# Novel Postprocessing Pipeline for Heatmaps in Weakly Supervised Segmentation in PET Imaging

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

**Abstract.** Quantitative FDG-PET/CT analysis depends on accurate tumour lesion delineation, but manual annotation is costly and variable, while fully supervised segmentation requires dense masks that are expensive to obtain at scale. We investigate weakly supervised FDG-PET lesion segmentation from patient-level labels alone, with no segmentation annotations used for classifier training or post-processing parameter tuning. VGG16 and ResNet50 classifiers are trained on multi-angle maximum-intensity projections, and class activation mapping (CAM) is used to obtain lesion localisation heatmaps. To convert these continuous heatmaps into binary masks we propose FLOWS (Fused Localisation-map Watershed Segmentation), an annotation-free post-processing method based on explicit and interpretable operations: multi-scale hotspot detection, paired foreground/background seed generation, and marker-controlled watershed segmentation. FLOWS can operate on CAM heatmaps alone, while also allowing PET intensity to be fused as an additional delineation cue. On the autoPET dataset, Layer-CAM variants provide the strongest localisation among the evaluated CAM methods. Compared with Otsu thresholding, FLOWS improves pixel-wise DSC on true-positive VGG16 scans from 42.0% to 48.0% using the CAM signal alone, with similar performance when PET intensity is fused (48.6% at $\alpha = 0.7$). The main benefit is reduced over-segmentation and fewer false positives relative to single-threshold heatmap binarisation. These results show that interpretable, annotation-free post-processing can improve heatmap-based weakly supervised lesion segmentation without relying on ground-truth masks. Code is available at: https://anonymous.4open.science/status/emerge-2026-3F2B

**Keywords:** Weakly supervised semantic segmentation · FDG-PET · Heatmap · Class activation mapping.

## 1 Introduction

Quantitative FDG-PET analysis requires tumour lesion segmentation. In practice, such masks are produced by manual delineation, which is labour-intensive and subject to inter- and intra-observer variability [18]. Fully supervised methods such as nnU-Net [10] can automate segmentation, but require large collections of dense pixel-level annotations. This annotation bottleneck motivates weakly

supervised semantic segmentation (WSSS), which replaces masks with cheaper supervision. Image-level labels provide the coarsest supervision, indicating only whether the target class is present. In volumetric PET, they can be assigned at different granularities: slice-level labels indicate tumour presence on individual 2D slices, whereas patient-level labels provide a single indicator for the whole scan, and are therefore the weakest and cheapest form, and are often available from routine diagnosis.

Class activation mapping (CAM) [23] is commonly used to recover spatial localisation from classifiers trained with image-level labels. CAM exploits the spatial structure preserved in convolutional feature maps to produce heatmaps that highlight image regions contributing to the classification decision. In medical imaging, CAM-based WSSS has been shown to be feasible with slice-level labels [7, 20, 4, 12] and patient-level labels [1, 16]. However, CAMs provide continuous localisation heatmaps rather than binary lesion masks. More elaborate heatmap generation can improve localisation before mask generation [4, 22], but conversion remains a critical step and existing strategies are limited.

Fixed global thresholds are commonly applied [7, 16, 12, 6], often selected by grid search using ground-truth masks, which reintroduces the annotation dependency that weak supervision aims to avoid and limits transfer across settings. Dynamic thresholds such as Otsu's method [15], as used by [22], avoid supervised tuning by estimating a threshold for each heatmap, making them annotation-free and adaptable. Nevertheless, both fixed and dynamic thresholding reduce each heatmap to a single cutoff, which cannot account for lesions with different activation strengths or spatial extents within the same image. As a result, weak activations may be missed while strong activations are over-segmented. Conditional random fields (CRFs) offer an alternative refinement strategy using spatial and image-intensity consistency, sharpening boundaries before final mask extraction [13]. However, dense CRFs introduce several hyperparameters that commonly require optimisation [19, 3], weakening the annotation-free setting and limiting transferability to FDG-PET.

The remaining challenge is therefore not merely to binarise CAM heatmaps, but to obtain accurate PET lesion masks from them without using ground-truth annotations for either training or post-processing. We address this with FLOWS, Fused Localisation-map Watershed Segmentation, an annotation-free post-processing method that converts CAM heatmaps into binary lesion masks without direct heatmap thresholding. Instead, FLOWS uses explicit, interpretable operations that combine the discriminative localisation of CAMs with PET intensity information for lesion delineation, while retaining a CAM-only pathway. No ground-truth masks are used for classifier training, parameter tuning, or post-processing.

## 2 Materials and Methods

The proposed pipeline produces a binary tumour lesion masks from FDG-PET volumes using only patient-level binary labels (normal/abnormal), with no pixel-

level annotations at any stage, including training, hyperparameter selection, and post-processing. This work examines what lesion segmentation is achievable from the PET signal alone, without introducing CT information.

To reduce training time and cost, the 3D volumes are collapsed into 2D through maximum intensity projection (MIP), enabling the use of established ImageNet-pretrained 2D classification backbones and CAM methods. Computed at 32 evenly spaced angles over 180°, these multi-angle MIPs aggregate the full volume into a single image, so a patient-level label supervises each projection directly, and retain a path back to the volume through reconstruction [1], though the latter is not explored here.

### 2.1 Class Activation Mapping

Two ImageNet-pretrained classifiers are evaluated, VGG16 and ResNet50, with the first convolution adapted to single-channel input using `timm`'s pretrained-weight conversion and the classification head replaced for binary output. These networks are chosen for their established role in CAM evaluation [23, 17, 2, 11, 21, 5] and prior weakly-supervised work [7, 12]. For VGG16, the two hidden fully connected layers are reduced from 4096 to 256 nodes and the first head kernel is resized to collapse the spatial dimensions of the final feature map for input shape $(1, 224, 384)$. In both networks, the first max-pooling layer is removed [7], doubling subsequent feature-map sizes so the deepest layer is $(28, 48)$ in VGG16 and $(14, 24)$ in ResNet50. Additionally adjusting the first convolution's stride in ResNet50 to match the resolution of the VGG16 network decreased both classification and segmentation performance and was thus omitted.

Class activation maps are computed for the abnormal class, upsampled to $(224, 384)$, and normalised to $[0, 1]$. Different methods are explored. Grad-CAM [17] and Grad-CAM++ [2] are evaluated on the final convolutional layer, where the latter improves localisation coverage under multiple object instances by replacing the uniform spatial average with a pixel-wise weighted average. Layer-CAM [11] assigns pixel-wise weights and fuses maps across depths, combining the semantic specificity of deep layers with the spatial detail of shallow ones, and is evaluated at blocks and cumulative fusions of increasing depth. As gradient-based methods can be noisy or vanish in networks operating in the saturated regime of their activation functions [5], Score-CAM [21] and Ablation-CAM [5] are explored as gradient-free alternatives that estimate feature-map importance based on change in class score, evaluated only on the final layer.

### 2.2 FLOWS: Annotation-Free Post-processing

FLOWS (Fused Localisation-map Watershed Segmentation) addresses the limitations of thresholding methods without annotations at any stage. It treats the normalised CAM as a topographic surface, detects activation hotspots at multiple scales as foreground seeds, and grows lesion regions from them by marker-controlled watershed on a joint CAM-PET surface that refines boundaries against competing background seeds.

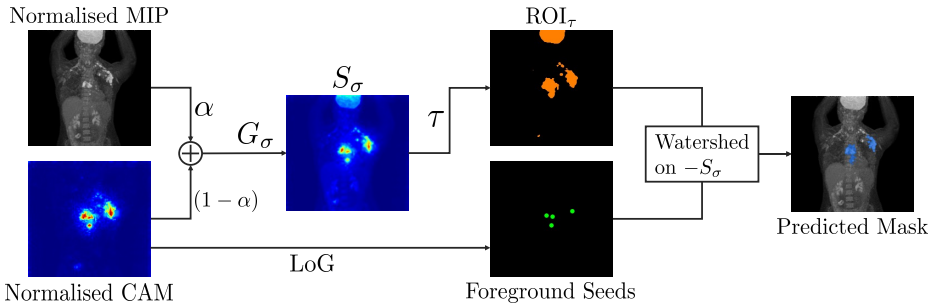


**Fig. 1.** Schematic of FLOWS. The MIP and CAM are combined into surface $S_\sigma$, used to obtain the ROI$_\tau$. Foreground seeds, obtained via the Laplacian of Gaussian (LoG) operation, drive the watershed bounded by ROI$_\tau$.

**1) Normalisation and smoothing.** The heatmap is min–max normalised to $[0, 1]$ and smoothed by a Gaussian of scale $\sigma_s$ to suppress upsampling artefacts.

**2) Multi-scale hotspot detection.** Because lesions activate at different scales and strengths, hotspots are detected with the $\sigma^2$-normalised Laplacian of Gaussian [14],

$$\mathrm{LoG}_\sigma(\mathbf{x}) = \sigma^2 \nabla^2 (G_\sigma * \widetilde{\mathrm{CAM}}_{\sigma_s})(\mathbf{x}), \tag{1}$$

evaluated across scales $\{\sigma_j\}_{j=1}^{J}$. Candidates are local maxima of $|\mathrm{LoG}_{\sigma_j}|$ that are maximal over a $3 \times 3$ neighbourhood and across adjacent scales, each with radius $r_i = \sigma_i^* \sqrt{2}$ at its peak scale $\sigma_i^*$.

**3) Joint surface and region of interest.** The PET projection is clipped to $[0, 20]$ SUV [6], normalised, and combined with the normalised CAM,

$$S(\mathbf{x}) = \alpha \cdot \widetilde{\mathrm{PET}}(\mathbf{x}) + (1-\alpha) \cdot \widetilde{\mathrm{CAM}}(\mathbf{x}), \tag{2}$$

where $\alpha$ weights PET intensity, which sharpens boundaries, against CAM activation, which provides discriminative specificity. The surface is smoothed by a Gaussian of width $\sigma_p$ and thresholded by Otsu's method to define a region of interest $\Omega$ that constrains all subsequent operations.

**4) Seeds and watershed.** Each hotspot contributes a foreground seed, a disk of radius $0.5\, r_i$ at $\mathbf{x}_i$, and a paired background seed at the lowest surface value in the surrounding ring,

$$\mathbf{x}_i^{\mathrm{bg}} =_{\mathbf{x}' \in \mathrm{ring}(\mathbf{x}_i,\, 2r_i)} S_{\sigma_p}(\mathbf{x}'), \tag{3}$$

with $\mathrm{ring}(\mathbf{x}_i, 2r_i) = \{\mathbf{x} \in \Omega \mid r_i \le \|\mathbf{x} - \mathbf{x}_i\| \le 2r_i\}$. Watershed is applied to the inverted surface $-S_{\sigma_p}$ within $\Omega$. The resulting regions are filtered by removing background-labelled regions, pixels below an SUV floor $\tau_{\mathrm{SUV}}$, and components not anchored to a hotspot, and small holes are filled before binarisation.

# 3 Experimental Results and Discussion

## 3.1 Dataset, Preprocessing, and Implementation Details

Experiments use the autoPET FDG-PET/CT dataset [8], comprising 1014 single centre whole-body scans from 900 patients (501 positive: lymphoma, melanoma, and non-small-cell lung cancer; 513 negative). Each scan yields 32 multi-angle MIPs that inherit the patient-level label. Data are split 85/15% (label-stratified) into training and test sets at the patient level, resulting in slightly more lung cancer cases in the test set.

Projections are resampled to $3\times3\times3$ mm and cropped or padded to $224\times384$ pixels (symmetrically in the transverse direction; longitudinally by cropping 15 head-side pixels and the remainder at the legs, or padding at the legs). Five scans are excluded because preprocessing removed all annotated lesions. Inputs are log1p-normalised and augmented with random horizontal/vertical flips, 180° rotation, intensity scaling, and additive Gaussian noise.

Classifiers are trained with AdamW (weight decay 0.01) and dropout 0.2 in the fully connected hidden layers. Learning rate, batch size, and epoch count (selected by convergence and cross-validation) are $8 \times 10^{-7}/64/21$ for VGG16 and $2 \times 10^{-7}/32/39$ for ResNet50. The fully supervised nnU-Net baseline [10] is trained for 1000 epochs on the same split with matched augmentation, using the automatically configured 2D `PlainConvUNet`: six encoder stages, feature maps $[32, 64, 128, 256, 512, 512]$, no dropout, $224 \times 352$ patches, and batch size 41.

FLOWS is implemented in `scikit-image` with CAM smoothing $\sigma_s = 1.5$, LoG detection over 10 scales ($\sigma \in [2, 15]$ pixels), surface smoothing $\sigma_p = 1.0$, $\tau_{\mathrm{SUV}} = 2.5$ [9], and $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$.

## 3.2 Classification

Both classifiers show strong discriminative performance, with AUC of 0.93, recall of 0.80, and precision of 0.91 and 0.92 for VGG16 and ResNet50, respectively. The nnU-Net achieves comparable scan-level detection when non-empty masks are treated as positive predictions, with precision of 0.92 and recall of 0.81. Because classification affects downstream localisation and segmentation, abnormal scans are analysed separately as true positive scans (TPS) and false negative scans (FNS), reflecting a clinical pipeline in which a correct patient-level label would be present.

## 3.3 Segmentation

**CAM Comparison and Otsu Baseline.** CAM localisation is compared using positive-class CAM heatmaps thresholded with Otsu's method across Grad-CAM, Grad-CAM++, Layer-CAM, Score-CAM, and Ablation-CAM. Layer-CAM was evaluated both per layer and as multi-layer fusion [11]. Across all evaluated CAM methods and layer configurations, the best segmentation performance was

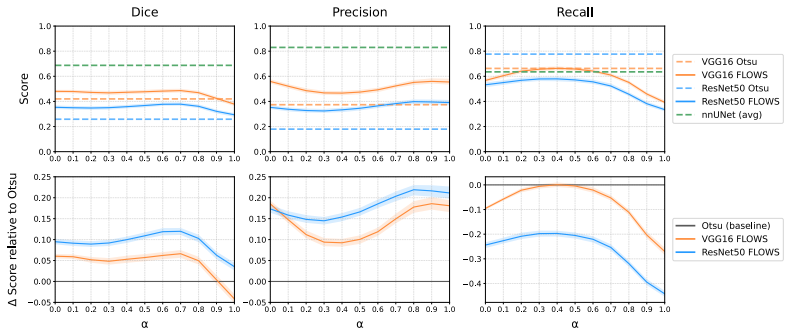


**Fig. 2.** Bootstrap ($n = 500$) of absolute (top) and relative (bottom) metrics over $\alpha$ values for TPS for VGG16 ($\text{LayerCAM}_{123}$, $n = 1860$) and ResNet50 ($\text{LayerCAM}_2$, $n = 1877$) (overlap: $n = 1753$).

obtained with the last three VGG16 layers ($\text{LayerCAM}_{123}$) and the second-to-last ResNet50 layer ($\text{LayerCAM}_2$), likely aided by their higher-resolution feature maps. These configurations are used for all subsequent evaluations.

Otsu thresholding performs similar to grid-optimised global thresholding for these heatmaps, while remaining annotation-free. For example, for VGG16 Layer-$\text{CAM}_{123}$, $\Delta_{\text{Dice}} = \text{Dice}_{\text{Otsu}} - \text{Dice}_{\text{global}}$ had a median of $-0.0141$ (95% bootstrap CI $[-0.0166, -0.0107]$; Wilcoxon signed-rank test on TPS, $p < 0.001$).

**FLOWS.** Figure 2 summarises bootstrap mean metrics across $\alpha$ values for VGG16 and ResNet50, with Otsu baselines and a fully supervised nnU-Net reference. Because the VGG16 and ResNet50 TPS sets have a normalised overlap of 94.25%, a single averaged nnU-Net reference is shown. The bottom row reports mean FLOWS–Otsu differences to highlight relative changes. Table 1 reports VGG16 segmentation results for TPS and FNS. Otsu thresholding is applied directly to the normalised CAM heatmap, while FLOWS is shown for the annotation-free default $\alpha = 0$ and for $\alpha = 0.7$. Dice remains stable for $\alpha \in [0, 0.7]$ (Fig. 2), where PET fusion trades some precision for recall without reducing dice, but decreases at higher $\alpha$ for both models. Since $\alpha$ is not selected using annotations, $\alpha = 0$ is used as the default configuration.

*1) True positive scans.* At $\alpha = 0$, FLOWS increases precision and reduces false positive pixels relative to Otsu, at the cost of recall. This shows that hotspot-based seed filtering and paired background seed watershed reduce over-segmentation without ground-truth annotations or PET fusion. The recall loss likely reflects lesions that do not receive foreground seeds or limited foreground region growing. Non-zero $\alpha$ values incorporate PET intensity into the CAM-PET surface, recovering recall by drawing high-SUV regions with weak CAM activation into the watershed, but may also introduce false positives from high-uptake healthy tissue.

**Table 1.** Segmentation metrics for VGG16 (LayerCAM$_{123}$) for TPS ($n$=1860) and FNS ($n$=476), reported as mean ± std. Dice, precision and recall are reported in percent.

| Scans | Method | Dice | Precision | Recall | FP (↓) | FN (↓) |
|---|---|---|---|---|---|---|
| **TPS** | Otsu | 42.0±20.0 | 37.4±20.9 | **66.3±25.3** | 1059±826 | 579±954 |
| | FLOWS$_{\alpha=0}$ | 48.0±21.6 | **55.9±28.2** | 56.7±26.8 | **364±293** | 695±1110 |
| | FLOWS$_{\alpha=0.7}$ | **48.6±25.8** | 52.3±31.9 | 61.0±26.3 | 688±1154 | **510±739** |
| | nnUNet | 69.4±28.1 | 83.4±26.7 | 64.3±30.3 | 82±139 | 336±456 |
| **FNS** | Otsu | 5.1±6.7 | 2.8±4.0 | **62.0±37.9** | 5108±1546 | **120±206** |
| | FLOWS$_{\alpha=0}$ | **11.3±15.6** | **8.1±13.0** | 40.5±39.3 | **1310±648** | 154±220 |
| | FLOWS$_{\alpha=0.7}$ | 8.7±13.0 | 5.8±10.7 | 42.1±39.8 | 2108±1203 | 148±219 |
| | nnUNet | 13.2±25.6 | 28.1±43.2 | 10.8±23.3 | 9±50 | 221±243 |

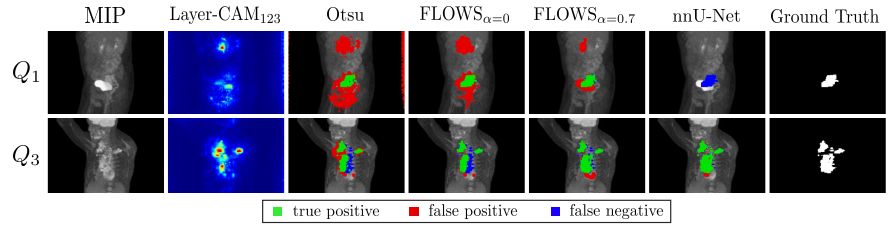


**Fig. 3.** Qualitative results for VGG16 (LayerCAM$_{123}$) showing the 25th ($Q_1$) and 75th ($Q_3$) percentile cases (cropped) by FLOWS$_{\alpha=0}$ Dice.

Qualitative examples in Figure 3 support these trends. Both cases show reduced false positives with FLOWS relative to Otsu, most clearly for FLOWS$_{\alpha=0.7}$ in the $Q_1$ example. In the $Q_3$ example, PET fusion recovers more of the lesion extent, but also introduces a false positive from a nearby high-uptake structure. The same case illustrates a remaining failure mode of CAM-driven watershed segmentation: disconnected lesion regions may be missed when they do not receive detected foreground seeds. Overall, the gap to nnU-Net is concentrated mainly in precision, with false positives arising from coarse CAM localisation and high-uptake structures being captured by nearby seeds at higher $\alpha$.

Performance varies by tumour type (Table 2). Melanoma remains the most challenging tumour type for WSSS, possibly because small scattered metastases produce coarse and diffuse CAM activations. In this group, default FLOWS provides the largest improvement over Otsu, whereas additional PET information yields only marginal further benefit. Lymphoma benefits most from increasing $\alpha$, consistent with discrete, metabolically active lesions. Aggregate TPS results (Table 1) are weighted toward lung cancer, which constitutes 53% of TPS and for which PET fusion reduces segmentation performance.

*2) False negative scans.* Segmentation is poor on scans misclassified as normal (Table 1). While FLOWS improves over Otsu, dice and precision remain low because the heatmaps are diffuse and weakly localised. Otsu thresholding also

**Table 2.** Per-tumour-type segmentation metrics for VGG16 ($\text{LayerCAM}_{123}$) for TPS, reported in percent as mean ± std.

| | Lung cancer ($n = 985$) | Lymphoma ($n = 454$) | Melanoma ($n = 421$) |
|---|---|---|---|
| Otsu | 49.6±16.1 | 40.2±19.7 | 26.1±18.5 |
| $\text{FLOWS}_{\alpha=0}$ | **52.9±18.1** | 47.5±22.8 | 37.1±23.6 |
| $\text{FLOWS}_{\alpha=0.7}$ | 48.6±21.4 | **57.9±28.8** | **38.7±28.1** |
| nnU-Net | 73.3±21.9 | 61.9±36.9 | 68.5±28.5 |

retains a larger foreground fraction in these scans (VGG16 $\text{LayerCAM}_{123}$: Mann–Whitney U = 837,758, $p < 0.001$). Despite poor segmentation, WSSS recall remains substantially higher than nnU-Net, by approximately 30–50 percentage points, because diffuse heatmaps can still overlap true lesions. In contrast, within these FNS, nnU-Net produces empty masks in 65.8% and 64.9% of cases for VGG16 and ResNet50, respectively. These cases therefore reflect intrinsically difficult scans rather than a failure specific to weak supervision.

## 4 Conclusion

We presented a weakly supervised pipeline for FDG-PET tumour lesion segmentation from patient-level labels alone, combining multi-angle PET projections, classifier-based CAM localisation, and FLOWS, an annotation-free post-processing method for converting localisation maps into binary lesion masks. The pipeline remains interpretable at each stage, from scan-level classification and CAM localisation to seed detection, background competition, and watershed-based mask generation.

FLOWS improves over direct heatmap binarisation by replacing a single threshold with a structured map-to-mask conversion. Importantly, this already works in the default CAM-only setting ($\alpha = 0$): on true-positive VGG16 scans, FLOWS increased Dice from 42.0% with Otsu thresholding to 48.0%, while precision rose from 37.4% to 55.9% and mean false-positive pixels decreased from 1059 to 364. PET intensity can be incorporated as a modality-specific cue when beneficial, giving similar overall Dice at $\alpha = 0.7$ (48.6%) while trading precision for recall. These gains are obtained without ground-truth masks for classifier training, threshold selection, or post-processing.

The remaining gap to the fully supervised nnU-Net baseline (69.4% Dice) is mainly driven by coarse CAM localisation, high-uptake healthy tissue, and missed lesions without foreground seeds. Future work should therefore focus on higher-resolution localisation backbones, masking recurrent high-uptake structures, and extending FLOWS from 2D projections to 3D volumes, where the same formulation could be applied without the lossy projection step.

## References


1. Amyar, A., Modzelewski, R., Vera, P., Morard, V., Ruan, S.: Weakly supervised tumor detection in PET using class response for treatment outcome prediction. Journal of Imaging **8**(5), 130 (May 2022). https://doi.org/10.3390/jimaging8050130
2. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Improved visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (Mar 2018). https://doi.org/10.1109/WACV.2018.00097
3. Chen, S., Sedghi Gamechi, Z., Dubost, F., Van Tulder, G., De Bruijne, M.: An end-to-end approach to segmentation in medical images with CNN and posterior-CRF. Medical Image Analysis **76**, 102311 (Feb 2022). https://doi.org/10.1016/j.media.2021.102311
4. Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S.: C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11676–11685 (June 2022)
5. Desai, S., Ramaswamy, H.G.: Ablation-CAM: Visual explanations for deep convolutional networks via gradient-free localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 972–980 (Mar 2020). https://doi.org/10.1109/WACV45572.2020.9093360
6. Diao, Z., Jiang, H.: An uncertainty-based Collaborative Weakly Supervised Segmentation Network for Positron emission tomography-Computed tomography images. Engineering Applications of Artificial Intelligence **148**, 110442 (May 2025). https://doi.org/10.1016/j.engappai.2025.110442
7. Früh, M., Fischer, M., Schilling, A., Gatidis, S., Hepp, T.: Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging. Journal of Medical Imaging **8**(5), 054003 (Oct 2021). https://doi.org/10.1117/1.JMI.8.5.054003
8. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (Oct 2022). https://doi.org/10.1038/s41597-022-01718-3
9. Hellwig, D., Graeter, T.P., Ukena, D., Groeschel, A., Sybrecht, G.W., Schaefers, H.J., Kirsch, C.M.: $^{18}$F-FDG PET for mediastinal staging of lung cancer: Which SUV threshold makes sense? Journal of Nuclear Medicine **48**(11), 1761–1766 (Nov 2007). https://doi.org/10.2967/jnumed.107.044362
10. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z
11. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: LayerCAM: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021). https://doi.org/10.1109/TIP.2021.3089943
12. Kim, J., Kim, H.J., Kim, C., Lee, J.H., Kim, K.W., Park, Y.M., Kim, H.W., Ki, S.Y., Kim, Y.M., Kim, W.H.: Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. Scientific Reports **11**(1), 24382 (Dec 2021). https://doi.org/10.1038/s41598-021-03806-7
13. Krähenbühl, P., Koltun, V.: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc. (2011)

14. Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics **21**(1–2), 225–270 (1994)
15. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (Jan 1979). https://doi.org/10.1109/TSMC.1979.4310076
16. Schmidt-Mengin, M., Benichoux, A., Belachew, S., Komodakis, N., Paragios, N.: ToNNO: Tomographic reconstruction of a neural network's output for weakly supervised segmentation of 3D medical images (2024)
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Feb 2020). https://doi.org/10.1007/s11263-019-01228-7
18. Shiyam Sundar, L.K., Beyer, T.: Is automatic tumor segmentation on whole-body $^{18}$F-FDG PET images a clinical reality? Journal of Nuclear Medicine **65**(7), 995–997 (Jul 2024). https://doi.org/10.2967/jnumed.123.267183
19. Spajić, M., Habijan, M., Marinčić, D., Galić, I.: Evaluating Semantic Segmentation Performance Using DeepLabv3+ with Pretrained ResNet Backbones and Multi-Class Annotations. Engineering Proceedings **125**, 23 (Feb 2026). https://doi.org/10.3390/engproc2026125023
20. Syed, S., Anderssen, K.E., Stormo, S.K., Kranz, M.: Weakly supervised semantic segmentation for MRI: exploring the advantages and disadvantages of class activation maps for biological image segmentation with soft boundaries. Scientific Reports **13**(1), 2574 (Feb 2023). https://doi.org/10.1038/s41598-023-29665-y
21. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 111–119. IEEE (Jun 2020). https://doi.org/10.1109/CVPRW50498.2020.00020
22. Zhang, W., Chen, J., Kanan, C.: INSIGHT: Explainable Weakly-Supervised Medical Image Analysis (Aug 2025). https://doi.org/10.48550/arXiv.2412.02012
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (Jun 2016). https://doi.org/10.1109/CVPR.2016.319