Mitigating Spurious Correlations in Weakly Supervised Semantic Segmentation via Cross-architecture Consistency Regularization

Industrial Exhaust Smoke emission-Oriented Pseudo label Refinement Method

Presented by:
Zheyuan Zhang
Supervised by:
Professor Yen-Chia Hsu

July 20, 2025

Universiteit van Amsterdam
VU Vrije Universiteit Amsterdam

Outline

Background

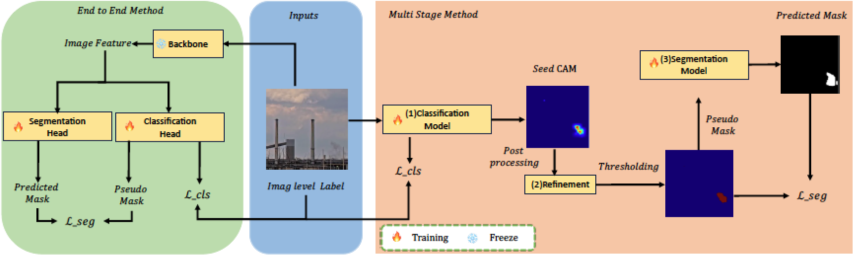
Goal

- Task: Industrial exhaust smoke segmentation.
- Challenge: Scarcity of pixel-level annotations.
- Approach: Multi-stage weakly supervised semantic segmentation based on image-level labels.

Universiteit van Amsterdam VU Vrije Universiteit Amsterdam

The pipeline of weakly supervised semantic segmentation

1. Train a Classifier using image-level labels.
2. Using class activation map to generate pseudo labels.
3. Train a segmentation model using pseudo labels.



Challenges

Observation: The classifier achieves very high accuracy, but the CAM is inaccurate or even fails to localize the foreground.

How to Address These Issues?

Post-processing

- Applied **after** CAM generation to improve pseudo mask quality.
- Encourages spatial consistency
  Limitations:
  - May amplify existing errors.
  - Effectiveness is bounded by the initial CAM quality.

Optimizing CAM Generation

- Improve the quality of Class Activation Maps **at the source**.
- Leads to better semantic localization and more accurate masks.

Previous Work: Addressing Spurious Correlations

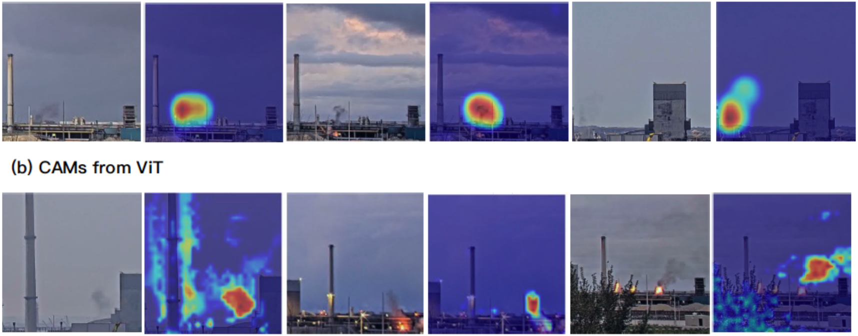
- **Data Augmentation**: Breaking object co-occurrence
  - **Image decomposition**: Separate foreground and background.
  - **Supplemental images**: Introduce diverse contexts.
- **Human Priors**:
  - **Human-in-the-loop**: Human feedback.
  - **Causality chain modeling**: Incorporate causal reasoning into training.
- **External Supervision / Additional Knowledge**:
  - **Saliency map**: Use saliency maps as guidance for pseudo label refinement.
  - **CLIP**: Leverage natural language supervision.

Key Observation

Biased knowledge extracted from both sides.

(a) CAMs from ResNet

(b) CAMs from ViT



Motivation

Intuition: CNNs and ViTs offer complementary strengths.

- **CNNs** leverage local convolutions and strong inductive biases, making them effective at precisely localizing foreground objects.
- **ViTs** utilize global self-attention mechanisms, enabling them to capture rich semantic context.

Table: Key architectural differences between ResNet and ViT

Aspect	**ResNet (CNN)**	**ViT (Transformer)**
Receptive Field	**Local**	**Global**
Inductive Bias	**Strong spatial priors**: • Locality • Spatial invariance	**Weak spatial priors**: • Learn from data • Global context modeling
CAM	**Precise localization**	**Semantic rich but diffused**

Outline

Research question

Question 1: Based on the fact that the classifier achieves very high accuracy, but the CAM is inaccurate or even fails to localize the foreground, how to maintain high accuracy at the same time generate high-quality pseudo labels?

Research question

Question 1: Based on the fact that the classifier achieves very high accuracy, but the CAM is inaccurate or even fails to localize the foreground, how to maintain high accuracy at the same time generate high-quality pseudo labels?

Question 2: Is it possible to address co-occurrence issue without external supervision or additional knowledge?

Research question

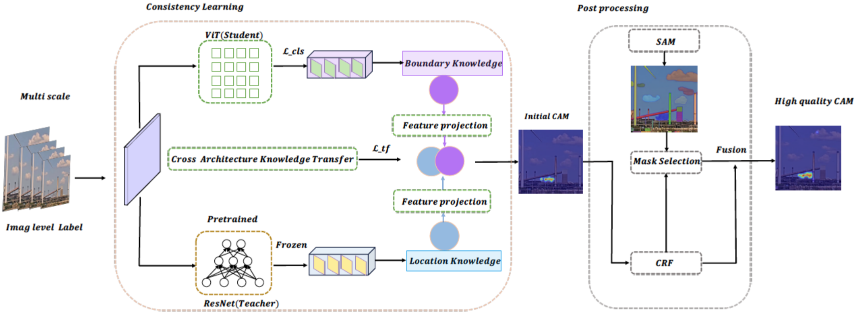
Question 1: Based on the fact that the classifier achieves very high accuracy, but the CAM is inaccurate or even fails to localize the foreground, how to maintain high accuracy at the same time generate high-quality pseudo labels?

Question 2: Is it possible to address co-occurrence issue without external supervision or additional knowledge?

Question 3: Can we collaboratively aggregate heterogeneous features from CNN based and Transformer based models to address co-occurrence issue?

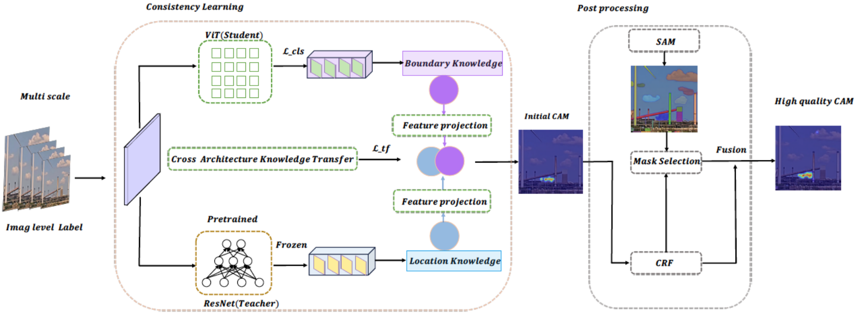
Outline

Framework



Consistency Learning
Post processing
ViT(Student)
£_cls
Boundary Knowledge
Multi scale
Feature projection
Initial CAM
SAM
High quality CAM
Cross Architecture Knowledge Transfer
£_tf
Mask Selection
Fusion
Feature projection
Imag level Label
Pretrained
Frozen
Location Knowledge
CRF
ResNet(Teacher)

Framework

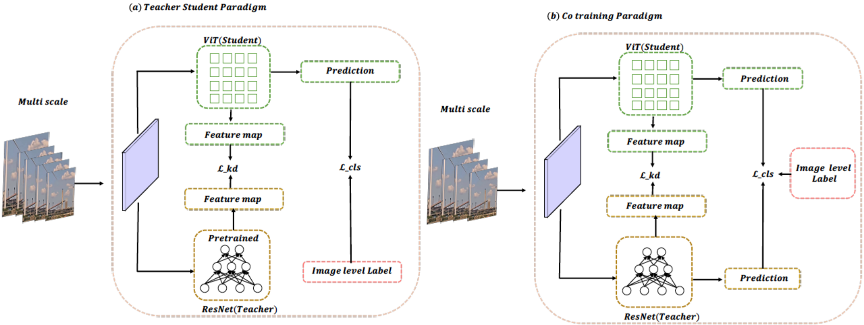


$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{tf}}$$

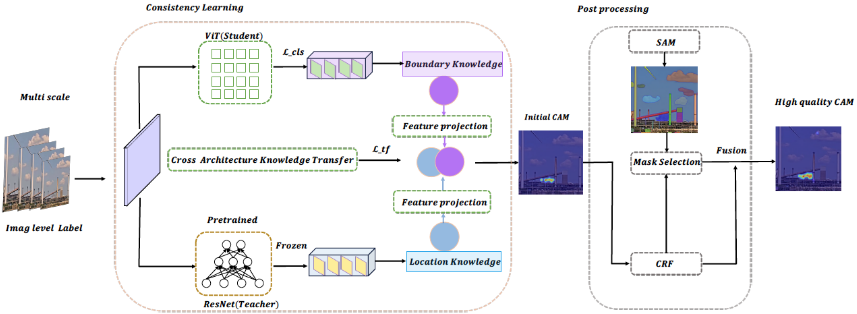
Cross-Architecture Feature

Challenge: Transferring knowledge between fundamentally different architectures—such as Transformers and CNNs—is more difficult. Their distinct design principles lead to divergent feature representations, making compact and effective knowledge transfer non-trivial.

Knowledge transfer training scheme



Framework


Consistency Learning
Post processing
ViT(Student)
£_cls
Boundary Knowledge
SAM
Multi scale
Feature projection
Initial CAM
High quality CAM
Cross Architecture Knowledge Transfer
£_tf
Mask Selection
Fusion
Feature projection
Imag level Label
Pretrained
Frozen
Location Knowledge
CRF
ResNet(Teacher)

Which part provides a more informative knowledge source?

The knowledge transfer performance is sensitive to how the knowledge is defined.

Logit-Based

- Uses the teacher's softmax predictions as pseudo labels for the student.
- Not suitable for our task, as it loses the spatial information and ignores how the internal representations are formed.

Feature-Based

- Minimizes the difference between the intermediate feature representations of the student and the teacher.
- Preserves semantic and spatial information.

How to align the mismatched representations

Spatial Map: Aggregates channel information into a 2D spatial map.

$$F : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{H\times W}$$

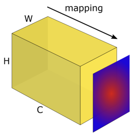


Figure: Spatial map: loses channel idms semantic information.

Inner Product: Computes pairwise channel relations to preserve semantic structure.

$$F : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{C\times C}$$

Cross-Architecture Feature Alignment Strategies

Table: Comparison of Feature Shapes and Their Properties

Shape	Keeps Channel Info?	Keeps Spatial Info?	Semantically Rich?
$[B, C, H \times W]$	✓ Yes	✓ Yes	✓ Yes
$[B, H \times W]$	No	✓ Yes	No
$[B, C, C]$	✓ Yes	No	No

[B, C, H×W] :flatten spatial layout while keeping semantic channels. Then use a learnable feature projection layer to align the feature.

Cross-Architecture Feature Alignment Strategies

Comparison Strategies:

- **Global Alignment:** Enforces consistency in holistic feature representations.
- **Channel-Wise Alignment:** Aligns feature responses along the channel dimension, helping match semantic filters between models.
- **Spatial Alignment:** precise pixel-to-pixel correspondence.

Various Post-processing Techniques

Problem: The initially generated CAMs are often redundant and incomplete.

- **Multi-scale Inference:** Aggregates CAMs from multiple input resolutions to improve robustness and capture multi-level semantics.
- **CRF (Conditional Random Field):** Models pixel-level relationships to enforce spatial consistency and sharpen object boundaries.
- **AffinityNet:** Learns pairwise pixel affinities and propagates CAMs to refine segmentation masks.
- **CAM Fusion:** Combines CAMs from different layers to increase coverage and completeness.

Emerging Trends

- **SAM-Enhanced** Leverage SAM for zero-shot pseudo masks generation, enhancing spatial consistency and boundary quality.
- **CLIP-Aided** Incorporate vision-language priors by using CLIP's text encoder to generate class-specific weights for CAM generation. The effectiveness relies heavily on well-crafted textual prompts, especially for abstract concepts like smoke.

Outline

Dataset description

Dataset

Source	Examples	Supervision Type	Class
IJmond	train	image-level	Smoke&non-smoke
IJmond	test	pixel-level	Smoke
Smoke5K	train	image-level	Smoke
RISE	train	image-level	Smoke&non-smoke

Comparison with baseline models

Table: Evaluate mIOU of pseudo masks with different backbones. T_1:Train dataset. T_2:Part of the test dataset. Gray rows indicate ours method.

Supervision	**Source**	**Backbone**	**mIOU**
image-level	T_1	ResNet50	26.10
image-level	T_1	ResNet101	21.29
image-level	T_1	ViT-S	13.18
image-level	T_1	Ours	47.37
image-level	$T_1 + T_2$	ViT-B	47.99
image-level+limited pixel-level	$T_1 + T_2$	ViT-B	✗

Fully supervised model vs Weakly supervised model

Table: Comparison of semantic segmentation methods. Fully supervised learning methods are trained with ground truth labels without any post-processing.

Method	**Backbone**	**mIOU**
Fully Supervised		
SERT[?]	Transformer	68.27
SAM-fine-tuning	ViT-B	54.68
Multi stage WSSS		
Ours+post-processing	ViT-S+ResNet50	52.93

Comparison with previous methods

Table: Comparison with previous methods.

Method	**backbone**	**mIOU**
TransCAM	Conformer	15.02
AffinityNet	ResNet50	24.28
PCM	ResNet50	33.56
Ours	ViT-S+ResNet50	47.37
Ours+post-processing	ViT-S+ResNet50	52.93

Post-processing

Table: Evaluation of pseudo labels with different post-processing techniques.

Method	mIoU
w/o post-processsing	37.42
+Multi scale	38.49
+AffinityNet[?]	34.00
+SAM-enhanced[?]	43.20
+CLIP[?]	✗
+CRF	43.27
+CAM fusion	46.91
+CRF+CAM fusion	37.81
+CRF+AffinityNet[?]	38.51
Optimal threshold	53.92

(a) CAMs generated by ours (Worse seed)

Method	mIoU
w/o post-processsing	46.25
+Multi scale	47.37
+CAM fusion	45.27
+CRF+AffinityNet[?]	49.16
+SAM-enhanced[?]	51.00
+CLIP[?]	✗
+CRF	52.52
+CRF+SAM-enhanced	52.93
Optimal threshold	57.15

(b) CAMs generated by ours (Best seed)

Ablation Studies

Table: Comparison of different knowledge transfer strategies

Paradigm	Teacher	Student	Metric	Level	mIOU
Teacher-Student	ResNet(Pre-trained)	ViT	Cosine	Global	47.37
Teacher-Student	ResNet(Pre-trained)	ViT	L_1	Global	38.39
Teacher-Student	ResNet(Pre-trained)	ViT	L_2	Global	33.74
Teacher-Student	ResNet(Pre-trained)	ViT	Cosine	Spatial	43.70
Teacher-Student	ResNet(Pre-trained)	ViT	Cosine	Channel	46.85
Co-training	ViT + ResNet(From scratch)	ViT+ResNet	Cosine	Global	45.93
Co-training	ViT + ResNet(From scratch)	ViT+ResNet	L_1	Global	18.51
Co-training	ViT + ResNet(From scratch)	ViT+ResNet	L_2	Global	0.27
Co-training	ViT + ResNet(From scratch)	ViT+ResNet	Cosine	Spatial	45.11
Co-training	ViT + ResNet(From scratch)	ViT+ResNet	Cosine	Channel	42.75

Table: The impact of the knowledge transfer Loss coefficient

λ	**mIOU**
0.3	38.55 $\uparrow$ 4.99
0.5	43.81 $\uparrow$ 10.25
0.8	45.04 $\uparrow$ 11.48
1.0	47.37 $\uparrow$ 13.81
1.3	46.62 $\uparrow$ 13.06
1.5	41.90 $\uparrow$ 8.34

Outline

Conclusion

- We propose a simple yet effective **knowledge transfer** method based on **cross-architecture consistency**, aimed at mitigating spurious correlations without relying on human priors or external supervision.
- Our approach successfully transfers **complementary knowledge** from the teacher model while preserving the strengths of the student model, thereby **reducing classifier bias**.
- In addition, we explore and evaluate various post-processing techniques to further enhance the quality of pseudo labels.

Limitations and Future Work

- Explore additional feature alignment strategies for better cross-architecture knowledge transfer.
- Investigate more effective integration of post-processing techniques.
- Extend the approach from binary to multi-class classification.
- Evaluate the generalization of the method on other datasets.
- Validate the effectiveness in a end-to-end WSSS setting.

End

Thanks for listening!