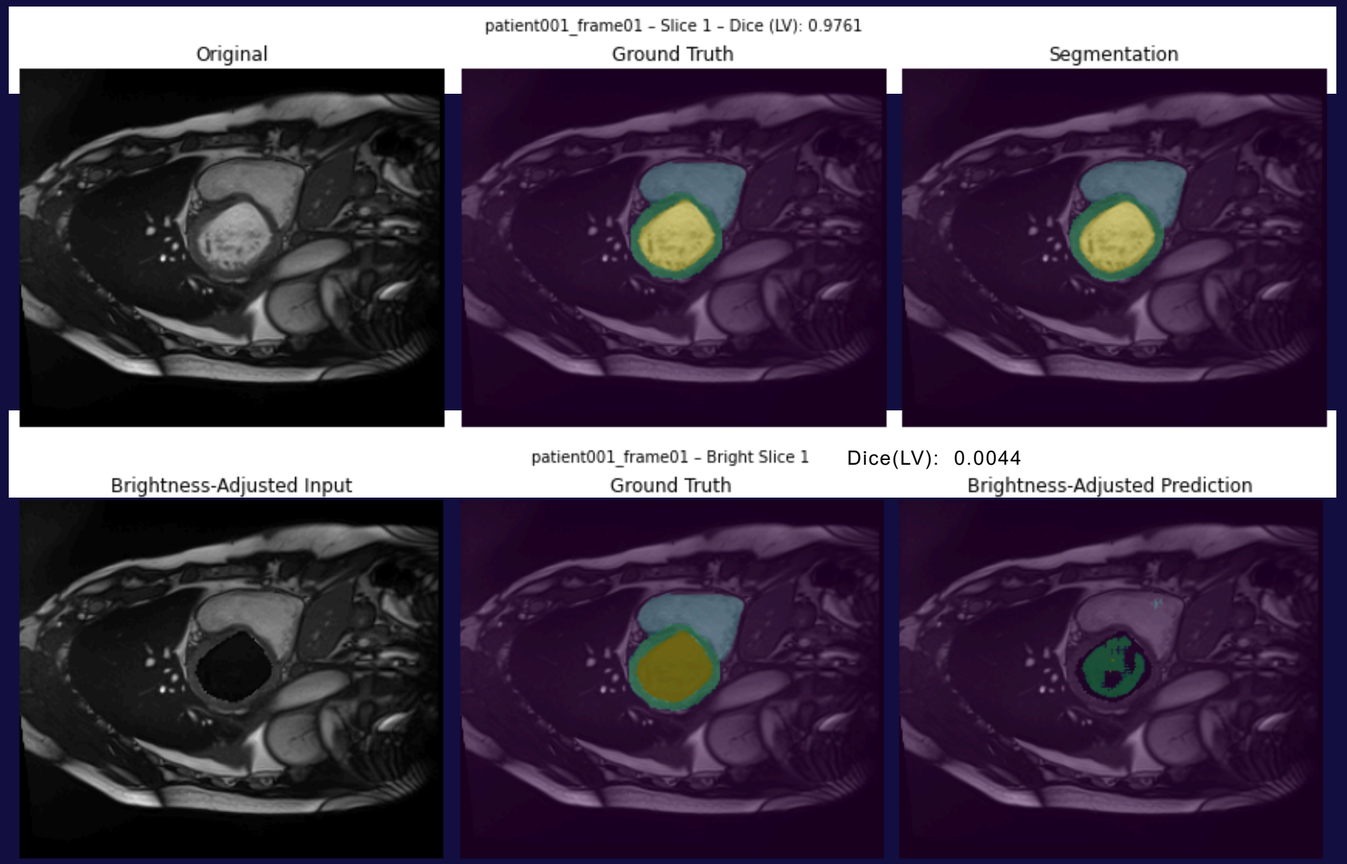
Systematically investigate how the output changes

Task 3

Investigate and visualize which and how important feature maps

- DL models are often referred to as black box models because it is not clear which strategies the models learn to solve a task.

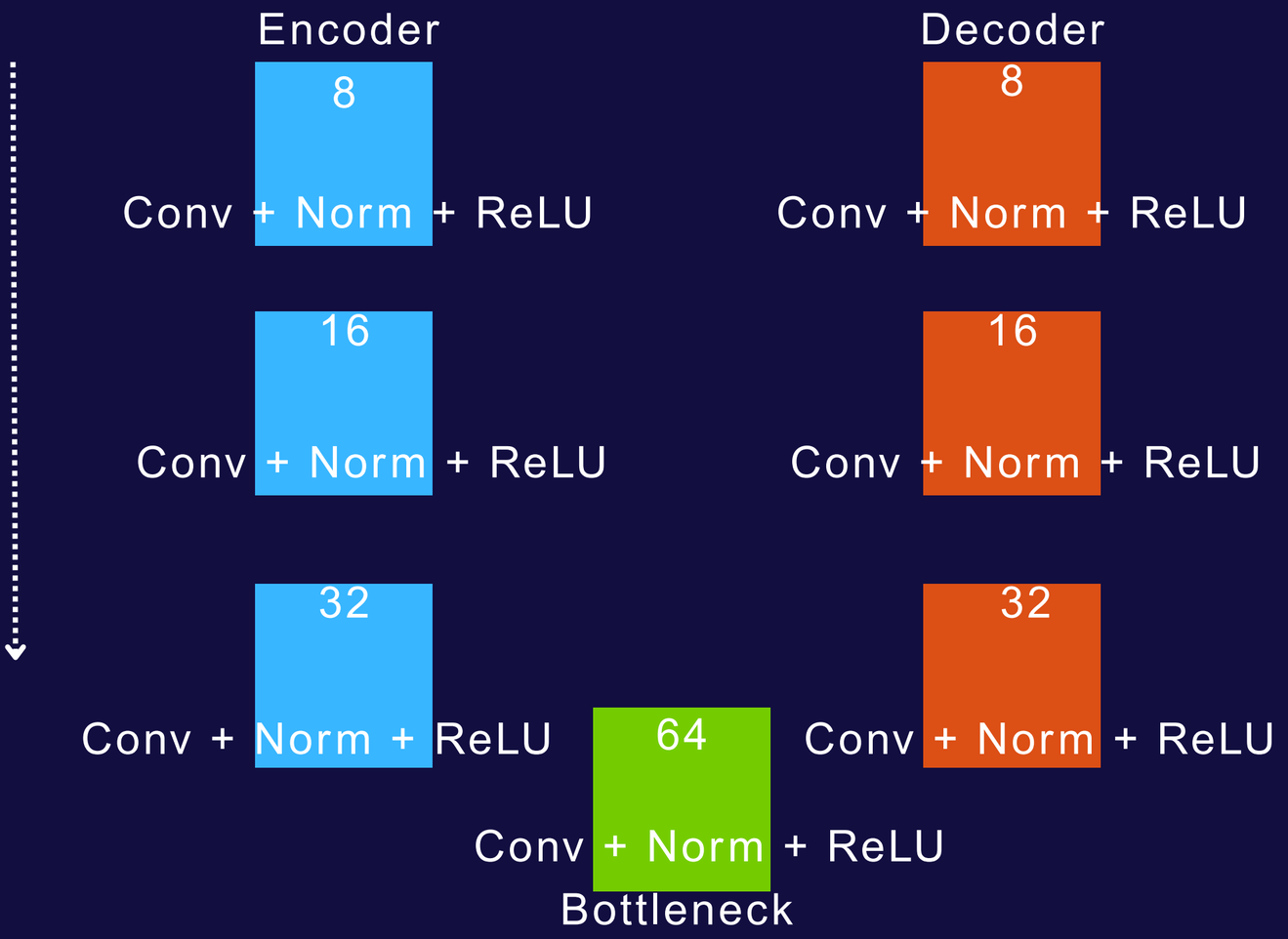
- Problem: How do class-specific transformations influence the activations inside of a DL segmentation model?



Unet Model

model = UNetHeart_8_4_4()

Learns more abstract features (edges → textures → shapes



Given the abstract features, where exactly are the pixels for each class (LV, RV, MYO)?

Uses skip connections: copies detail from encoder to decoder, so boundaries are sharp.

Data Processing Steps

10 times 3D Images

10 2D images for each 10 patient images

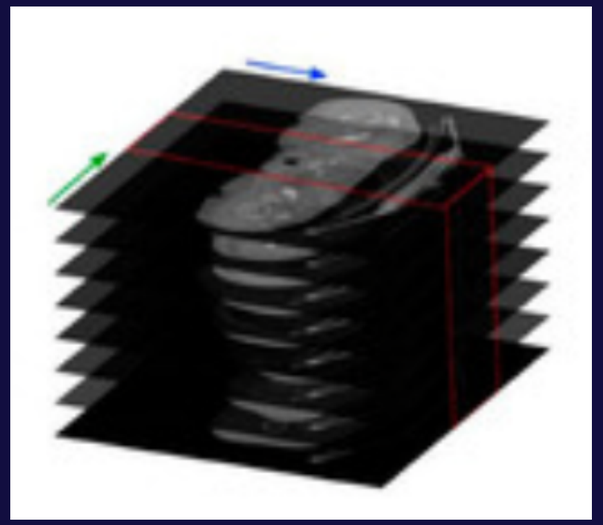
4-64 chanels/feature maps

Filter out images without the specified class

LV:98

RV:79

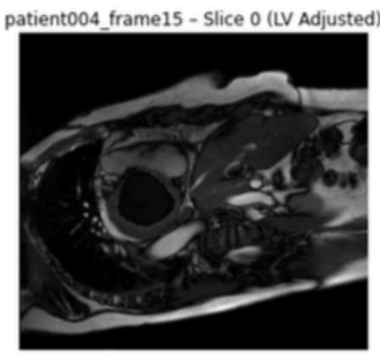
MYO:98



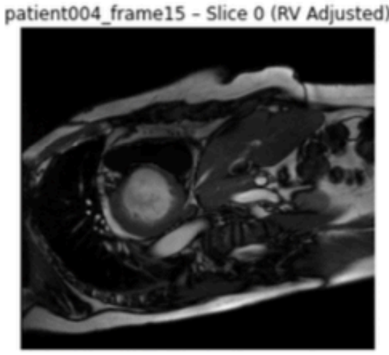
(Bertolini, Rossoni, & Colombo, 2021)

Brightness Adjustments

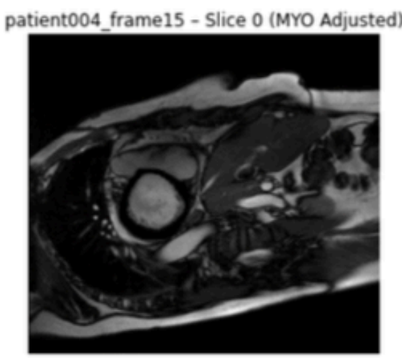
Dice Score Calculations



(a) LV Brightness Adjusted



(b) RV Brightness Adjusted



(c) MYO Brightness Adjusted

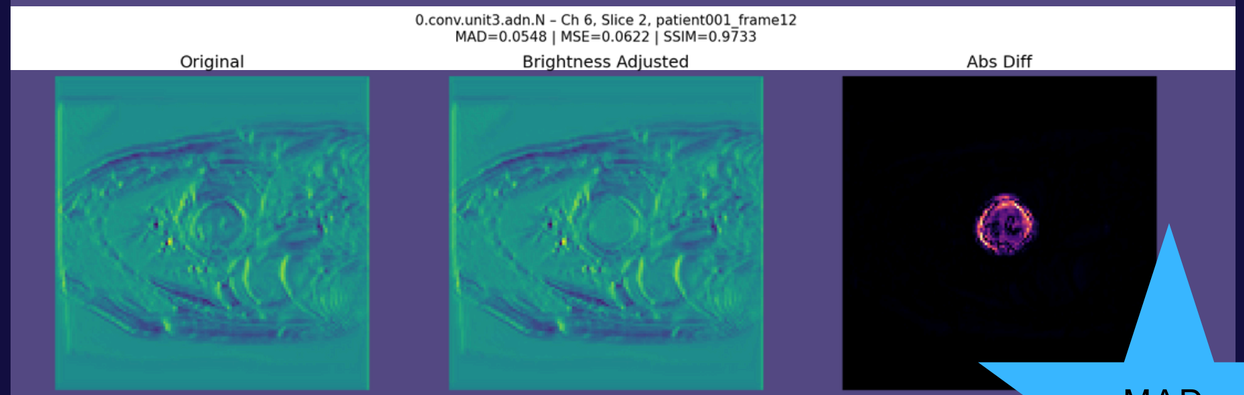
Hooks on adn.N why?

21 Hooks to different Layers

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

MAD

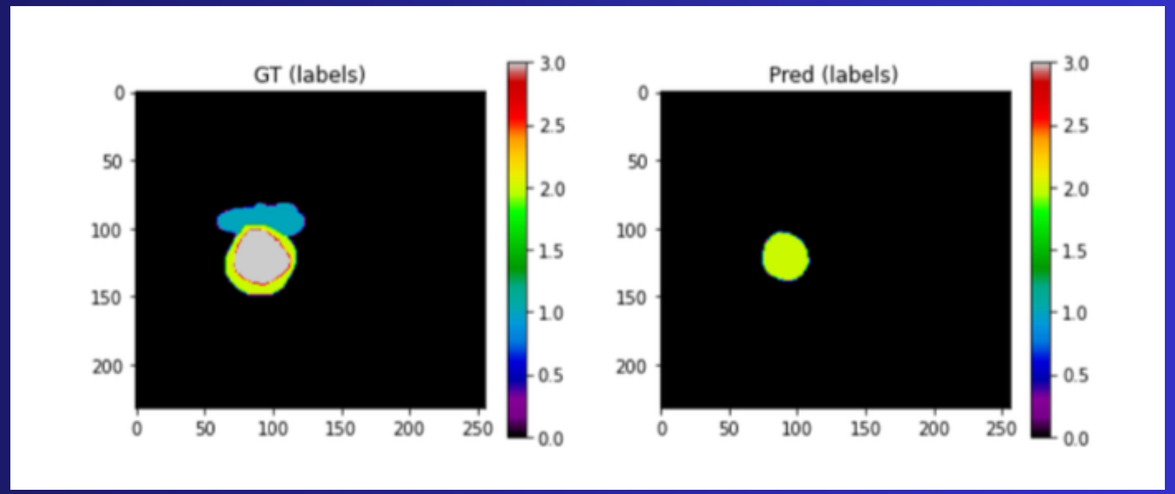
- **MAD – Mean Absolute Difference**
  Measures the average absolute pixel-wise difference between two images.



MAD: 0.0548

🧪 Case Categorization of LV

Empty Prediction – <35 total foreground pixels
Only LV – LV present, MYO & RV absent
Only RV – RV >100px, LV <10px, MYO ≤20px
LV → MYO Misclassification – ≥10% LV misclassified as MYO
Extreme LV → MYO Misclassification – >50% LV misclassified as MYO
LV + MYO – Both inside LV mask, ≥70% MYO within LV
Every Class – LV, RV, and MYO all appear inside LV mask
Error – fallback if thresholds not met

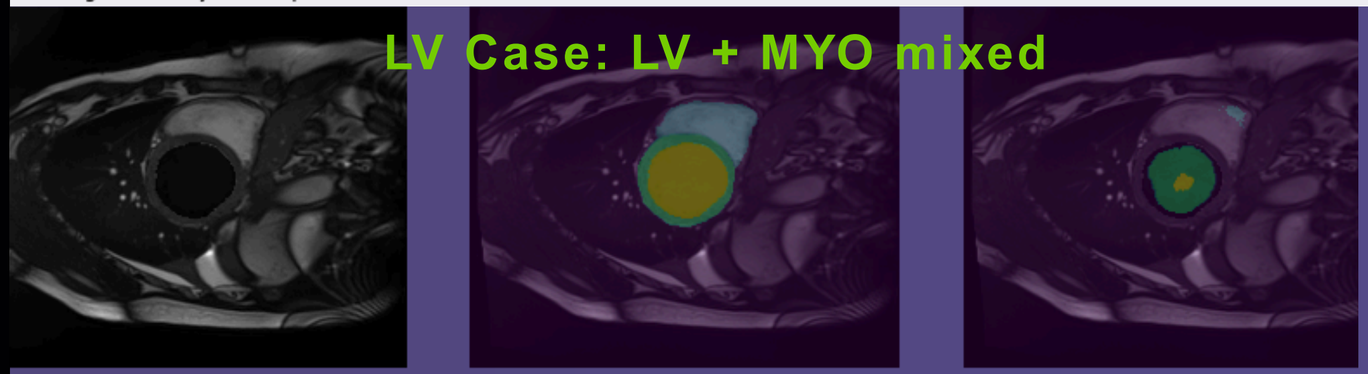


LV Selected Cases

LV Not Selected Cases

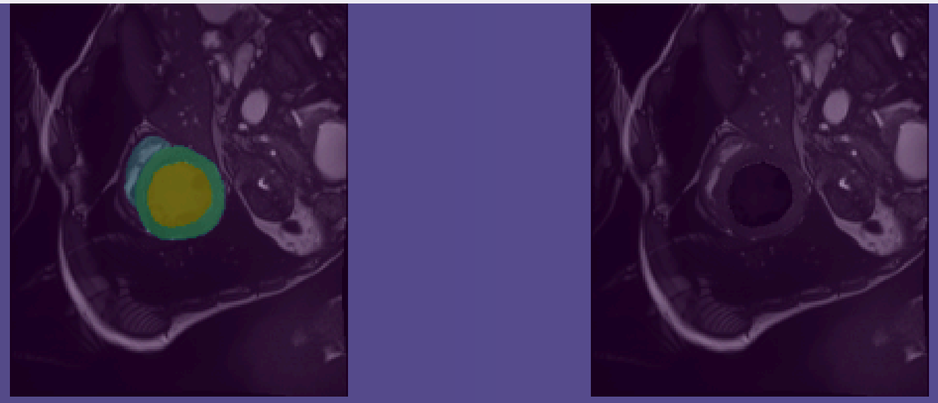
patient001_frame01 – Bright Slice 2
Dice=0.1042 | Group: LV + MYO Mixed | LV fully inside and MYO mostly inside GT LV (rv=73), lv_total=91, myo_total=1124, myo_inside_ratio=1.00
GT: {0.0: 51506, 1.0: 1254, 2.0: 880, 3.0: 1656} | Pred: {0: 54008, 1: 73, 2: 1124, 3: 91}
LV→MYO frac=0.68

Brightness-Adjusted Input | Ground Truth | Prediction



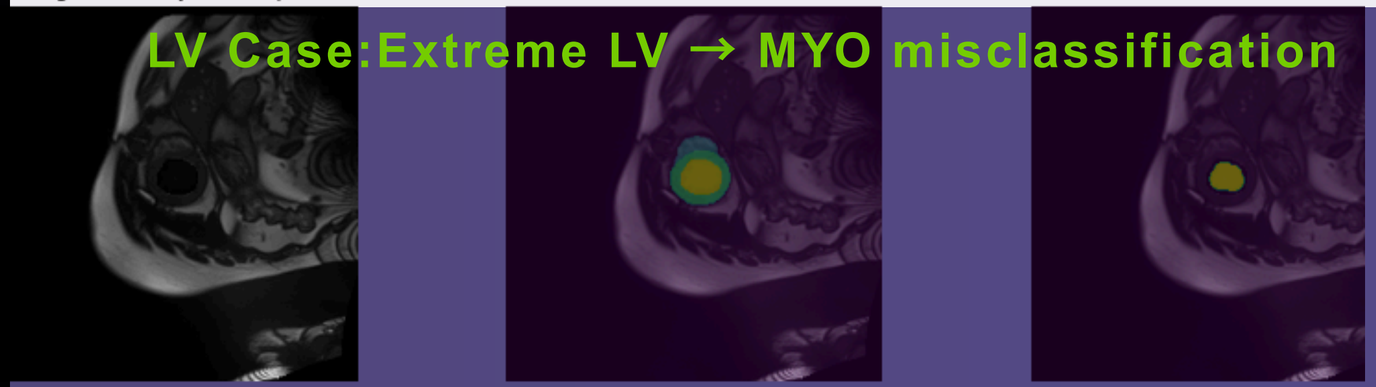
LV Case: LV + MYO mixed

LV Case: Empty Prediction



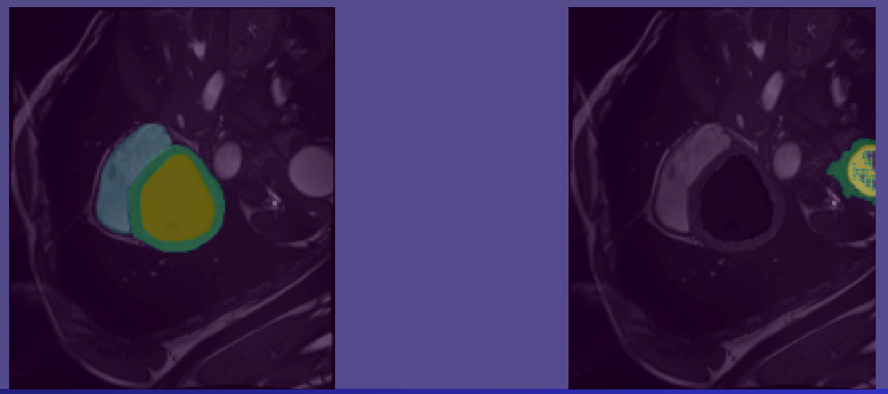
patient003_frame15 – Bright Slice 8
Dice=0.0000 | Group: Extreme LV→MYO Misclassification | 372 LV pixels misclassified as MYO (0.67 of GT LV)
GT: {0.0: 64086, 1.0: 254, 2.0: 641, 3.0: 555} | Pred: {0: 65164, 2: 372}
LV→MYO frac=0.67

Brightness-Adjusted Input | Ground Truth | Prediction

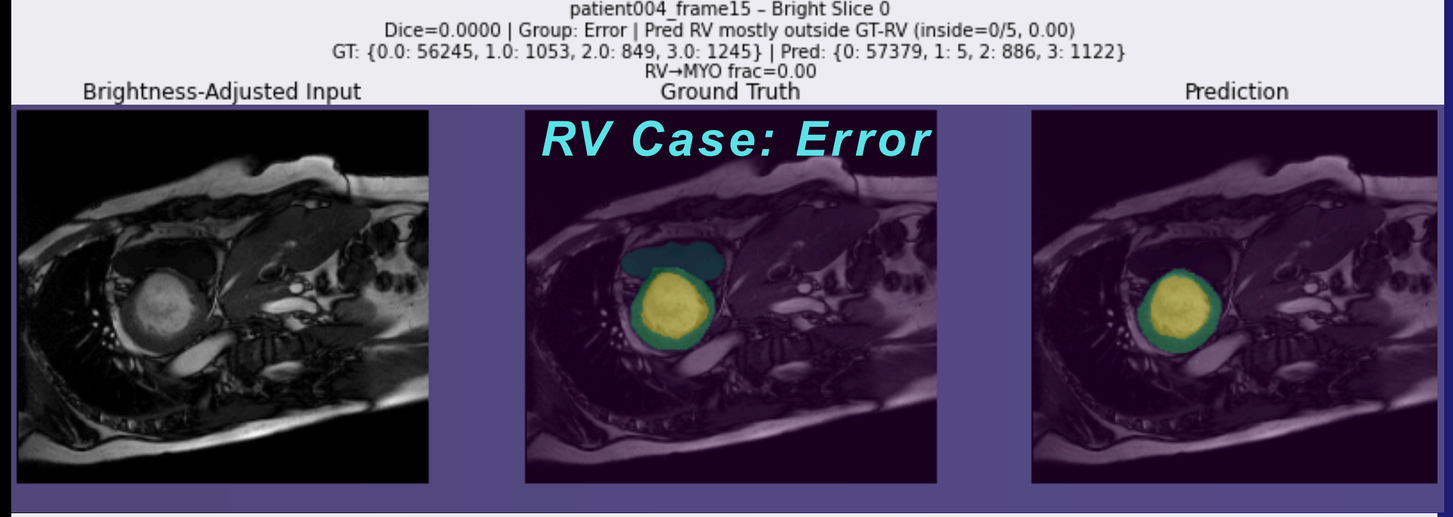


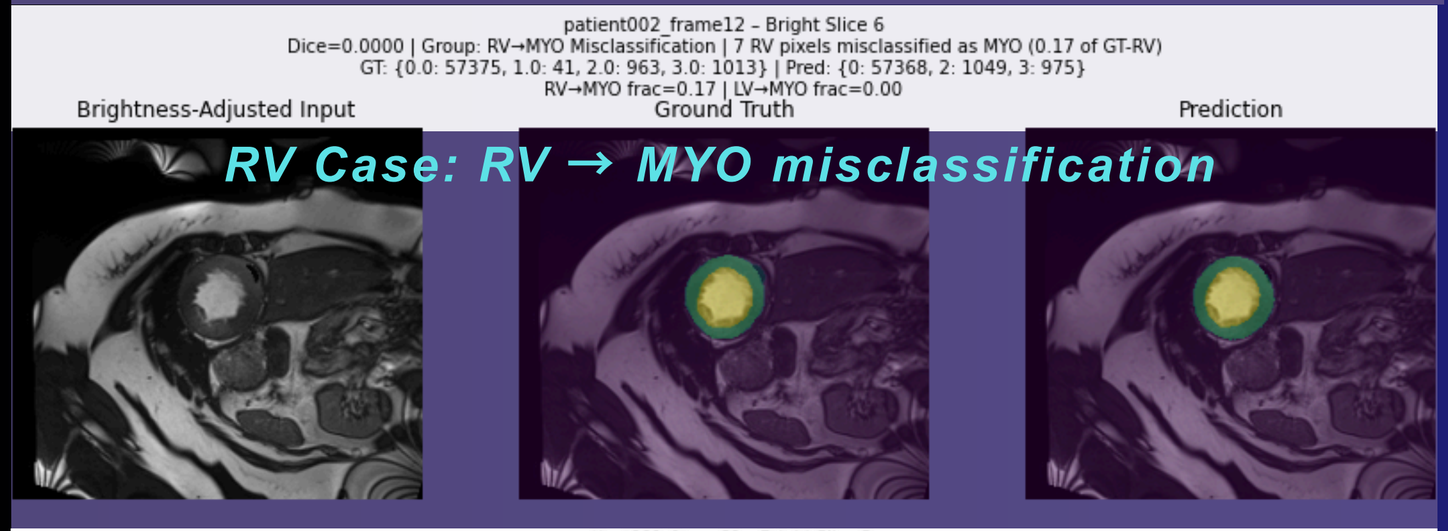
LV Case:Extreme LV → MYO misclassification

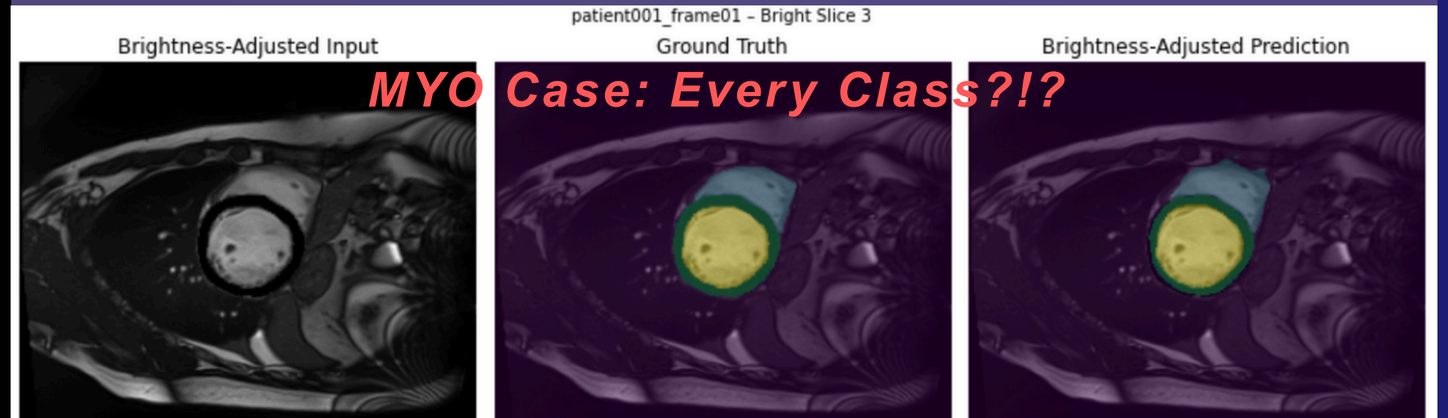
LV Case: Error



Where are the Cases for RV and MYO?



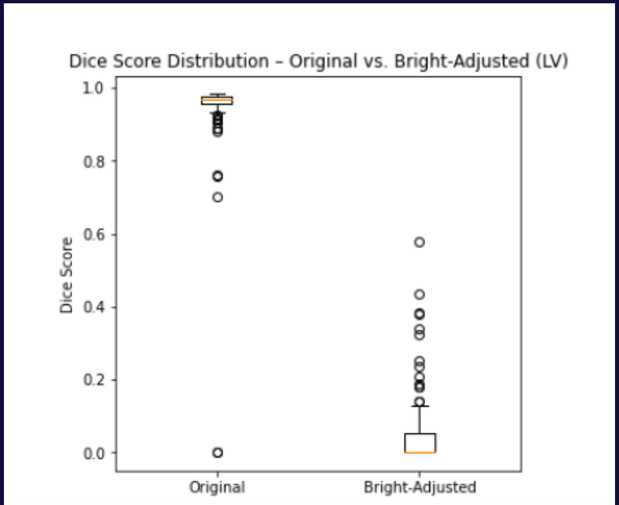


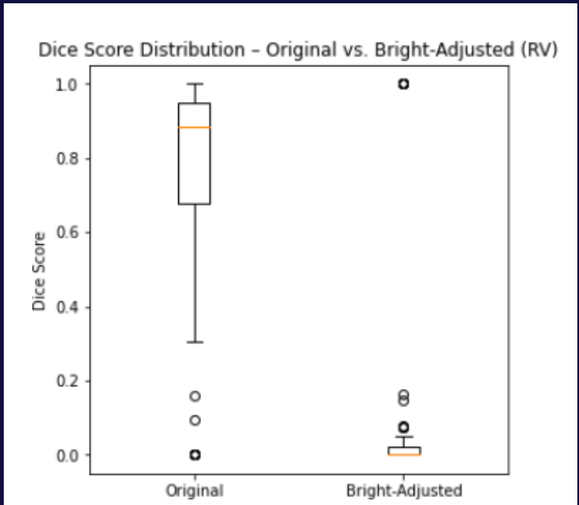


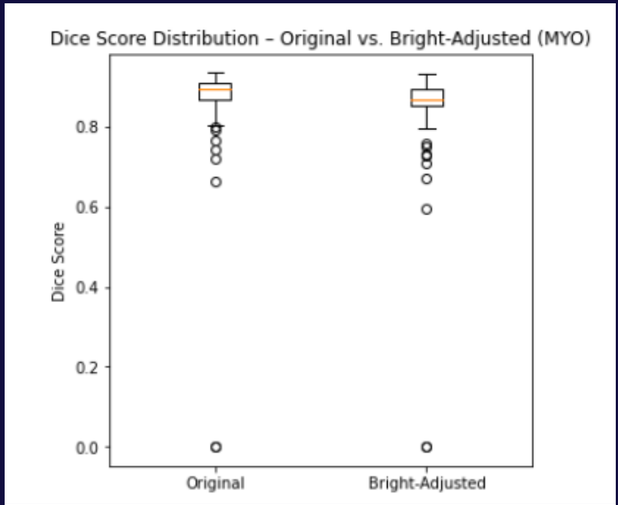
7 pixels misclassified as MYO but it looks almost same as the case Error

We see every classes in MYO segmentation results

Dice Score Distribution - Original vs Adjusted







LV Dice Score Change

- Original: 0.94 → Adjusted: 0.05
- Drop ≈ −0.88
- Almost complete segmentation failure

RV Dice Score Change

- Original: 0.73 → Adjusted: 0.12
- Drop ≈ −0.61
- Strong disruption, but not total failure

MYO Dice Score Change

- Original: 0.86 → Adjusted: 0.84
- Drop ≈ −0.02
- Largely robust, only few outlier failures

LV and RV are highly sensitive, MYO is comparatively robust against darkening.

Calculation of the Top Feature Maps

Goal

Identify model activations sensitive to brightness adjustments.

Class-Specific Analysis

- LV: compare with case slices (e.g., LV→MYO misclassification)
- MYO & RV: compare frequencies across all three classes

Method

- Under each layer, for each patients and their slices, select Top-3 MAD channels
  - Top-1 = too noisy
  - Top-5+ = too diluted
  - Top-3 = balance
- Count how often each channel appears overall → Top-7 most frequent per layer

```
for each layer:
    for each patient, each slice:
        compute MAD(channel) for all channels
        select Top-3 channels with highest MAD
        record these channels

for each layer:
    count frequency of channels across all patients & slices
    select Top-7 most frequent channels

for each case (e.g., LV+MYO, Extreme LV→MYO):
    repeat Top-3 selection restricted to case slices
    compute Top-7 frequencies within this subset
```

Insights

- General distribution vs case-specific distribution
- Shows which channels are:
  - 🔹 Consistently responsible for class detection
  - 🔹 Most sensitive to brightness adjustments

Evaluation of the Top Channel Maps

LV Frequency Map

Figure 1

LV+MYO case

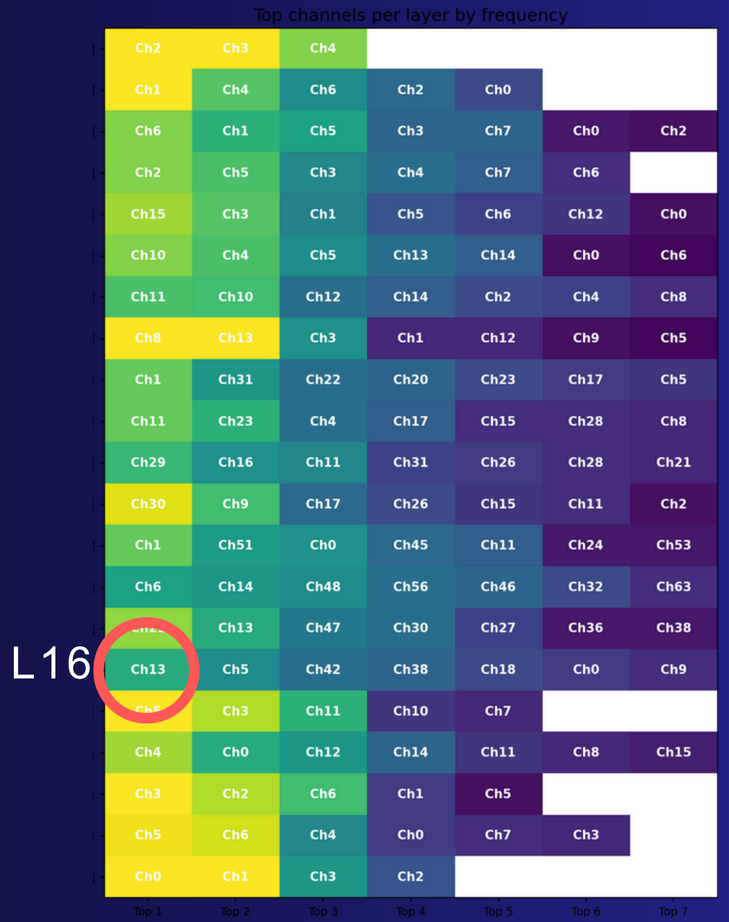


Figure 2

Extreme LV-MYO misclassification

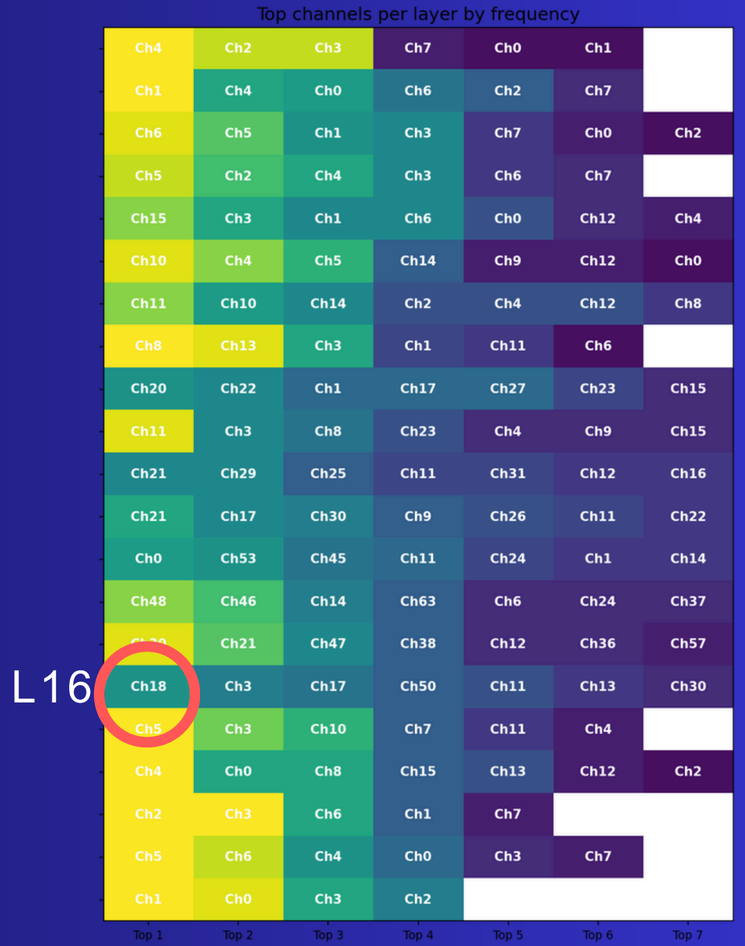
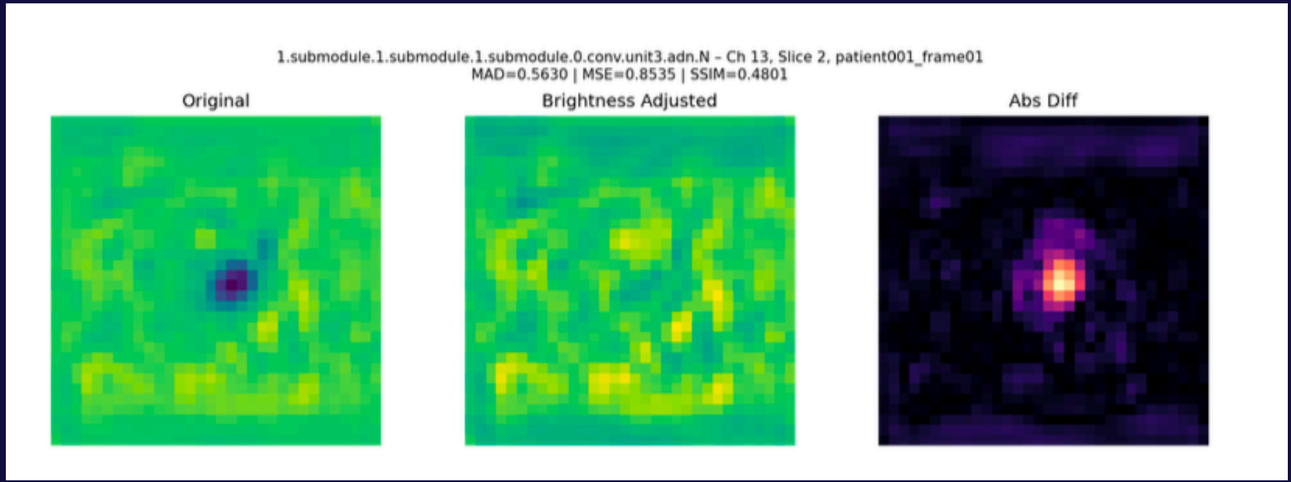
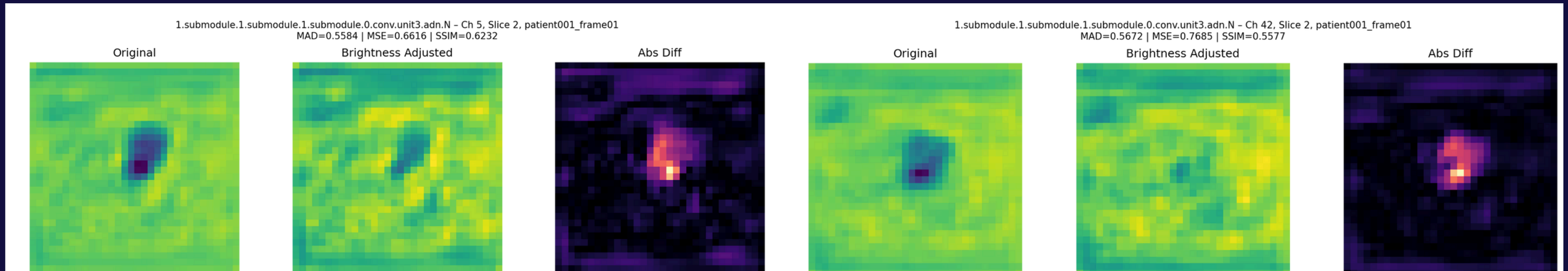


Figure 3

LV-Brightess Sensitive LV class dependent

L16–Ch13, L16–Ch42, and L16–Ch05





LV Frequency Map

Figure 1

RV Frequency Map

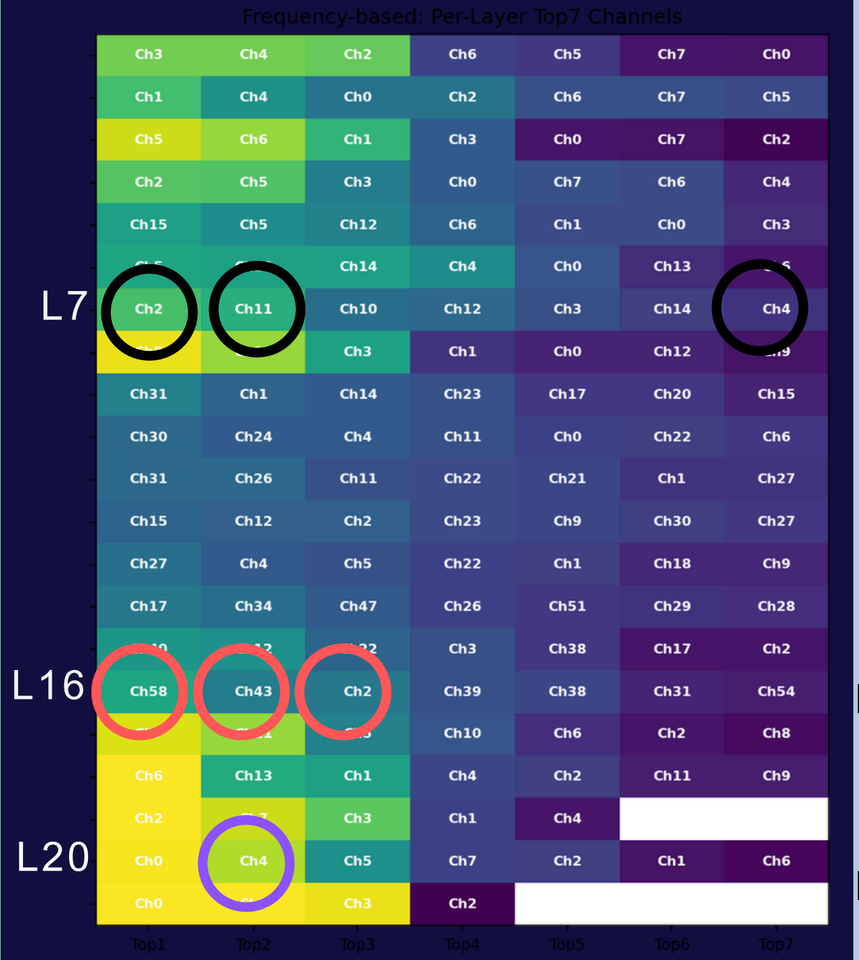


Figure 4

MYO Frequency Map

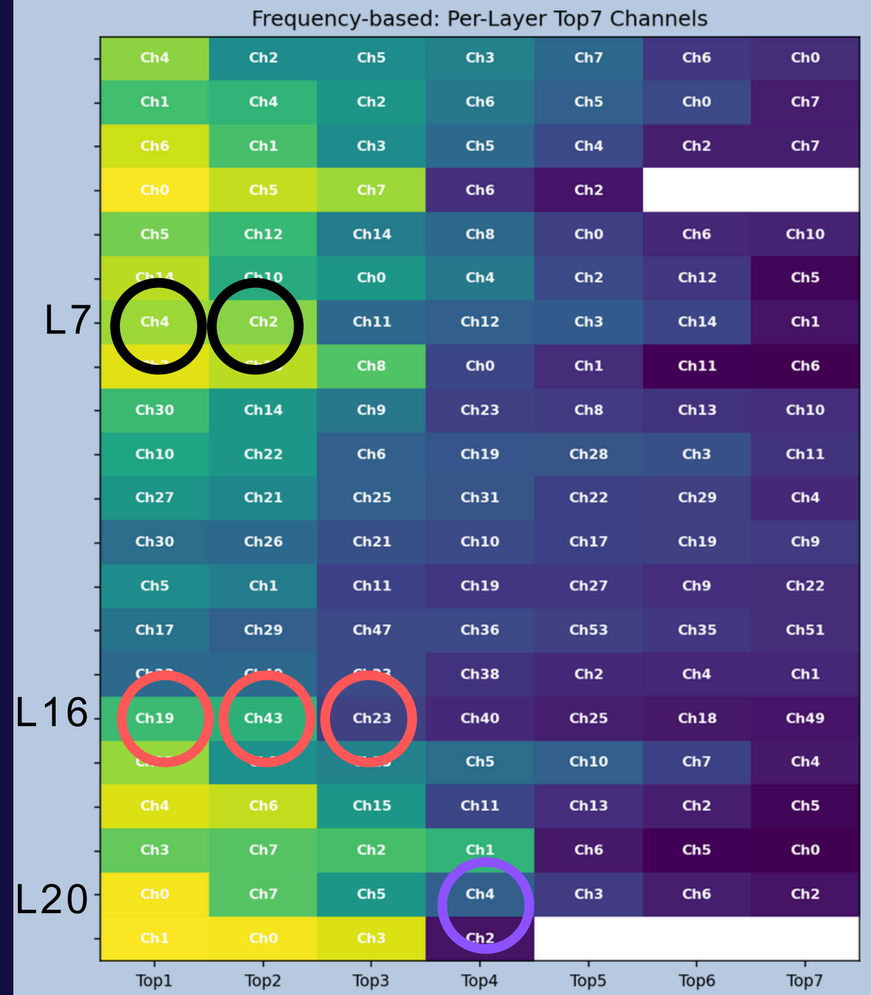
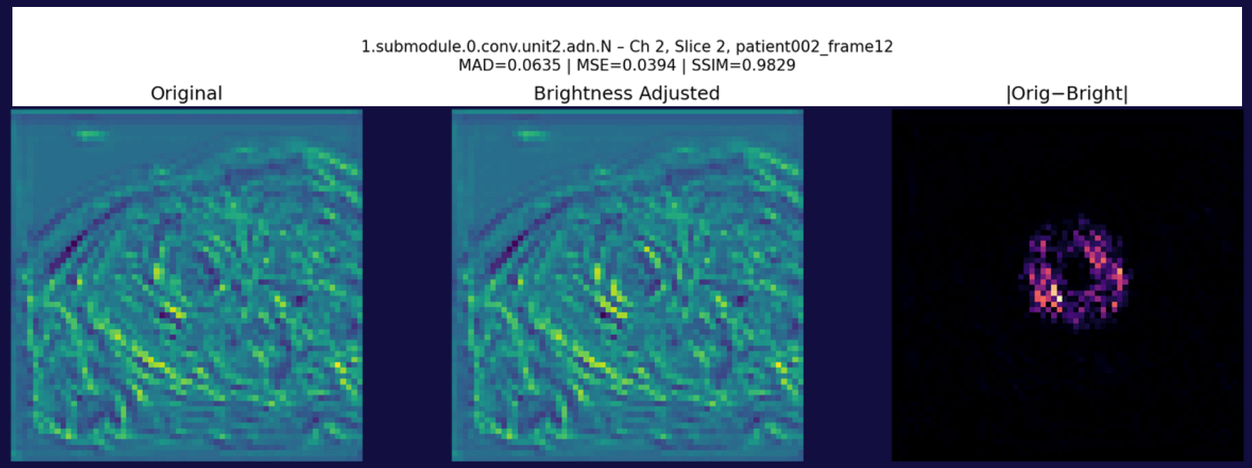
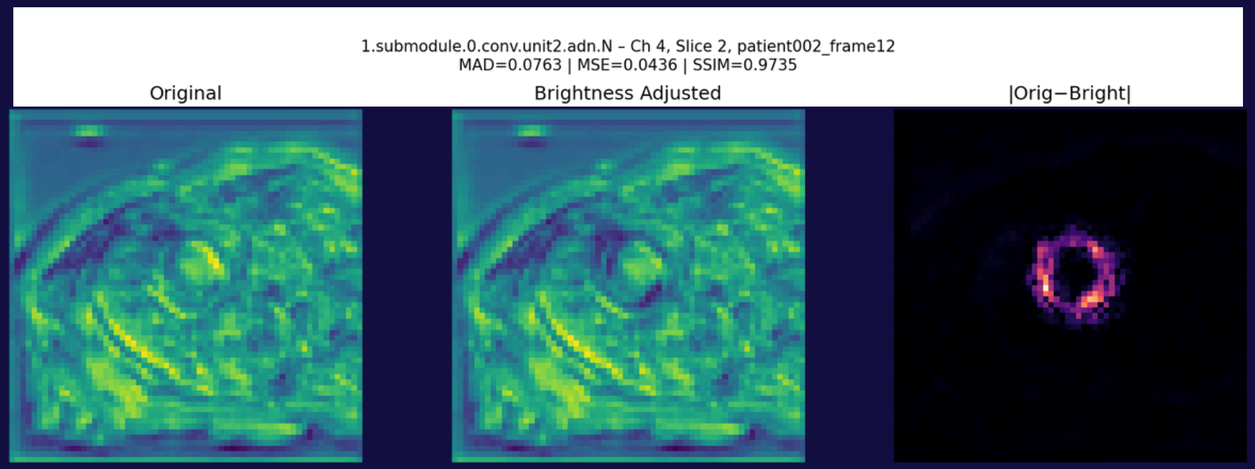


Figure 5

MYO-Brightess Sensitive MYO class dependent

L7-Ch04





Thank you for listening!

Why do we have brightness specific channels just for a particular patient group (patient1 frame1 and frame12)?

Some layers have different MAD value ranges what does this indicate?

References

Bertolini, M., Rossoni, M., & Colombo, G. (2021). Operative Workflow from CT to 3D Printing of the Heart: Opportunities and Challenges. Bioengineering, 8(10), 130.
Bertolini, M., Rossoni, M., & Colombo, G. (2021). Operative Workflow from CT to 3D Printing of the Heart: Opportunities and Challenges. Bioengineering, 8(10), 130.