Evaluating Threshold Selection for CAM Refinement in Weakly Supervised Semantic Segmentation

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

Weakly Supervised Semantic Segmentation (WSSS) enables pixel-level segmentation using image-level labels, a less labor-intensive form of supervision than dense annotations. The PSA (Pixel-Semantic Affinity) pipeline is a commonly used multi-stage approach for WSSS, where class activation maps (CAMs) are refined to generate pseudo-labels for training a fully supervised model. A key challenge within this pipeline is accurately determining confident foreground and background regions in CAMs, as these regions impact the quality of pseudo-labels and ultimately segmentation accuracy. To address this bottleneck, we introduce two novel metrics: the Pseudolabel Quality Score (PQS), which measures the impact of confident region thresholds on segmentation accuracy, and the Threshold Evaluation Consistency Metric (TECM), a novel weakly supervised criterion for selecting optimal thresholds without ground truth annotations. Experiments on the PASCAL VOC 2012 dataset demonstrate a strong correlation between these metrics and segmentation accuracy, underscoring the importance of confident region determination in the WSSS pipeline. Our approach improves the performance of the WSSS pipeline by optimizing confident region selection and providing insights for future advancements.

1. Introduction

Semantic segmentation is a fundamental task in computer vision with diverse applications, including autonomous driving, remote sensing, and medical imaging. While fully supervised models have achieved remarkable success in this semantic segmentation [3, 4, 15], their dependence on pixel-level annotations introduces challenges. These annotations are resource-intensive and time-consuming to obtain, particularly for large-scale datasets or domains requiring expert knowledge. To address these challenges, weakly supervised semantic segmentation (WSSS) has emerged as a promising research direction, leveraging less labor-intensive supervisory signals such as image-level labels, bounding boxes, or scribbles. Among these, image-level supervision stands out for its ease of acquisition. However, WSSS pipelines face inherent limitations due to the weak nature of supervision. In this study we focus on WSSS pipelines using image-level labels, identifying a key bottleneck in the pipeline and proposing a novel evaluation criterion to address this limitation.

A typical multi-stage WSSS pipeline consists of two main stages: Stage 1 and Stage 2. In Stage 1, a classifier is trained using image-level labels, and class activation maps (CAMs) [24] are extracted from it. These CAMs provide an initial localization of object regions within an image but are often noisy or incomplete due to the weak nature of supervision. To address this, a refinement network is introduced to refine the CAMs and improve segmentation accuracy. The refined CAMs are then transformed into pseudo-labels by assigning each pixel to the class with the highest activation score. In Stage 2, these pseudo-labels serve as supervision for training a fully supervised semantic segmentation model, further improving segmentation accuracy. Importantly, the quality of the pseudo-labels and therefore the performance of the entire pipeline is heavily dependent on the refinement process conducted by the refinement network. This network is typically trained using only the confident regions of the CAMs. Therefore, accurately identifying these confident regions is a critical step that directly impacts the overall performance of the WSSS pipeline.

We experimentally demonstrate that the selection of confident regions in CAMs significantly impacts the performance of the fully supervised model trained in Stage 2 of the WSSS pipeline. To systematically define these confident regions, we employ two thresholds: a confident background threshold and a confident foreground threshold. A pixel is classified as confident background if its highest CAM score

is less than or equal to the background threshold, and as confident foreground if its score meets or exceeds the foreground threshold. Pixels with scores falling between these thresholds are classified as neutral pixels, representing non-confident regions excluded from refinement network training. Building on these insights, we propose a novel evaluation criterion for determining optimal confident background and foreground thresholds. This criterion improves the refinement of pseudo-labels, thereby improving the overall performance of the WSSS pipeline. Importantly, our approach operates in a weakly supervised setting and does not require additional annotations.

Our main contributions are as follows:

- Identifying the determination of confident regions in CAMs as a critical bottleneck in the WSSS pipeline.
- Proposing a novel evaluation criterion to select the optimal confident background and foreground thresholds, allowing effective confident region determination without additional annotations.

2. Related Work

2.1. Weakly Supervised Semantic Segmentation

Weakly Supervised Semantic Segmentation (WSSS) methods aim to achieve pixel-level segmentation using limited supervision, addressing the high cost and effort associated with dense pixel-level annotations. Various forms of weak supervision have been explored, including image-level labels [11, 16, 17], bounding boxes [5, 10], and scribbles [14, 19]. Among these, methods utilizing image-level labels have gained widespread attention, as they offer a more accessible and cost-effective alternative compared to the more precise but labor-intensive annotations required for bounding boxes or scribbles.

2.1.1 Single-Stage Methods

Single-stage methods [2, 21, 23] address weakly supervised semantic segmentation in a single, end-to-end process, combining classification and segmentation within a unified framework. By eliminating the need for multiple training stages, these methods offer a streamlined structure. However, the joint optimization of classification and segmentation inherently limits the flexibility to further refine segmentation quality. Consequently, despite their simplicity and efficiency, single-stage methods typically achieve lower performance compared to their multi-stage counterparts.

2.1.2 Multi Stage Methods

Multi-stage methods, in contrast to single-stage approaches, typically achieve superior segmentation performance by progressively refining initial segmentation masks. These methods for weakly supervised semantic segmentation often begin with the generation of class activation maps (CAMs) to create initial segmentation masks, which are then used to train a fully supervised segmentation model. However, initial masks derived from CAMs suffer from two primary limitations: they often fail to cover the complete object area and may incorrectly activate non-relevant regions. To address these limitations and improve CAM quality, several methods have been proposed. For instance, Su et al. [18] introduce a novel data augmentation strategy to reduce object-context dependency, Wang et al. [20] leverage a pixel correlation module to refine predictions using neighboring pixel information, Kweon et al. [12] adopt an adversarial approach by coupling a classifier with a reconstruction model, and Ahn et al. [1] propose AffinityNet to capture semantic affinity between adjacent pixels.

Among multi-stage methods, the PSA (Pixel-Semantic Affinity) pipeline, introduced by AffinityNet, is a widely recognized framework in WSSS. The PSA pipeline typically consists of two stages: first, a classifier is trained to generate CAMs, highlighting potential object regions within an image; second, these CAMs are refined using a random walk guided by affinities predicted by Affinity Net, which learns pixel-level semantic affinities from the confident regions of the CAMs. Numerous methods [9, 18] have adopted the PSA pipeline as a base framework, often incorporating modifications to improve the quality of the initial CAMs. In this study, we evaluate the impact of confident region determination within the PSA pipeline, focusing on its implementation with the widely used AffinityNet model. While our analysis centers on AffinityNet, the proposed approach is generalizable to other methods that rely on confident CAM regions for supervision. The PSA pipeline structure is illustrated in Figure 1, showing the stages from initial image classification to final segmentation prediction.

2.2. Confident Region Determination in Affinity Net

Defining confident foreground and background regions in CAMs is critical for effectively training Affinity Net within the PSA pipeline. In their formulation, the authors of Affinity Net introduced a parameter, α, to define the background CAM as $M_{\text{bg}}(x, y) = \left\{1 - \max_{c \in C} M_c(x, y)\right\}^{\alpha}$, where C represents the set of object classes. To determine confident regions, they proposed using lower α values for confident foreground regions and higher α values for confident background areas. While they demonstrated that α is not highly sensitive to overall model performance, they did not evaluate its robustness in accurately identifying confident foreground and background regions, which are essential for training the Affinity Net.

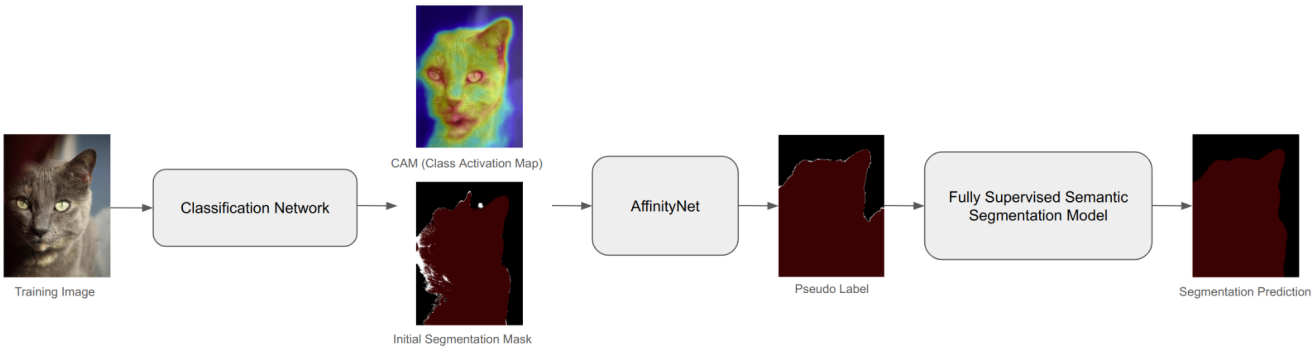


Figure 1. PSA Pipeline

Some studies have modified threshold selection within the PSA pipeline. For example, PuzzleCAM [9] follows the PSA framework but does not provide a specific formulation for generating the background CAM. Instead, they adopt a constant thresholding approach for all pixels, classifying pixels with scores below a background threshold as confident background, those above a foreground threshold as confident foreground, and those in between as neutral. These thresholds are used to define confident regions from which affinity labels are extracted to train the Affinity Net, guiding it in learning pixel affinities. To evaluate the chosen foreground and background thresholds, they calculate the mIoU scores of these masks against ground truth segmentation masks. However, this reliance on ground truth masks for threshold selection introduces challenges in a weakly supervised setting. In contrast, we propose a novel evaluation criterion for selecting the optimal confident background and foreground thresholds, maintaining a fully weakly supervised framework by avoiding reliance on ground truth segmentation masks.

3. Methodology

3.1. Motivation

As highlighted in our related work, multi-stage approaches like the PSA pipeline demonstrate strong potential in WSSS. This study focuses on the PSA pipeline with Affinity Net, which refines initial class activation maps (CAMs) to produce high-quality pseudo-labels for training fully supervised segmentation models. Since Affinity Net relies on confident foreground and background regions in the CAMs for learning, the determination of these regions is critical. The quality of these confident regions directly influences the refinement process, affecting pseudo-label quality and, ultimately, the performance of the fully supervised segmentation model. To address this dependency, we experimentally demonstrate the impact of confident region selection on the WSSS pipeline and propose a method for optimizing confident region determination in CAMs.

3.2. Impact of Confident Region Determination on WSSS Performance

Optimizing the PSA pipeline requires selecting appropriate confident foreground and background thresholds in CAMs. To evaluate the effectiveness of different thresholds in improving segmentation quality, we introduce the Pseudolabel Quality Score (PQS). This metric assesses the quality of pseudo-labels generated from CAMs, serving as a proxy for predicting the segmentation accuracy achievable in Stage 2. The PQS is calculated as:

$$\text{PQS} = \text{mIoU} \times (1 - \text{Neutral Pixel Ratio}) \tag{1}$$

where:

- **mIoU**: The *mean Intersection-over-Union (mIoU)* measures segmentation quality by comparing pseudo-labels with ground truth segmentation masks, excluding neutral pixels. Neutral pixels, whose scores fall between the background and foreground thresholds, represent uncertain regions and are excluded from mIoU computation, as they do not contribute to Affinity Net training. This ensures that only confidently labeled pixels are used to assess segmentation quality. The mIoU is defined as:

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^{C} \frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{2}$$

  where:

  - C is the total number of classes.
  - P_i represents the set of pixels classified as class i in the pseudo-labels.
  - G_i represents the set of pixels classified as class i in the ground truth.

- **Neutral Pixel Ratio**: The *Neutral Pixel Ratio* is the proportion of neutral pixels relative to the total number of pixels in the pseudo-labels. Neutral pixels are

excluded from Affinity Net training to reduce noise but excessive neutral pixels reduce effective training data, potentially limiting the model's ability to learn semantic affinities. Balancing a high mIoU score while minimizing the neutral pixel ratio is critical for robust training. The Neutral Pixel Ratio is calculated as:

$$\text{Neutral Pixel Ratio} = \frac{\text{Number of Neutral Pixels}}{\text{Total Number of Pixels}} \quad (3)$$

In this study, we use PQS as an experimental metric to demonstrate how confident region determination—defined by background and foreground thresholds—affects the performance of the entire WSSS pipeline. PQS provides insights into pipeline dependencies, highlighting the impact of confident region selection on downstream segmentation quality. However, it is not intended for direct threshold optimization, as ground truth masks are typically unavailable in weakly supervised settings. To validate PQS as an indicator of segmentation quality, we analyze its correlation with the mIoU scores of a fully supervised segmentation model trained in Stage 2. Detailed results and analysis of this correlation are presented in the Experiments section.

3.3. Proposed Evaluation Criterion for Confident Region Thresholds

The Pseudolabel Quality Score (PQS) relies on access to ground truth segmentation masks for mIoU calculation, making it impractical for weakly supervised settings where such annotations are unavailable. To overcome this limitation, we propose a novel evaluation criterion tailored for the weakly supervised context. This criterion enables the assessment of confident background and foreground thresholds without requiring ground truth segmentation masks, providing a practical alternative for evaluating threshold effectiveness.

Our Threshold Evaluation Consistency Metric (TECM) leverages the refinement behavior observed in CAMs before and after processing by AffinityNet. By analyzing the changes in the neutral pixel ratio and tracking pixel transitions from neutral to foreground or background during refinement, TECM provides a meaningful indicator of confident region selection quality. This metric aims to optimize the training of AffinityNet, thereby improving the overall performance of the segmentation pipeline.

$$\text{TECM} = \max\left(0, \min\left(2, 1 + \frac{\text{NPR}_{\text{initial}} - \text{NPR}_{\text{refined}}}{\text{NPR}_{\text{initial}} + \epsilon}\right)\right) \times \left(1 - \frac{|\text{N-to-B} - \text{N-to-F}|}{\text{N-to-B} + \text{N-to-F}}\right) \quad (4)$$

where:

- **$\text{NPR}_{\text{initial}}$** represents the Neutral Pixel Ratio in the initial CAMs, defined as the proportion of pixels classified as neutral based on the chosen background and foreground thresholds.
- **$\text{NPR}_{\text{refined}}$** represents the Neutral Pixel Ratio in the refined CAMs generated after processing with Affinity Net.
- **N-to-B** denotes the number of pixels initially classified as neutral in the initial CAMs but reclassified as background in the refined CAMs.
- **N-to-F** denotes the number of pixels initially classified as neutral in the initial CAMs but reclassified as foreground in the refined CAMs.
- ϵ is a small constant added to avoid division by zero when $\text{NPR}_{\text{initial}}$ is close to zero.

This metric incorporates two main factors:

1. **Neutral Pixel Ratio Reduction**: The first component,

$$\max\left(0, \min\left(2, 1 + \frac{\text{NPR}_{\text{initial}} - \text{NPR}_{\text{refined}}}{\text{NPR}_{\text{initial}} + \epsilon}\right)\right)$$

measures the relative reduction in the neutral pixel ratio from the initial to refined CAMs. This term ranges from 0 to 2, penalizing cases where the neutral pixel ratio increases or remains largely unchanged after refinement, while rewarding cases where it decreases, indicating improved confident region determination.

2. **Balanced Transition from Neutral to Foreground/Background**: The second component,

$$1 - \frac{|\text{N-to-B} - \text{N-to-F}|}{\text{N-to-B} + \text{N-to-F}}$$

assesses the balance in transitions from neutral pixels to foreground and background. Balanced transitions indicate that the refined CAMs are consistently reassigning neutral regions, avoiding bias toward either background or foreground. A perfect balance (N-to-B = N-to-F) yields a value of 1, while extreme bias results in lower scores.

TECM combines these two components to provide a robust measure of confident region quality in CAMs. This two-component approach allows the TECM to capture both the extent of neutral pixel reduction and the balance in pixel reassignment. Unlike metrics requiring ground truth masks, TECM evaluates confident region determination directly from the refinement dynamics of CAMs, making it suitable for weakly supervised pipelines.

3.3.1 Limitations of the Proposed Metric

While the metric provides a robust way to evaluate confident region thresholds in the weakly supervised setting, it has limitations. Specifically, when there are no neutral pixels (i.e., if the foreground and background thresholds are

identical), the metric yields a score of 1. This result does not offer meaningful insight into the quality of the threshold pair, as the metric is fundamentally based on the change in neutral pixels and their reassignment. However, in our experiments, we observed that having a proportion of neutral pixels often leads to better performance, as it allows the Affinity Net to focus on high-confidence areas, which improves overall segmentation quality. Therefore, the metric remains valuable and applicable for most threshold configurations.

4. Experiments

4.1. Experimental Setup

In this study, we evaluate our approach using the PASCAL VOC 2012 [6] dataset, which contains 1,464 training images, 1,449 validation images, and 1,456 test images, covering 20 object categories and a background class. We include additional annotations from the Semantic Boundary Dataset [8], resulting in an augmented set of 10,582 images, as used in prior work [1, 13, 20].

We use the PuzzleCAM model for classifier training and CAM extraction, following the experimental setup described in their work. PuzzleCAM employs the PSA pipeline introduced by AffinityNet. Using ResNeSt-101 [22] as the backbone, we train the PuzzleCAM model to generate CAMs and refine them using AffinityNet to create pseudo-labels. PuzzleCAM applies a single-thresholding mechanism to classify pixels in pseudo-labels as either background or foreground, effectively disallowing any neutral pixels. While this approach simplifies the thresholding process, it risks introducing noisy labels into pseudo-labels, especially in regions where CAM scores are uncertain.

In contrast, we employ the same threshold combination used for confident region determination in CAMs to identify confident pixels in pseudo-labels. This ensures consistency across the pipeline and allows uncertain regions to remain neutral, reducing the likelihood of incorporating noisy labels. Non-confident regions in pseudo-labels are excluded from training the fully supervised segmentation model, ensuring that only confident regions guide the learning process. Notably, our experiments show that allowing neutral pixels in pseudo-labels generally results in higher mIoU scores compared to excluding them entirely (see Table 1), further validating our approach. The refined pseudo-labels obtained through this method are then used to train DeepLabv3+ [4] as a fully supervised semantic segmentation model.

4.2. Validation of PQS and TECM Metrics

To evaluate the impact of confident region determination on the WSSS pipeline's performance, we conduct a two-part validation using the Pseudolabel Quality Score (PQS) and the Threshold Evaluation Consistency Metric (TECM). PQS quantitatively assesses the segmentation quality of pseudo-labels by combining segmentation accuracy (mIoU) with data utilization (neutral pixel ratio). TECM, on the other hand, provides a fully weakly supervised criterion to evaluate threshold selection by analyzing refinement behavior in CAMs. Together, these metrics offer complementary perspectives: PQS serves as a direct approximation for segmentation accuracy, while TECM assesses the consistency and effectiveness of the thresholding process in the absence of ground truth annotations.

4.2.1 Validation of PQS

In our first validation, we analyze the relationship between the Pseudolabel Quality Score (PQS) of the pseudo-labels and the mIoU scores of the fully supervised segmentation model on the PASCAL VOC 2012 *val* set. By applying RANSAC [7] to exclude outliers, we observe a Pearson correlation coefficient of **0.97** with a p-value of **3.81e-12**, demonstrating a strong relationship between the PQS of pseudo-labels and the mIoU performance of the fully supervised model across various foreground and background threshold combinations. Table 1 provides a detailed comparison of PQS and mIoU scores for the fully supervised model, while Figure 2 visually illustrates their correlation. These results underscore the critical role of pseudo-label quality, as measured by PQS, in determining the performance of the fully supervised segmentation model.

To further demonstrate the impact of confident region determination on pseudo-label quality and the effectiveness of the WSSS pipeline, we compare the differences in PQS scores between pseudo-labels and initial segmentation masks derived from CAMs with the differences in mIoU scores of the fully supervised models trained on these respective labels. After applying RANSAC to eliminate outliers, we observe a Pearson correlation coefficient of **0.92** with a p-value of **3.23e-9**, indicating a strong relationship between the PQS differences of pseudo-labels and initial masks and the corresponding mIoU score differences of the fully supervised models. Table 2 provides a detailed comparison of these differences, while Figure 3 visually illustrates this correlation. Additionally, the qualitative examples in Figure 5 demonstrate this relationship, showing how an increase in PQS scores from initial masks to pseudo-labels corresponds to an increase in mIoU scores in the segmentation predictions and vice versa. These findings highlight that confident region determination in CAMs has a progressive effect on the segmentation performance of the final model. This underscores the critical role of selecting optimal confident regions in initial CAMs to improve the overall performance of the WSSS pipeline.

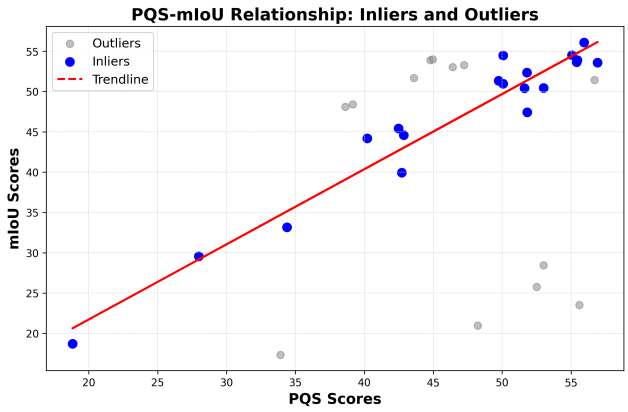


Figure 2. PQS-mIoU Relationship: Inliers and Outliers.

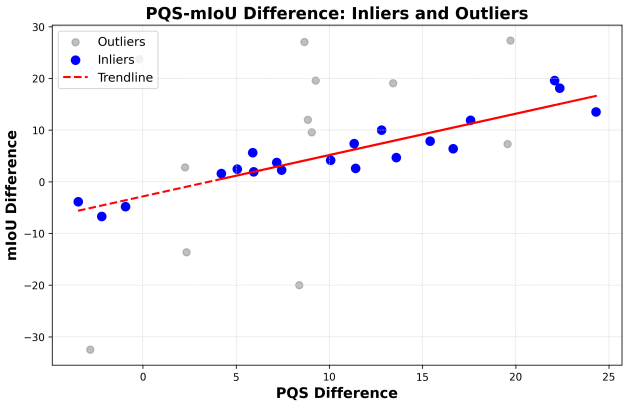


Figure 3. PQS-mIoU Difference: Inliers and Outliers.

BG Threshold	FG Threshold	Neutral Pixel Ratio	mIoU	PQS
0.05	0.05	0.00	39.92	42.71
	0.15	24.42	**44.58**	**42.85**
	0.3	48.01	17.33	33.92
	0.45	73.18	18.72	18.82
0.1	0.1	0.00	**50.40**	51.62
	0.2	12.18	25.78	**52.51**
	0.35	26.18	20.96	48.23
	0.5	41.21	44.19	40.21
0.15	0.15	0.00	23.51	55.60
	0.25	8.51	53.88	**55.46**
	0.4	19.14	47.43	51.82
	0.55	30.75	**53.97**	44.96
0.2	0.2	0.00	53.57	**56.92**
	0.3	6.80	**56.06**	55.94
	0.45	15.74	52.33	51.79
	0.6	25.59	53.88	44.78
0.25	0.25	0.00	51.41	**56.69**
	0.35	5.66	**54.51**	55.07
	0.5	13.51	54.46	50.07
	0.65	22.27	45.41	42.48
0.3	0.3	0.00	**53.64**	**55.40**
	0.4	5.00	50.43	53.00
	0.55	11.82	53.28	47.23
	0.7	19.88	48.08	38.63
0.35	0.35	0.00	28.45	**53.01**
	0.45	4.46	50.96	50.06
	0.6	10.64	**51.66**	43.59
	0.75	18.25	33.17	34.38
0.4	0.4	0.00	51.33	**49.74**
	0.5	3.97	**53.03**	46.41
	0.65	9.74	48.41	39.16
	0.8	17.33	29.56	27.99

Table 1. Comparison of mIoU and PQS scores across different background (BG) and foreground (FG) threshold configurations. The **Neutral Pixel Ratio** represents the proportion of neutral pixels in the pseudo-labels relative to the total pixels. The **mIoU** values are calculated on the PASCAL VOC 2012 *val* set from the fully supervised segmentation models trained with pseudo-labels generated using the respective threshold combinations. The **PQS (Pseudolabel Quality Score)** evaluates pseudo-label quality as defined in Equation 1.

4.2.2 Validation of TECM

Building on our validation with PQS, we assess the effectiveness of TECM in evaluating threshold pairs. The TECM metric provides a weakly supervised approach for threshold selection, eliminating the reliance on ground truth labels by focusing on neutral pixel changes and the balance of pixel reassignment.

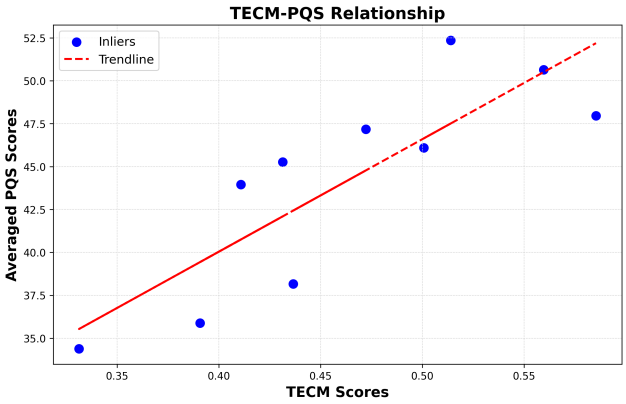


Figure 4. TECM-PQS Relationship.

To validate TECM, we calculate its correlation with the averaged PQS scores of initial masks and pseudo-labels, which serve as an approximation for segmentation accuracy. The correlation analysis excludes cases where no neutral pixels exist, i.e., when the foreground threshold equals the background threshold, as such configurations do not provide meaningful neutral region dynamics for TECM evaluation. A high correlation between TECM and PQS demonstrates that TECM is a reliable metric for identifying threshold values that enhance segmentation quality, even without access to ground truth annotations. After applying RANSAC to eliminate outliers, we observe a Pearson correlation coefficient of **0.84** with a p-value of **0.0024** across different background and foreground threshold combinations. Table 3 presents these results, while Figure 4 provides a visual representation of the correlation, excluding the identified outliers for clarity. Additionally, Figure 6 offers qualitative examples indicating the relationship between TECM and mIoU, further validating the metric's effectiveness.

In conclusion, PQS validates that confident region determination impacts the WSSS pipeline's segmentation accu-

racy, while TECM offers a weakly supervised criterion for optimal threshold selection without requiring ground truth. Together, PQS and TECM enable robust and effective confident region determination, improving the pipeline's performance.

5. Conclusion

In this study, we addressed the challenge of confident region determination in CAMs to improve the PSA pipeline for Weakly Supervised Semantic Segmentation (WSSS). We introduced two key metrics: the Pseudolabel Quality Score (PQS), which demonstrates the impact of confident region thresholds on final segmentation accuracy, and the Threshold Evaluation Consistency Metric (TECM), which provides a reliable and fully weakly supervised criterion for threshold selection, eliminating the need for ground truth annotations. Our results show that optimizing confident region thresholds, guided by TECM, leads to improved pseudo-label quality and improved segmentation performance. We hope this work will provide insights for future research following the PSA pipeline in determining optimal confident regions in CAMs.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 2, 5

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4253–4262, 2020. 2

[3] Liang-Chieh Chen. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 5

[5] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 2

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[8] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 5

[9] Sanghyun Jo and In-Jae Yu. Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE international conference on image processing (ICIP)*, pages 639–643. IEEE, 2021. 2, 3

[10] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2

[11] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 695–711. Springer, 2016. 2

[12] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 2

[13] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5267–5276, 2019. 5

[14] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 2

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[16] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015. 2

[17] 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. 2

[18] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021. 2

[19] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 2

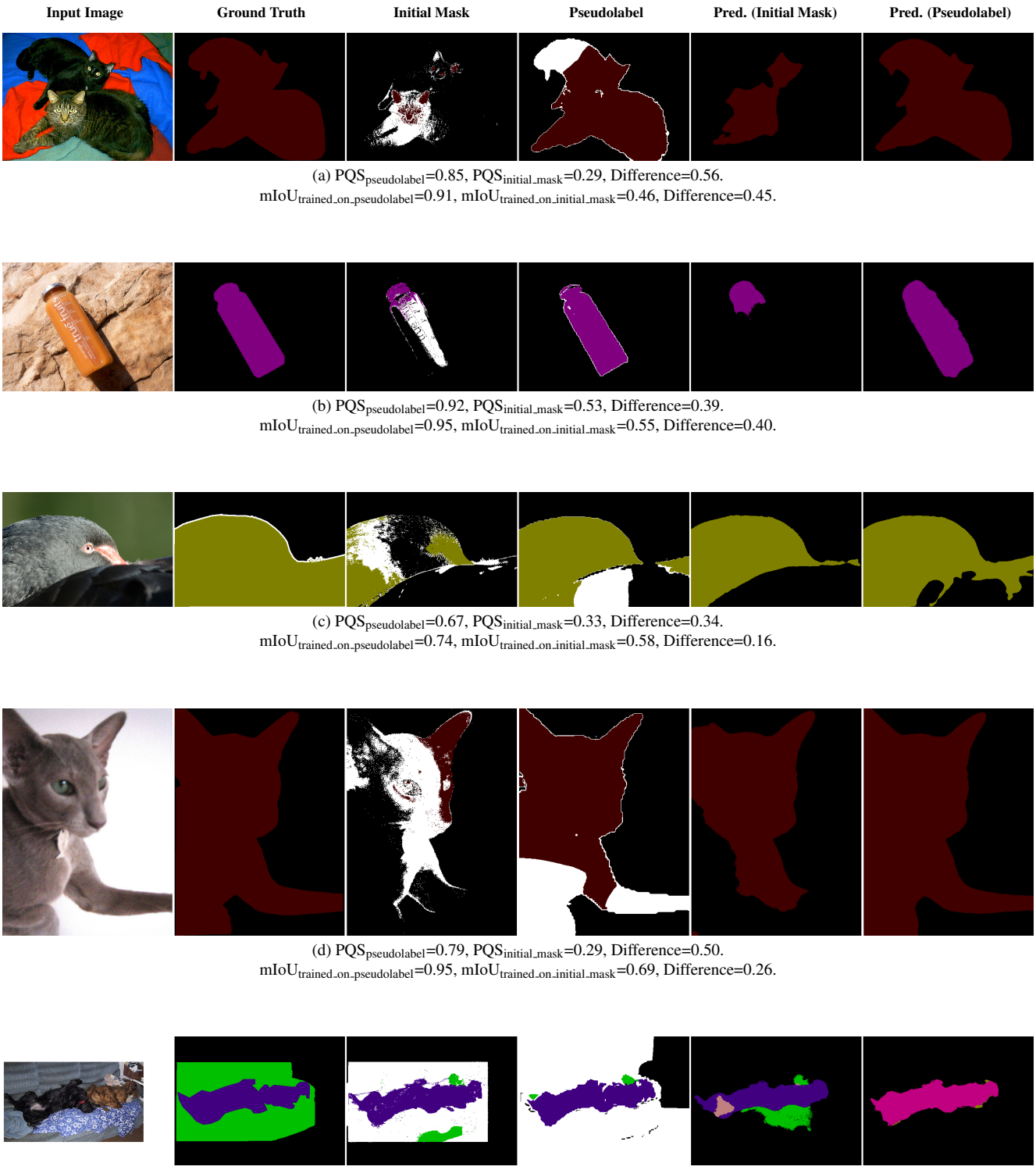
[20] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12275–12284, 2020. 2, 5

[21] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772, 2020. 2

[22] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 5

[23] Xiangrong Zhang, Zelin Peng, Peng Zhu, Tianyang Zhang, Chen Li, Huiyu Zhou, and Licheng Jiao. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5463–5472, 2021. 2

[24] 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. 1



(a) $\text{PQS}_{\text{pseudolabel}}$=0.85, $\text{PQS}_{\text{initial\_mask}}$=0.29, Difference=0.56.
$\text{mIoU}_{\text{trained\_on\_pseudolabel}}$=0.91, $\text{mIoU}_{\text{trained\_on\_initial\_mask}}$=0.46, Difference=0.45.

(b) $\text{PQS}_{\text{pseudolabel}}$=0.92, $\text{PQS}_{\text{initial\_mask}}$=0.53, Difference=0.39.
$\text{mIoU}_{\text{trained\_on\_pseudolabel}}$=0.95, $\text{mIoU}_{\text{trained\_on\_initial\_mask}}$=0.55, Difference=0.40.

(c) $\text{PQS}_{\text{pseudolabel}}$=0.67, $\text{PQS}_{\text{initial\_mask}}$=0.33, Difference=0.34.
$\text{mIoU}_{\text{trained\_on\_pseudolabel}}$=0.74, $\text{mIoU}_{\text{trained\_on\_initial\_mask}}$=0.58, Difference=0.16.

(d) $\text{PQS}_{\text{pseudolabel}}$=0.79, $\text{PQS}_{\text{initial\_mask}}$=0.29, Difference=0.50.
$\text{mIoU}_{\text{trained\_on\_pseudolabel}}$=0.95, $\text{mIoU}_{\text{trained\_on\_initial\_mask}}$=0.69, Difference=0.26.

(e) $\text{PQS}_{\text{pseudolabel}}$=0.19, $\text{PQS}_{\text{initial\_mask}}$=0.57, Difference=-0.38.
$\text{mIoU}_{\text{trained\_on\_pseudolabel}}$=0.12, $\text{mIoU}_{\text{trained\_on\_initial\_mask}}$=0.37, Difference=-0.25.

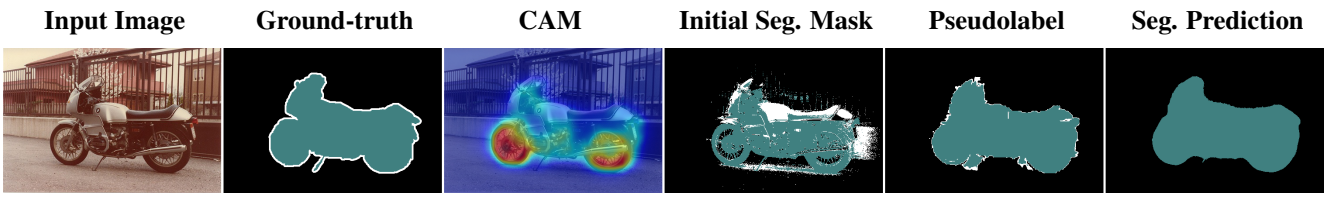
Figure 5. Qualitative examples from the PASCAL VOC 2012 *val* set illustrating PQS and mIoU scores for initial segmentation masks and pseudo-labels. The figure demonstrates improvements in segmentation quality with pseudo-labels, highlighting the differences in PQS and mIoU scores. Specifically, the **PQS Difference** is calculated as $PQS_{\text{pseudolabel}} - PQS_{\text{initial\_mask}}$, and the **mIoU Difference** is computed as $mIoU_{\text{trained\_on\_pseudolabel}} - mIoU_{\text{trained\_on\_initial\_mask}}$. **Pred. (Initial Mask)** represents the prediction made by the fully supervised model trained on initial segmentation masks, while **Pred. (Pseudolabel)** represents the prediction made by the fully supervised model trained on pseudo-labels. Positive differences indicate improved segmentation accuracy, while negative differences highlight performance degradation.

BG Threshold	FG Threshold	PQS Pseudo	PQS Initial	mIoU Pseudo	mIoU Initial	PQS Diff	mIoU Diff
0.05	0.05	42.71	46.19	39.92	43.79	-3.48	-3.87
	0.15	42.85	45.06	44.58	51.31	-2.21	-6.73
	0.3	33.92	34.85	17.33	22.13	-0.93	-4.80
	0.45	18.82	21.65	18.72	51.21	-2.83	-32.49
0.1	0.1	51.62	51.82	50.40	26.66	-0.20	23.74
	0.2	52.51	50.25	25.78	22.97	2.26	2.81
	0.35	48.23	42.29	20.96	19.04	5.94	1.92
	0.5	40.21	31.55	44.19	17.17	**8.66**	**27.02**
0.15	0.15	55.60	53.26	23.51	37.15	2.34	-13.64
	0.25	55.46	50.40	53.88	51.42	5.06	2.46
	0.4	51.82	42.55	47.43	27.82	9.27	**19.61**
	0.55	44.96	31.38	53.97	49.27	**13.58**	4.70
0.2	0.2	56.92	52.72	53.57	51.96	4.20	1.61
	0.3	55.94	48.76	56.06	52.34	7.18	3.72
	0.45	51.79	40.39	52.33	49.73	11.40	2.60
	0.6	44.78	28.13	53.88	47.47	**16.65**	**6.41**
0.25	0.25	56.69	50.80	51.41	45.76	5.89	5.65
	0.35	55.07	46.22	54.51	42.51	8.85	12.00
	0.5	50.07	36.64	54.46	35.40	13.43	**19.06**
	0.65	42.48	22.92	45.41	38.09	**19.56**	7.32
0.3	0.3	55.40	47.96	53.64	51.38	7.44	2.26
	0.4	53.00	42.93	50.43	46.28	10.07	4.15
	0.55	47.23	31.83	53.28	45.41	15.40	7.87
	0.7	38.63	16.55	48.08	28.50	**22.08**	**19.58**
0.35	0.35	53.01	44.62	28.45	48.45	8.39	-20.00
	0.45	50.06	38.73	50.96	43.55	11.33	7.41
	0.6	43.59	26.02	51.66	39.78	17.57	11.88
	0.75	34.38	10.08	33.17	19.64	**24.30**	**13.53**
0.4	0.4	49.74	40.69	51.33	41.75	9.05	9.58
	0.5	46.41	33.61	53.03	43.03	12.80	10.00
	0.65	39.16	19.45	48.41	21.09	19.71	**27.32**
	0.8	27.99	5.63	29.56	11.46	**22.36**	18.10

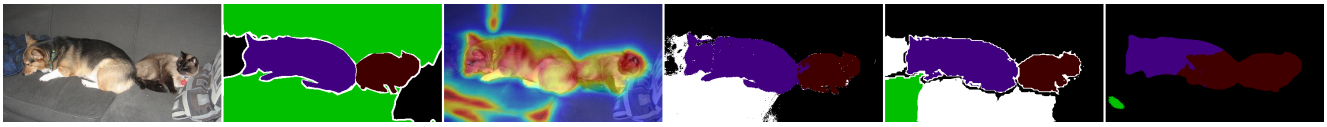
Table 2. Detailed comparison of PQS and mIoU scores across different threshold configurations for confident background (BG) and foreground (FG). The **BG Threshold** and **FG Threshold** columns represent the thresholds used to define confident background and foreground regions, respectively. **PQS Pseudo** and **PQS Initial** are the Pseudolabel Quality Scores for pseudo-labels and initial masks derived from CAMs, respectively. **mIoU Pseudo** and **mIoU Initial** are the mIoU scores calculated on the PASCAL VOC 2012 *val* set for fully supervised models trained on pseudo-labels and initial masks, respectively. **PQS Diff** and **mIoU Diff** represent the differences (**PQS Pseudo** - **PQS Initial**) and (**mIoU Pseudo** - **mIoU Initial**).

BG Threshold	FG Threshold	PQS Initial	PQS Pseudo	PQS Average	NPR Initial (%)	NPR Pseudo (%)	$\frac{\text{N-to-B}-\text{N-to-F}}{\text{N-to-B}+\text{N-to-F}}$	TECM (x100)
0.05	0.15	45.06	42.85	**43.955**	23.44	24.42	0.57	**40.81**
	0.3	34.85	33.92	34.385	45.86	48.01	0.66	32.69
	0.45	21.65	18.82	20.235	66.19	73.18	0.70	27.36
0.1	0.2	50.25	52.51	**51.38**	13.46	12.18	0.55	**49.83**
	0.35	42.29	48.23	45.26	28.75	26.18	0.60	44.07
	0.5	31.55	40.21	35.88	43.47	41.21	0.62	39.71
0.15	0.25	50.40	55.46	**52.93**	9.93	8.51	0.56	**51.32**
	0.4	42.55	51.82	47.185	22.29	19.14	0.58	48.44
	0.55	31.38	44.96	38.17	35.02	30.75	0.60	45.06
0.2	0.3	48.76	55.94	**52.35**	8.06	6.80	0.56	**51.92**
	0.45	40.39	51.79	46.09	18.85	15.74	0.57	51.37
	0.6	28.13	44.78	36.455	30.43	25.59	0.59	48.59
0.25	0.35	46.22	55.07	**50.645**	6.89	5.66	0.53	**56.68**
	0.5	36.64	50.07	43.355	16.85	13.51	0.55	55.57
	0.65	22.92	42.48	32.7	27.24	22.27	0.56	53.33
0.3	0.4	42.93	53.00	**47.965**	6.12	5.00	0.52	**59.09**
	0.55	31.83	47.23	39.53	15.59	11.82	0.55	58.92
	0.7	16.55	38.63	27.59	24.55	19.88	0.53	58.42
0.35	0.45	38.73	50.06	**44.395**	5.70	4.46	0.52	61.76
	0.6	26.02	43.59	34.805	14.76	10.64	0.55	**61.94**
	0.75	10.08	34.38	22.23	21.94	18.25	0.49	60.94
0.4	0.5	33.61	46.41	**40.01**	5.51	3.97	0.53	65.60
	0.65	19.45	39.16	29.305	13.91	9.74	0.54	65.27
	0.8	5.63	27.99	16.81	19.04	17.33	0.38	**68.25**

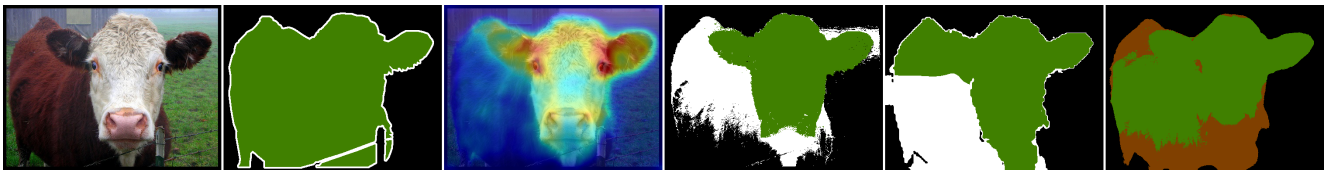
Table 3. Detailed analysis of PQS and TECM scores across different background (BG) and foreground (FG) threshold configurations. The **BG Threshold** and **FG Threshold** columns represent the thresholds used to define confident background and foreground regions, respectively. **PQS Initial** refers to the Pseudolabel Quality Score of the initial segmentation masks derived from CAMs, while **PQS Pseudo** refers to the PQS scores of the pseudo-labels. **PQS Average** is the average of PQS Initial and PQS Pseudo. **NPR Initial** and **NPR Pseudo** represent the neutral pixel ratio (%) in the initial segmentation masks and pseudo-labels, respectively. **N-to-B** denotes the number of pixels reclassified as background in pseudo-labels while neutral in initial masks, and **N-to-F** denotes the number of pixels reclassified as foreground in pseudo-labels while neutral in initial masks. **TECM** refers to the Threshold Evaluation Consistency Metric score (scaled by 100), as defined in Equation 4.



(a) $TECM = 1.39$, $mIoU = 0.98$

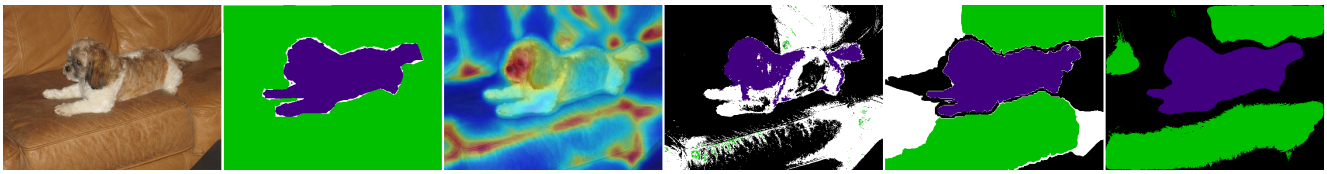


(b) $TECM = 0.28$, $mIoU = 0.32$



(c) $TECM = 0.36$, $mIoU = 0.56$

(d) $TECM = 0.60$, $mIoU = 0.68$



(e) $TECM = 0.29$, $mIoU = 0.47$

Figure 6. Qualitative examples from PASCAL VOC 2012 *val* set with corresponding $mIoU$ and $TECM$ scores.