Beyond Classification: Elaborating Network Predictions for Better Weakly Supervised Quantization

Chih-Chieh Chen[1] JACKFRANK@GMAIL.COM
Chang-Fu Kuo[1,2,3] ZANDIS@GMAIL.COM
[1] *Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[2] *Medical Education Department, Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[3] *Division of Rheumatology, Allergy and Immunology, Chang Gung Memorial Hospital, Taoyuan, Taiwan*

Editors: Under Review for MIDL 2026

Abstract

For clinical applications, more detailed information such as specific locations and ROI volumes is preferred. However, most of the time only classification annotations are available. Class Activation Mapping (CAM) and its variants are the most commonly used techniques for weakly supervised localization tasks. In this study, we assessed both traditional and modern network architectures regarding classification accuracy and CAM visualization. Although all networks achieved high AUROC scores and their heatmaps closely corresponded to pathology locations, we observed that the heatmaps were influenced by the particular network architectures and pretrained weights used. Additionally, current models produce heatmaps from small latent spaces (e.g. 16×16), which limits the precision of these heatmaps for further detailed analysis.

Based on the observations mentioned above, we designed a UNet-style architecture that utilizes pretrained classification networks as the encoder and produces heatmaps within a latent space of size 128×128. We observed that the generated heatmaps are more detailed and suitable for weakly supervised segmentation. We validated the effectiveness of our approach using the intracerebral hemorrhage (ICH) dataset.

Keywords: ICH classification, heatmap generations, weakly supervised localization.

1. Introduction

The localization of pathologies is just as crucial as their classification for clinical use. By utilizing heatmaps or bounding boxes, clinicians can revisit patient data to verify the presence of pathologies in specific areas, thereby lowering the false negative rates in diagnoses. However, creating bounding box annotations is time-consuming. Consequently, it is valuable to explore how weakly supervised methods can minimize the need for extensive labeling. In medical image analysis, (Zhou et al., 2016; Selvaraju et al., 2017) are currently the most widely used techniques. Given a neural network classifier, a target class c, and a chosen layer A, which is usually the layer before the final global average pooling layer, (Zhou et al., 2016; Selvaraju et al., 2017) generate the heatmap of the chosen layer by weighted summing the channel values, where the weight of the channel k is given as the summations of partial derivatives of final outputs with respect to $A_{i,j}$, the layer value at location (i, j). Due to the spatial biases inherent in convolutional neural networks, locations corresponding to class-related features tend to have higher output values and growth rates, resulting in elevated heatmap scores at these positions.

In this study, we explore the application of a GradCAM-based method for weakly supervised quantification in a clinical context. We assessed the GradCAM approach on quantifying subtypes of intracranial hemorrhage (ICH). The five subtypes considered are intraparenchymal hemorrhage (IPH), intraventricular hemorrhage (IVH), subarachnoid hemorrhage (SAH), subdural hematoma (SDH), and epidural hematoma (EDH). Typically, these subtypes can be differentiated by their distinct locations and shapes (Heit et al., 2016). Our goal is to evaluate how effectively GradCAM-generated heatmaps can identify the locations of these subtypes and measure the volumes of the corresponding hemorrhages.

Our primary focus is on two groups of architectures: ResNet (He et al., 2016), DenseNet (Huang et al., 2017), along with their variants such as ResNet-RS (Bello et al., 2021) and RDNet (Kim et al., 2024). We trained these models to classify ICH subtypes, and all achieved AUROC scores exceeding 0.95. For ICH subtype localization, we observed that all models could produce heatmaps accurately pinpointing the locations of each subtype, especially when multiple subtype features are completely separate (see Fig. 1 for more details.)

Regarding the estimation of stroke volume subtypes, as shown in Fig. 2, we found there is room for improvement. With ResNet152, the produced heatmaps lack strong location sensitivity, occasionally missing large strokes. In the case of DenseNet121, likely due to numerous skip-connections, the heatmaps are focused but the regions of high intensity are too broad for detailed features such as SAH. When training DenseNet from scratch without the ImageNet pretrained weights (referred to as DenseNet* in Fig. 2), the heatmaps become less focused, indicating that pretraining on a large and diverse image set aids heatmap quality. For newer models like ResNet-RS and RDNet, the heatmaps are more detailed; however, since they are derived from low-resolution latent spaces, their representational capacity is limited. Additional examples can be found in Fig. 6.

Modern architectures achieve impressive discriminative performance with fewer parameters and FLOPs by incorporating modern components like the squeeze-and-excitation block (Hu et al., 2018) and depthwise convolutions, along with improved training and scaling methods. Building on this insight, our work aims to leverage features from hidden layers to produce more detailed heatmaps. Our approach involves applying GradCAM to higher-resolution latent spaces. To do this, we first train the model on the target dataset using these modern architectures. Next, we construct a U-shaped network that uses the pretrained model as the encoder. Since segmentation labels are unavailable, inspired by (Pinheiro and Collobert, 2015), we replace the final average pooling and linear layers with a Log-Sum-Exp (LSE) pooling layer, which sums the exponentials of patch outputs to generate predictions. This means that the weights for each patch are updated in proportion to the exponential of its output (see Section 2.3 for details). We also incorporate intermediate supervision and regularization on each layer's output to ensure the generated heatmaps align with human intuition. The decoder's architecture is simple, and our intuition is to use the pretrained encoder to guide the nearly linear, untrained decoder in producing more fine-grained heatmaps.

We demonstrated the effectiveness of our methods using ICH datasets. Our model was trained on the RSNA intracranial hemorrhage dataset (Flanders et al., 2020) and its performance was assessed on the BHSD dataset (Wu et al., 2023). Our empirical results indicate that although the classification accuracy of our proposed architecture is comparable

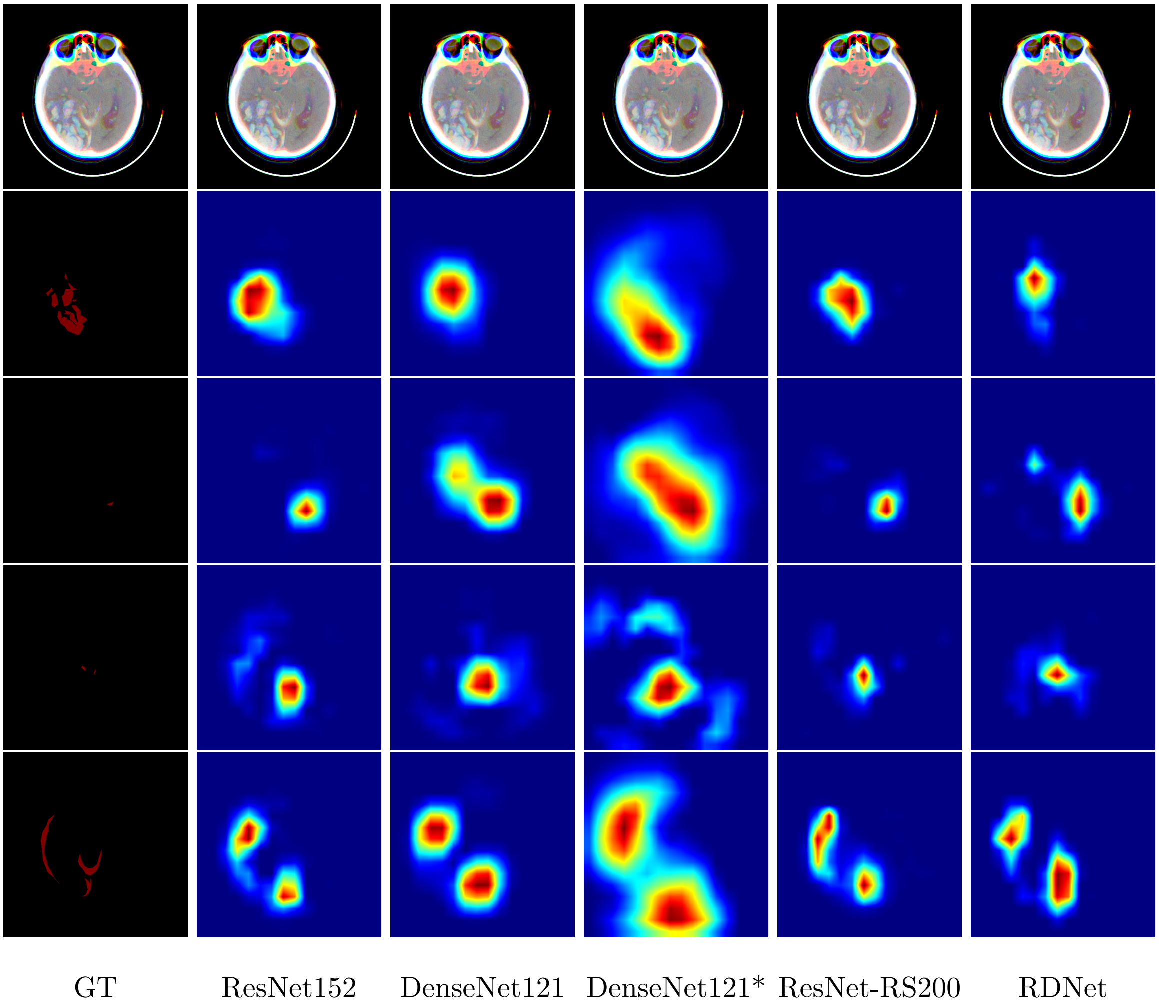

Figure 1: Heatmap generations for ICH subtypes from different network architectures. From the above to below: IPH, IVH, SAH, SDH.

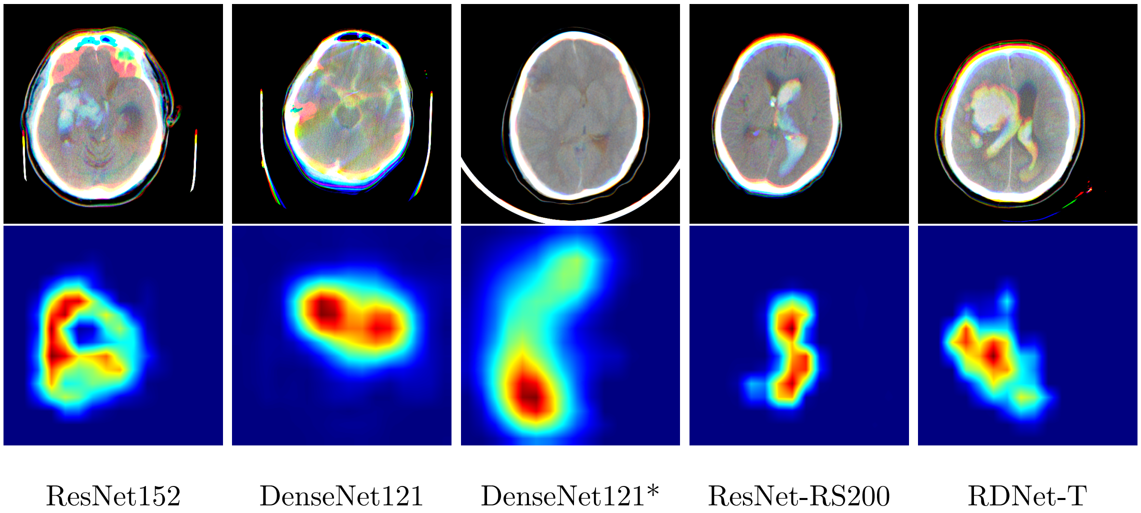

Figure 2: Illustrations of heatmap generations of different network architectures.

to or surpasses that of standard models, the heatmaps produced by the GradCAM algorithm are significantly more accurate.

In sum, our contributions are as follows:

- We empirically highlight both the strengths and limitations of the GradCAM algorithm when applied to clinical tasks.
- To produce more detailed heatmaps, we introduced a method that also leverages the shallow layers of a pretrained network by constructing a UNet-like architecture and generating heatmaps from the final layer of this U-shaped network.
- We evaluated our proposed architecture on the classification of intracranial hemorrhage subtypes, showing that it achieves comparable classification performance while producing more precise heatmaps.

2. Related Works

2.1. ICH related Tasks

In recent years, deep learning-based approaches have been proposed. (Kuo et al., 2019) developed a joint classification and segmentation framework based on 4,396 CT scans and demonstrated a case-level sensitivity of 100% and specificity of 90% on 200 randomly selected CT scans. In their related work (Kuo et al., 2018), DICE score of 76.6% was reported. In a similar study (Monteiro et al., 2020), edema was included as an additional ground truth class. A simplified annotation was used, merging SDH, EDH, and SAH into a single category named extra-axial hemorrhage (EAH). The testing dataset, which excludes lesions of 1 mL or smaller, yielded a case-level mean Dice score of 59.3% for ICH.

Although these methods yield promising results, they require extensive fine-grained annotations. Therefore, it is desirable to create techniques that demand less annotation effort.

2.2. GradCAM

Given a classifier ending with a global average pooling layer followed by a fully connected layer, the final output corresponding to class c is given by

$$y^c = w_l^c \cdot \frac{1}{Z}\Sigma_i\Sigma_j A_{i,j}^l, \tag{1}$$

where $A_{i,j}^l$ represents the output of the l-th feature at spatial location (i,j) in the last convolutional layer and w_l^c is the weight of the fully connected layer associated with class c for unit l. (Zhou et al., 2016) proposed to generate heatmaps M^c as

$$M_c(i,j) = \Sigma_l w_l^c A_{i,j}^l. \tag{2}$$

(Selvaraju et al., 2017) further generalizes the idea to arbitrary neural networks by letting

$$w_l^c = \frac{1}{Z}\Sigma_i\Sigma_j \frac{\partial y^c}{\partial A_{i,j}^l}, \tag{3}$$

and the equivalence of these two approaches when the target layer is right before the final global classification layer is also demonstrated in (Selvaraju et al., 2017).

Lastly, to create visualizable heatmaps, M_c in Equation 2 will be passed through a ReLU layer and further normalized by its maximum value.

2.3. LSE Pooling Layer

Let

$$LSE_r(x_1, \ldots, x_n) = \log(\Sigma_i \exp(r \cdot x_i)) \tag{4}$$

be the scaled LogSumExp function. Recall that the partial derivatives

$$\frac{\partial LSE_r(x1, \ldots, x_n)}{\partial x_j} = \frac{\exp(r \cdot x_j)}{\Sigma_i \exp(r \cdot x_i)}. \tag{5}$$

Thus, when the global average pooling layer is replaced by the LSE pooling layer, according to Equation 3, the contributions to the weight w_l^c with respect to the spatial locations are proportional to their scaled softmax values. This implies that the final heatmap scores at each spatial location are proportional to the product of their patch-level outputs and the scaled softmax values. Compared to heatmaps generated by networks using global average pooling, these heatmaps are significantly more sensitive to patch-level outputs.

3. Proposed Architecture and Algorithm

3.1. Architecture

The main proposed architecture is illustrated in Fig. 3. It is a UNet-like architecture with encoder that is first pretrained on the target dataset. For the encoder backbone, we use RDNet-T (Kim et al., 2024), an updated version of DenseNet designed to enhance accuracy and scaling strategy compared to the original. Modern architectures like patchification stem, base block from ConvNeXt (Liu et al., 2022) are adopted. And the transition layer is redesigned and applied after every three blocks instead of only after each stage. Working on images with size 512 × 512, we added additional parameters at the end of each stage, orresponding to spatial dimensions of $128 \times 128, 64 \times 64, 32 \times 32$, and 16×16, with channel sizes of $256, 440, 744, 1040$, respectively.

For the decoder, each stage primarily includes three components: a convolution block, an upsampling block, and the final classification layer. The convolution block applies 1D convolution to convert the concatenated features into the desired channel dimension, followed sequentially by a Layer Normalization layer, a ReLU activation, and a squeeze-and-excitation block. The upsampling block first uses 1D convolution to adjust the channel dimension to match that of the previous stage, then applies 2D convolution to increase the number of channels by four times, and performs upsampling using the depth-to-space operation. A Layer Normalization layer is added at the end of the upsampling block. Lastly, the final classification layer produces multi-label scores, which are then fed into the LSE pooling layer. Heatmaps are generated just before the top LSE pooling layer. As discussed in Section 2.3, when patch-level predictions are trained with high accuracy, the resulting

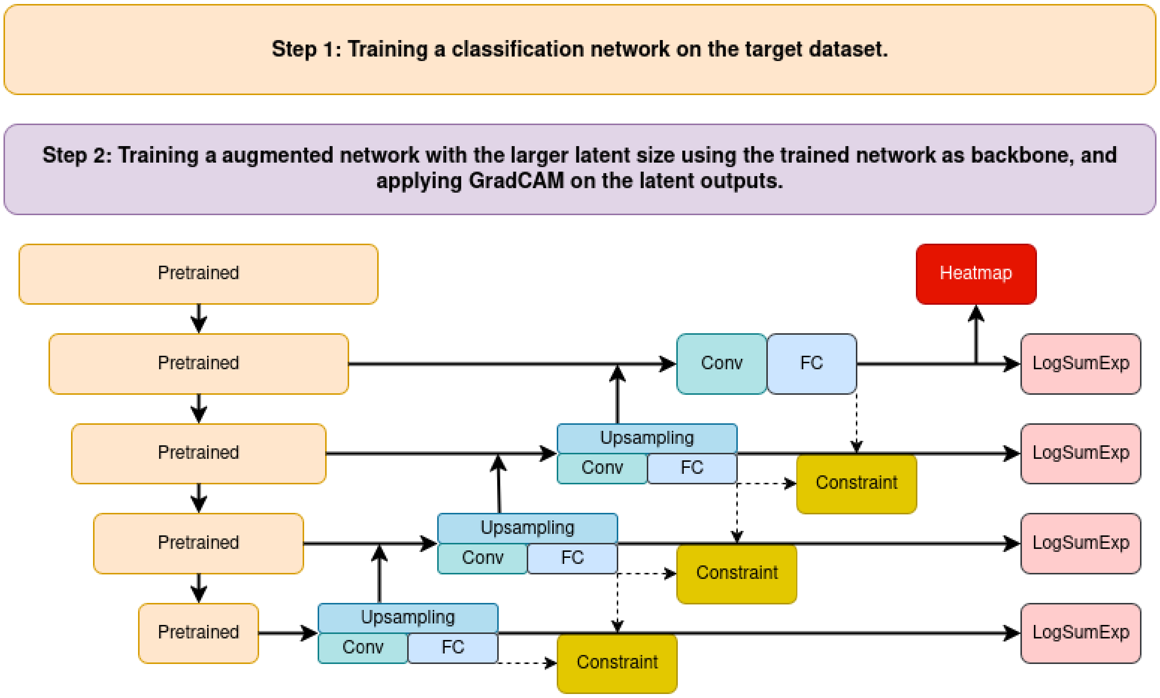
Step 1: Training a classification network on the target dataset.
Step 2: Training a augmented network with the larger latent size using the trained network as backbone, and applying GradCAM on the latent outputs.
Pretrained
Pretrained
Pretrained
Pretrained
Pretrained
Heatmap
Conv
FC
LogSumExp
Upsampling
Conv
FC
Constraint
LogSumExp
Upsampling
Conv
FC
Constraint
LogSumExp
Upsampling
Conv
FC
Constraint
LogSumExp

Figure 3: Proposed Architecture

heatmaps tend to be sharper compared to those produced by networks that use global average pooling layers.

3.2. Loss Function

Let $f_0, \ldots, f_3$ represent the inputs of the LSE pooling layer, and $l_0, \ldots, l_3$ the outputs following the LSE pooling layer, from above to below. Denote gt the ground truth label. Then we have the classification loss

$$\text{CLoss} = \alpha_0 \text{BCE}(l_0, gt) + \alpha_1 \text{BCE}(l_1, gt) + \alpha_2 \text{BCE}(l_2, gt) + \alpha_3 \text{BCE}(l_3, gt), \tag{6}$$

where BCE represents the binary cross entropy loss function and $\alpha_0, \ldots, \alpha_3$ are weight coefficients. Additionally, to ensure consistency in heatmap generation at each level, a heatmap consistency loss is applied,

$$\text{HLoss} = \beta_0 \text{MSE}(\text{Pool}(f_0), f_1) + \beta_1 \text{MSE}(\text{Pool}(f_1), f_2) + \beta_2 \text{MSE}(\text{Pool}(f_2), f_3), \tag{7}$$

where MSE represents the mean square error loss, Pool refers to average pooling with a size of 2, and $\beta_0, \ldots, \beta_2$ are weight coefficients. The overall loss is the sum of the classification loss (CLoss) and the heatmap consistency loss (HLoss).

3.3. Weakly Supervised Quantization Algorithm

Based on the observations that the generated heatmaps for different disease subtypes are nearly mutually exclusive, in this work, we propose to generate segmentation masks by initially generating heatmaps for subtypes identified as positive by neural network models. Then, heatmap scores below a certain threshold are set to zero. Finally, for each pixel, the predicted mask is determined by selecting the subtype with the highest score. The pseudocode outlining this process is provided in Algorithm 1.

Algorithm 1: Pseudocode for Subtype Segmentation Mask Generation

```
Input: Class thresholds $c_{\mathrm{IPH}}, \ldots, c_{\mathrm{EDH}}$. Mask thresholds $m_{\mathrm{IPH}}, \ldots, m_{\mathrm{EDH}}$. Neural
       network $N$, image $I$, image size $(h, w)$.
Output: Segmentation masks $M$
$M \leftarrow$ np.zeros((6, h, w))
for $i \leftarrow$ ICH subtypes do
    if $N(I)[i] > c_i$ then
        $M[i] \leftarrow$ np.where(GradCAM$(I) > m_i$, GradCAM$(I), 0)$;
end
$M \leftarrow$ np.argmax(M, 0)
return $M$
```

4. Experiments

In this work, we validate our proposed approach for intracranial hemorrhage subtype classification. In 2019, the RSNA hosted a competition centered on classifying acute intracranial hemorrhage subtypes (Flanders et al., 2020), providing over 20,000 CT scans along with slice-level classification labels. The class annotations include five primary ICH subtypes: intraparenchymal hemorrhage (IPH), intraventricular hemorrhage (IVH), epidural hematoma (EDH), subdural hematoma (SDH), and subarachnoid hemorrhage (SAH). We randomly selected $2,734$ CT scans ($91,844$ slices) for training, and 89 CT scans ($3,000$ slices) for validation of subtype classification. The class thresholds used in Algorithm 1 were chosen based on the threshold that maximized the difference between true positive rate and false positive rate on the validation set. For testing, we utilized the BHSD dataset (Wu et al., 2023), which contains 192 CT scans with subtype segmentation labels, as the test dataset.

For image input, we followed the approach used by (Kuo et al., 2019): all CT slices were adjusted with a window width of 130 and a window center of 25. We opted to stack three consecutive slices as input and use the segmentation mask of the middle slice as the output, instead of employing a three-dimensional model. To the best of our knowledge, using three consecutive slices is generally sufficient for accurately detecting and classifying subtype hemorrhages in almost all cases, except for a small number of instances that require distinguishing between EDH and SDH.

All experiments were carried out using a single Nvidia V100 GPU. For models that utilized pretrained checkpoints, we applied SGD with a step learning rate schedule over 20 epochs. For models without pretrained checkpoints, we employed Adam optimizer with a learning rate of $1e-4$ for 30 epochs. For the hyperparameters in Equations 6 and 7, in this work, we set $\alpha_0 = 2, \alpha_1 = 0.5, \alpha_2 = 0.25, \alpha_3 = 0.125, \beta_0 = 0.5, \beta_1 = 0.25$, and $\beta_2 = 0.125$. Additionally, the scaling factor r in the LSE pooling layer (Equation 5) is set to 3.

5. Results

The AUROCs on the validation dataset are shown in Table 1. Although our method achieves the highest performance, we observed that all models yield similar results. However, as shown in Table 2, our method significantly outperforms all other models across all disease subtypes. The heatmaps generated by our proposed method are illustrated in Fig. 4 and Fig. 5. When compared to the ground truth labels and the heatmaps produced by other network architectures shown in Fig. 2, our method clearly performs better in accurately capturing the shapes and locations of hemorrhages.

Table 1: AUROCs on the validation dataset

Model	**IPH**	**IVH**	**SAH**	**SDH**	**EDH**	**AVG**
ResNet152	0.970	0.997	0.949	0.923	0.941	0.956
DenseNet121	**0.974**	0.997	0.957	0.918	0.954	0.960
DenseNet121*	0.973	0.997	0.960	0.915	0.963	0.961
ResNetRS	0.971	0.996	0.956	0.926	0.956	0.961
RDNet	0.973	0.996	0.948	0.916	**0.975**	0.962
Ours	0.973	**0.998**	**0.959**	**0.934**	0.969	**0.967**

Table 2: DICE coefficients on the BHSD dataset, using the class thresholds of the models to generate heatmaps.

Model	**Normal**	**IPH**	**IVH**	**SAH**	**SDH**	**EDH**	**AVG**
ResNet152	0.992	0.242	0.093	0.056	0.075	0.060	0.253
DenseNet121	0.990	0.256	0.113	0.060	0.070	0.078	0.261
DenseNet121*	0.989	0.260	0.132	0.047	0.076	0.054	0.260
ResNetRS	0.996	0.445	0.182	0.138	0.144	0.158	0.344
RDNet	0.996	0.380	0.211	0.108	0.133	0.141	0.328
Ours	**0.997**	**0.551**	**0.387**	**0.225**	**0.245**	**0.356**	**0.460**

For ICH strokes, all types except IPH are irregular, and SAH lesions are typically small and difficult to capture with low-resolution heatmaps. On the other hand, because we generate heatmaps in a latent space of size 128×128, our approach can better approximate irregular or small, fine-grained shapes compared to other methods. Additionally, we found that heatmaps produced by other models tend to generate many false positives. When using GradCAM solely for visualization or explanation—meaning heatmaps are generated without applying class thresholds but based on ground truth labels—our method still outperforms others, but ResNetRS and RDNet-tiny also show competitive results. Further details can be found in Table 3. We observed that ResNetRS, RDNet, and our model all achieved a

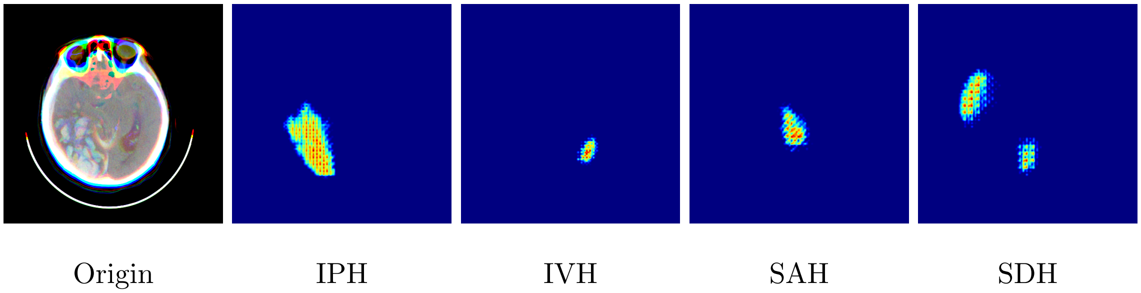

Figure 4: Illustrations of subtype heatmaps generations by our proposed method.

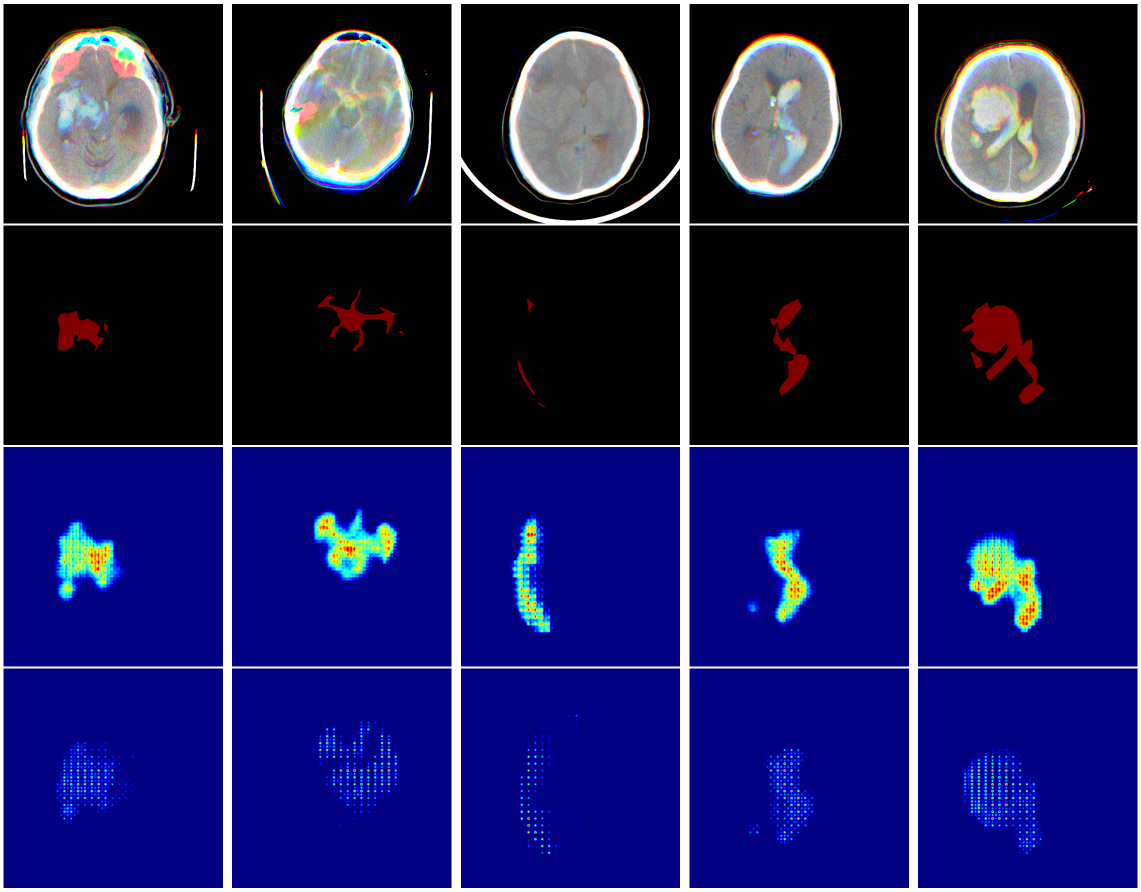

Figure 5: Illustrations of the heatmap generations of ICH by our proposed method. From above to below: original images, ground truths, heatmap generated by our proposed method, heatmap generated by our proposed method without the heatmap consistency loss (HLoss).

Table 3: DICE coefficients on the BHSD dataset, using the ground truth labels to generate heatmaps.

Model	Normal	IPH	IVH	SAH	SDH	EDH	AVG
ResNet152	0.996	0.359	0.116	0.105	0.158	0.281	0.336
DenseNet121	0.995	0.425	0.162	0.122	0.166	0.303	0.362
DenseNet121*	0.993	0.286	0.151	0.084	0.131	0.205	0.308
ResNetRS	0.998	0.552	0.245	0.193	0.243	0.425	0.443
RDNet	0.997	0.502	0.264	0.173	0.234	0.373	0.424
Ours	**0.998**	**0.607**	**0.424**	**0.254**	**0.281**	**0.445**	**0.502**

Dice coefficient above 0.5 for IPH. For our proposed method, the difference in performance between Table 2 and Table 3 is relatively small, indicating the method's robustness.

6. Ablation Study

We want to highlight that the heatmap consistency loss described in Equation 7 is essential. Without this regularization during training, as shown in Fig. 5, the resulting heatmaps tend to have a grid-like pattern and are less focused on the stroke areas.

7. Conclusion

Nowadays, there is a growing number of advanced models and pretrained backbones. Although these models have achieved increasingly better performance on computer vision benchmarks such as ImageNet, we would like to ask: Are these techniques equally beneficial for clinical applications? In this study, we use intracerebral hemorrhage (ICH) subtype classification and visualization as a case example. Surprisingly, we found that while all models showed similar classification accuracy, the quality of the heatmaps they generated varied significantly (See the thresholds in Appendix A for more details). Although every model correctly identified the subtype location, advanced models with pretrained checkpoints produced more detailed heatmaps closely matched subtype stroke segmentation masks.

However, all heatmaps were generated from relatively low-resolution feature maps, for example, 16×16. To harness the full potential of these advanced models and pretrained checkpoints, we developed a decoder for these pretrained models and retrained them on higher-resolution feature maps using the Log-Sum-Exp (LSE) pooling layer. We carefully enforced consistency across latent spaces at each level and ultimately generated heatmaps from the latent space with the highest resolution. Our findings indicate that even without segmentation labels, well-trained classifiers may possess some capacity for quantification. We believe our approach can be useful in situations where training a segmentation model is costly. We believe these results are sufficiently accurate for some situations where clinicians need to include quantified information in radiology reports, but was previously performed using less precise estimation methods. In future research, we intend to conduct a systematic investigation of this approach and extend its application to a variety of disease types.

Acknowledgments

Both authors acknowledge funding from the Center for Artificial Intelligence in Medicine at Chang Gung Memorial Hospital, via grant agreement no. CLRPG3H0016.

References

Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34:22614–22627, 2021.

Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2 (3):e190211, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jeremy J Heit, Michael Iv, and Max Wintermark. Imaging of intracranial hemorrhage. *Journal of stroke*, 19(1):11, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Donghyun Kim, Byeongho Heo, and Dongyoon Han. Densenets reloaded: paradigm shift beyond resnets and vits. In *European Conference on Computer Vision*, pages 395–415. Springer, 2024.

Weicheng Kuo, Christian Häne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Patchfcn for intracranial hemorrhage detection. *arXiv preprint arXiv:1806.03265*, 2018.

Weicheng Kuo, Christian Häne, Pratik Mukherjee, Jitendra Malik, and Esther L Yuh. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proceedings of the National Academy of Sciences*, 116(45):22737–22745, 2019.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Miguel Monteiro, Virginia FJ Newcombe, Francois Mathieu, Krishma Adatia, Konstantinos Kamnitsas, Enzo Ferrante, Tilak Das, Daniel Whitehouse, Daniel Rueckert, David K

Menon, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. *The Lancet Digital Health*, 2(6):e314–e322, 2020.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Biao Wu, Yutong Xie, Zeyu Zhang, Jinchao Ge, Kaspar Yaxley, Suzan Bahadir, Qi Wu, Yifan Liu, and Minh-Son To. BHSD: A 3D Multi-class Brain Hemorrhage Segmentation Dataset. In *International Workshop on Machine Learning in Medical Imaging*, pages 147–156. Springer, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix A. Performance for Pure Visualization

Fig. 6 illustrates examples of heatmaps corresponding to various network architectures. We aim to demonstrate that when the ground truth labels are known (i.e., after eliminating all false negative cases), although our model achieves the highest performance, advanced models such as ResNetRS and RDNet are also suitable for ICH quantification. First, we described how we selected our mask thresholds. We rescaled all heatmaps to the range $[0, 255]$, and then applied grid search on thresholds $m = 25, 50, 75, 100, 125, 150, 175, 200$ to evaluate subtype performance. As shown in Fig. 1, we empirically observed that these subtype heatmaps are mutually exclusive. Therefore, we tested the subtype thresholds jointly. The final mask thresholds are summarized in Table 4.

Table 4: Selected mask thresholds.

Model	$\mathbf{m_{IPH}}$	$\mathbf{m_{IVH}}$	$\mathbf{m_{SAH}}$	$\mathbf{m_{SDH}}$	$\mathbf{m_{EDH}}$
ResNet152	150	150	175	150	175
DenseNet121	175	200	200	175	200
DenseNet121*	200	200	200	200	200
ResNetRS	150	175	150	150	175
RDNet	175	200	175	175	200
Ours	75	125	125	75	25

The best results are already presented in Table 3. We discovered that mask thresholds play a key role in the final heatmap quantization. Some of the results are presented in Table 5, Table 6, and Table 7.

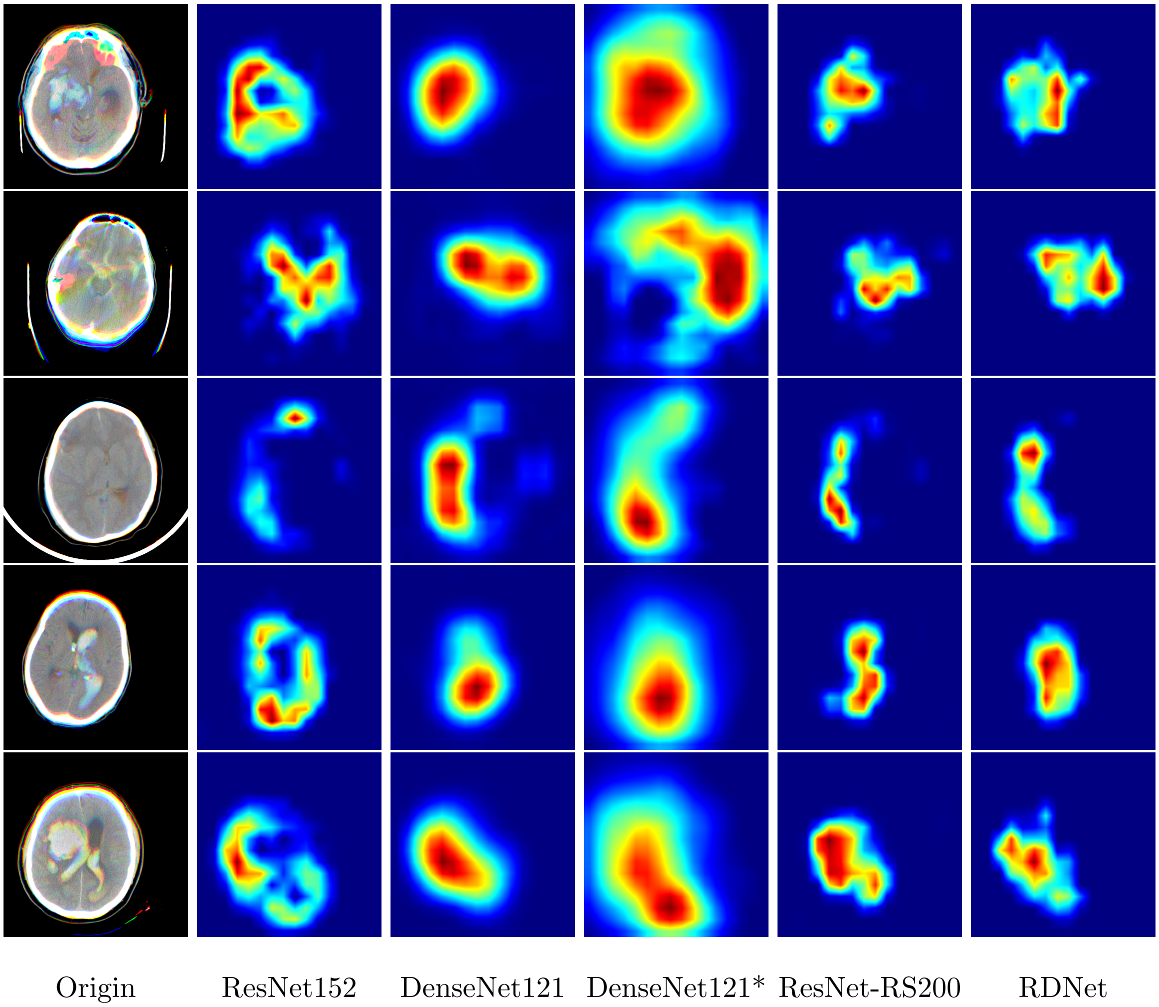

Figure 6: More examples for different network architectures.

Table 5: DICE coefficients on the BHSD dataset, using the ground truth labels to generate heatmaps and with all mask thresholds equal to 125.

Model	**Normal**	**IPH**	**IVH**	**SAH**	**SDH**	**EDH**	**AVG**
ResNet152	0.993	0.354	0.112	0.094	0.161	0.242	0.326
DenseNet121	0.990	0.352	0.106	0.854	0.140	0.186	0.310
DenseNet121*	0.978	0.169	0.057	0.051	0.085	0.123	0.244
ResNetRS	0.996	0.539	0.211	0.183	0.249	0.361	0.423
RDNet	0.994	0.468	0.182	0.143	0.213	0.295	0.383
Ours	0.998	0.576	0.424	0.254	0.236	0.255	0.457

Table 6: DICE coefficients on the BHSD dataset, using the ground truth labels to generate heatmaps and with all mask thresholds equal to 75.

Model	**Normal**	**IPH**	**IVH**	**SAH**	**SDH**	**EDH**	**AVG**
ResNet152	0.988	0.293	0.089	0.071	0.138	0.171	0.291
DenseNet121	0.983	0.263	0.078	0.062	0.108	0.128	0.270
DenseNet121*	0.960	0.100	0.035	0.035	0.062	0.079	0.211
ResNetRS	0.993	0.450	0.158	0.140	0.211	0.265	0.369
RDNet	0.990	0.377	0.126	0.105	0.166	0.212	0.329
Ours	0.997	0.607	0.366	0.207	0.281	0.381	0.473

Table 7: DICE coefficients on the BHSD dataset, using the ground truth labels to generate heatmaps and with all mask thresholds equal to 25.

Model	**Normal**	**IPH**	**IVH**	**SAH**	**SDH**	**EDH**	**AVG**
ResNet152	0.973	0.197	0.057	0.042	0.083	0.099	0.242
DenseNet121	0.967	0.167	0.047	0.037	0.071	0.070	0.226
DenseNet121*	0.924	0.053	0.021	0.022	0.041	0.043	0.184
ResNetRS	0.986	0.322	0.097	0.088	0.143	0.161	0.299
RDNet	0.980	0.262	0.076	0.064	0.101	0.125	0.268
Ours	0.996	0.577	0.293	0.168	0.253	0.445	0.455

Appendix B. Heuristics about the Heatmap Consistency Loss

In this section we would like to provide an intuitive explanation for the heatmap consistency loss in Equation 7. Following the notation in Section 3.2, locally the loss will be

$$f_{i,j}^{k,1} + f_{i,j}^{k,2} + f_{i,j}^{k,3} + f_{i,j}^{k,4} = f_{i,j}^{k+1}, \tag{8}$$

where $f_{i,j}^{k+1}$ is the value of f^{k+1} at spatial location (i,j), and $f_{i,j}^{k,1}, \ldots, f_{i,j}^{k,4}$ are values of f^k lying above (i,j). Now we assume $f_{i,j}^{k,l} >= f_{min}$, for a designed minimum f_{min}. For $r > 0$, we rewrite the LSE function in Equation 4 as

$$\log(\Sigma_{i,j}(\exp(r \cdot f_{i,j}^{k,1}) + \exp(r \cdot f_{i,j}^{k,2}) + \exp(r \cdot f_{i,j}^{k,3}) + \exp(r \cdot f_{i,j}^{k,4})) = \log(\Sigma_{i,j} L_{i,j}^k). \tag{9}$$

On the other hand, from the inequality of arithmetic and geometric means,

$$\frac{L_{i,j}^k}{4} \geq (\exp(r \cdot (f_{i,j}^{k,1} + f_{i,j}^{k,2} + f_{i,j}^{k,3} + f_{i,j}^{k,4}))^{\frac{1}{4}} = \exp(\frac{r \cdot f_{i,j}^{k+1}}{4}), \tag{10}$$

and the unique minimum is achieved when $f_{i,j}^{k,1} = f_{i,j}^{k,2} = f_{i,j}^{k,3} = f_{i,j}^{k,4}$. On the other hand, to find the maximum value, without loss of generalization, we assume $f_{i,j}^{k,1} > f_{i,j}^{k,2} \geq f_{i,j}^{k,3} \geq f_{i,j}^{k,4}$, then we rewrite $f_{i,j}^{k,4} = f_{i,j}^{k+1} - (f_{i,j}^{k,1} + f_{i,j}^{k,2} + f_{i,j}^{k,3})$, then

$$\frac{\partial L_{i,j}^k}{\partial f_{i,j}^{k,1}} = r * (\exp(r \cdot f_{i,j}^{k,1}) - \exp(r \cdot f_{i,j}^{k,4})) > 0, \tag{11}$$

which means that we can increase $L_{i,j}^k$ by increasing $f_{i,j}^{k,1}$ until $f_{i,j}^{k,4}$ achieves the designed lower bound f_{min}. Recursively following the argument, we find the maximum value is achieved when

$$f_{i,j}^{k,1} = f_{i,j}^{k+1} - 3 \cdot f_{min}, \quad f_{i,j}^{k,2} = f_{i,j}^{k,3} = f_{i,j}^{k,4} = f_{min}. \tag{12}$$

In sum, if the spatial location (i,j) belongs to the pathological region, then only one of $f_{i,j}^{k,l}$ achieves the extreme value while the remaining ones tends to achieve the minimum value. On the other hand, if the spatial location (i,j) belongs to the normal region, all the four values tends to achieve the $\frac{f_{i,j}^{k+1}}{4}$, hence the heatmap will be more fine-grained in the outputs of the upper levels.