Exploring Explaining Methods in Multi-Label Problems and Complementary Regularization Strategies in Weakly Supervised Semantic Segmentation

Candidate: Lucas Oliveira David
Advisor: Prof. Dr. Zanoni Dias
Co-advisor: Prof. Dr. Hélio Pedrini

Schedule

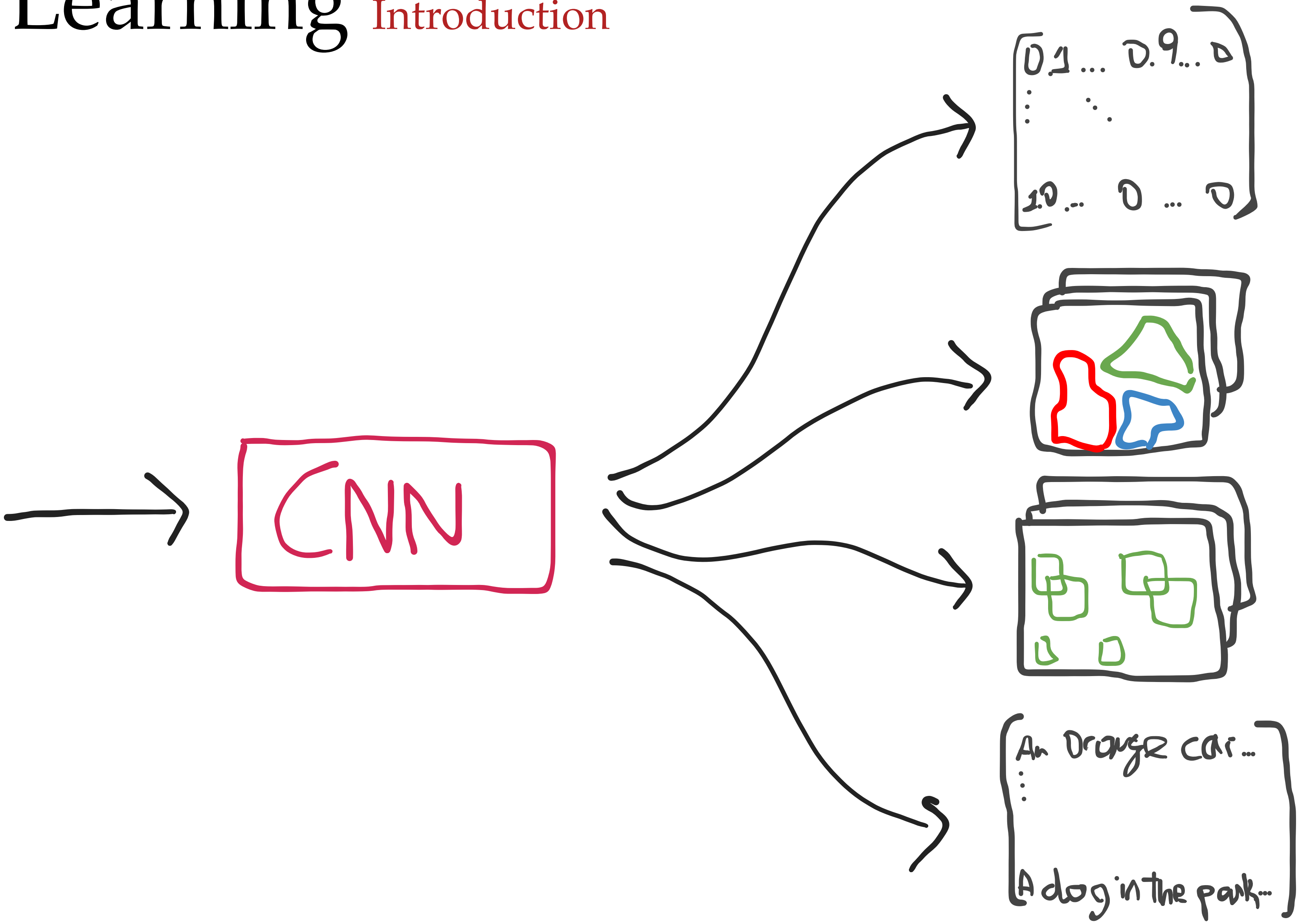
Schedule

Schedule

Representation Learning Introduction

Figure 1: Samples in the ImageNet 2012 dataset[1].

Source: cs.stanford.edu/people/karpathy/cnnembed.



[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and A.C. Berg. Imagenet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision*, 115, pp.211-252, 2015.

Representation Learning Introduction

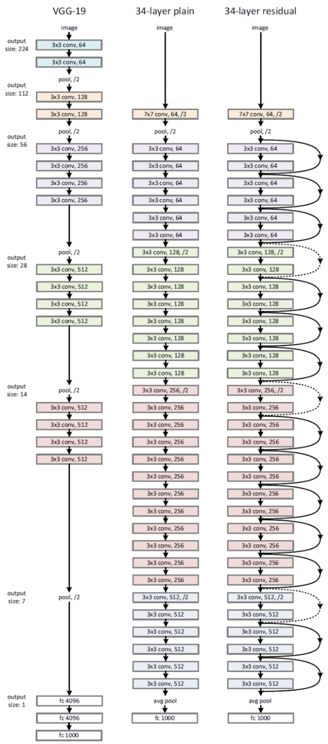


Figure 2: VGG-19, 34Plain and ResNet34 architectures[1].

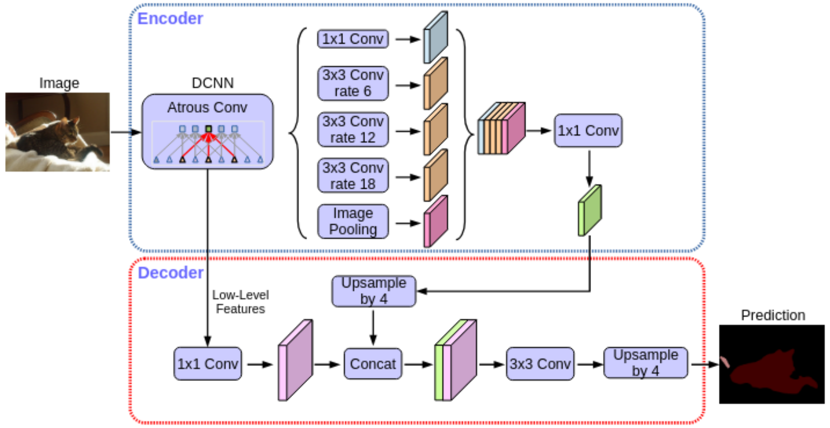


Figure 3: DeepLabV3+ architecture[2].

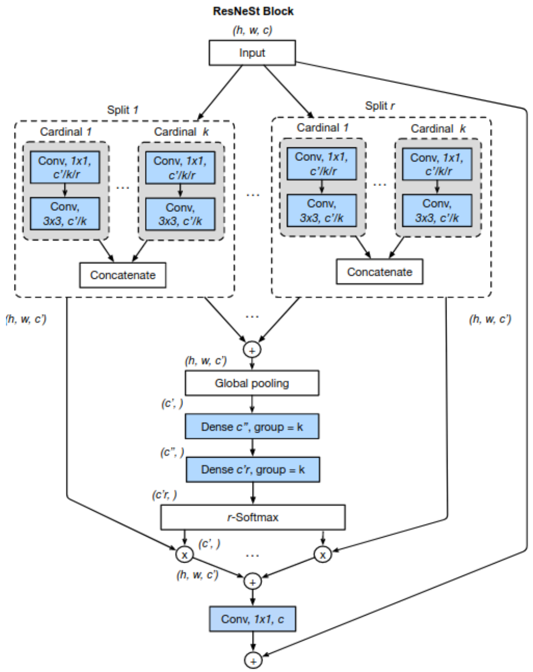


Figure 4: Split-Attention Block in the ResNeSt architecture.[3]

[1] Source: K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. 2016.
[2] Source: L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 801-818. 2018.
[3] Source: H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun et al. ResNeSt: Split-Attention Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2736-2746. 2022.

Complex Architectures Representation Learning

Models with millions of parameters are now the **standard**.

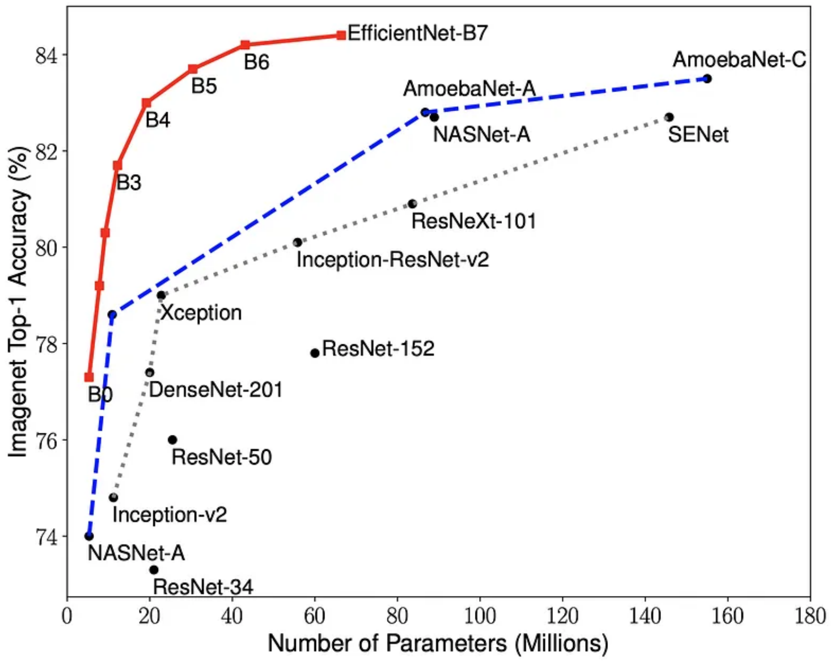


Figure 5: Models of various architectures, pre-trained over ImageNet. Source: Tan and Le[2].

[1] N. Burkart, and M.F. Huber. A survey on the explainability of supervised machine learning. In *Journal of Artificial Intelligence Research*, 70, pp.245-317., 73, pp.1-15. 2018.

[2] M. Tan, and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR. 2019.

Complex Architectures

Models with millions of parameters are now the **standard**.

But can we thrust their predictions?

And why do we have to?[1]

- Critical operations
- Medical diagnostics
- Finance systems
- Accountability and failure mitigation

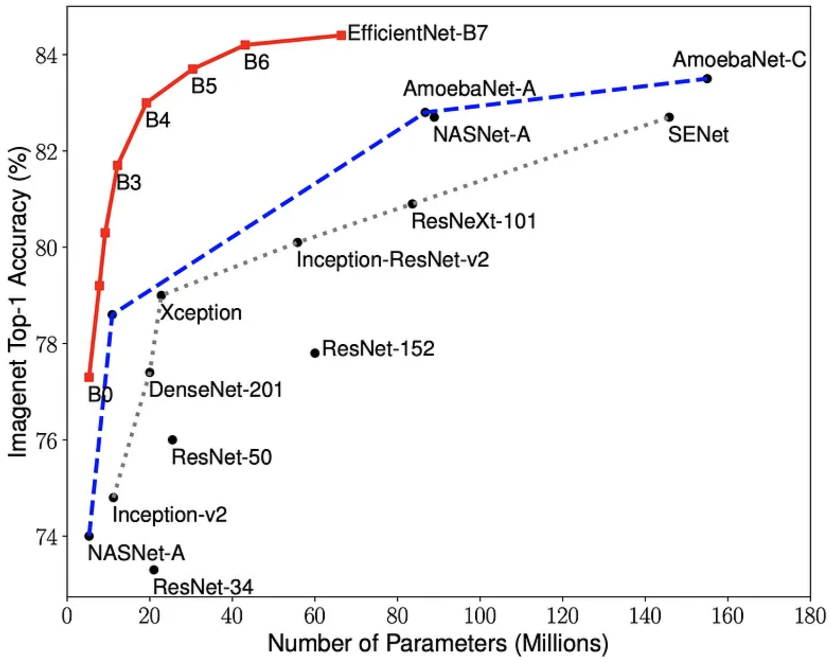


Figure 5: Models of various architectures, pre-trained over ImageNet. Source: Tan and Le[2].

[1] N. Burkart, and M.F. Huber. A survey on the explainability of supervised machine learning. In *Journal of Artificial Intelligence Research*, 70, pp.245-317., 73, pp.1-15. 2018.

[2] M. Tan, and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR. 2019.

Schedule

Explaining and Interpreting Models Introduction

"An interpretation is the mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of.[1]"

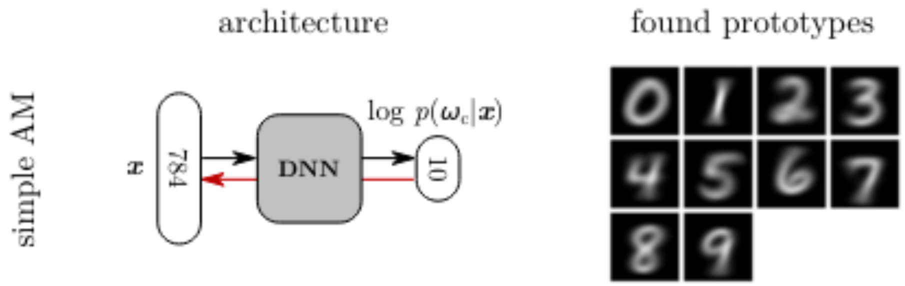


Figure 6: Illustration of Activation Maximization[2] applied to finding the prototypes for each class in the MNIST dataset. Source: Montavon et al.[1]

"An explanation is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression).[1]"

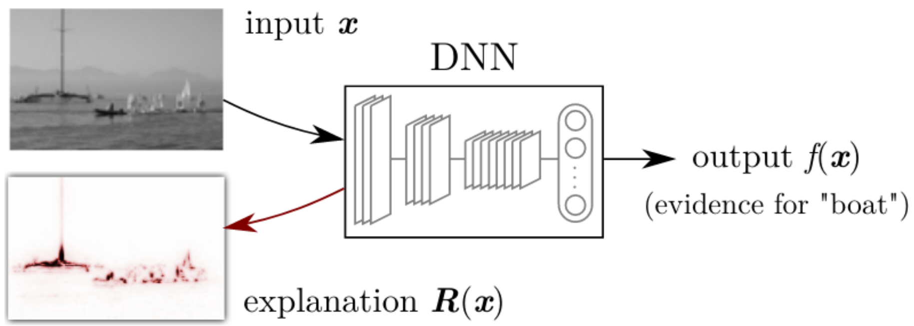


Figure 7: Example of the LRP method being applied to explain the prediction of class boat, given the image x. Source: Montavon et al.[1]

[1] G. Montavon, W. Samek, and K.R. Müller. Methods for Interpreting and Understanding Deep Neural Networks. In *Digital Signal Processing*, 73, pp.1-15. 2018.

[2] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

In Computer Vision Explainable AI

Explainability and explainable predictions:

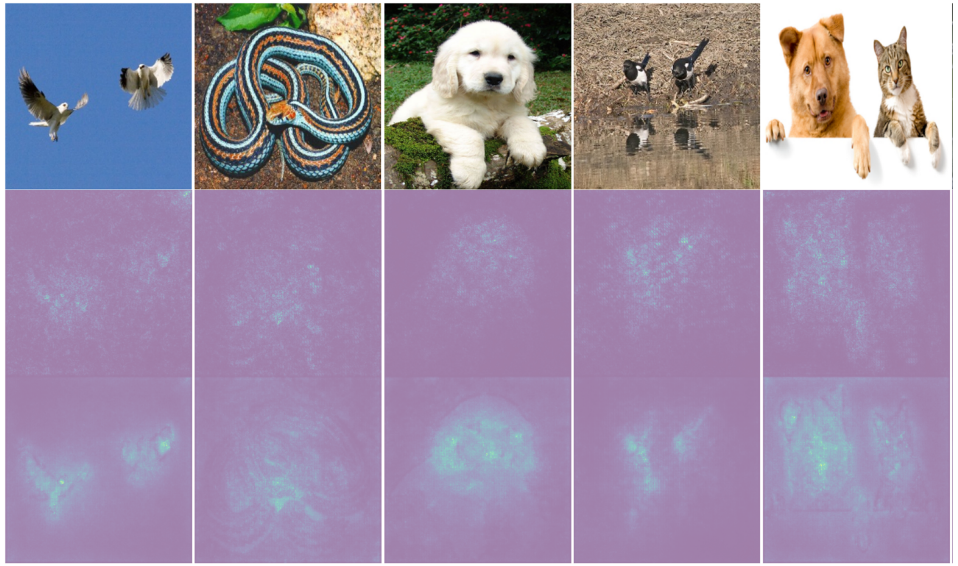
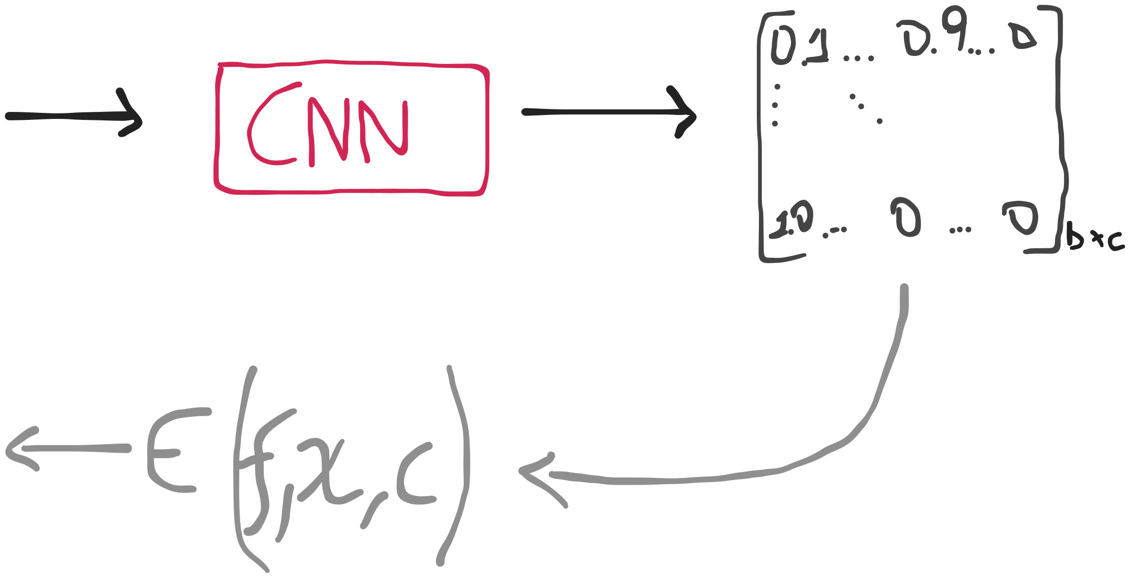


Figure 8: Sensitivity maps produced by Vanilla Gradient[1] (second row) and Smooth-Grad[2] (third row), when employed to explain the predictions made by a Xception model. Source: keras-explainable/methods/saliency/smoothgrad.



[1] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.
[2] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg. *SmoothGrad: removing noise by adding noise*. arXiv preprint arXiv:1706.03825. 2017.

In Computer Vision Explainable AI

Explainability and explainable predictions:

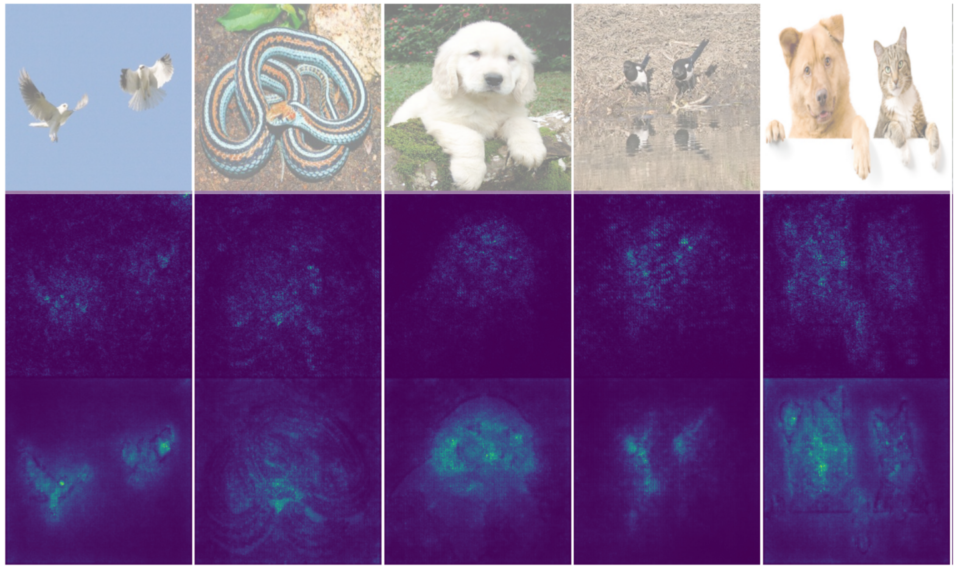
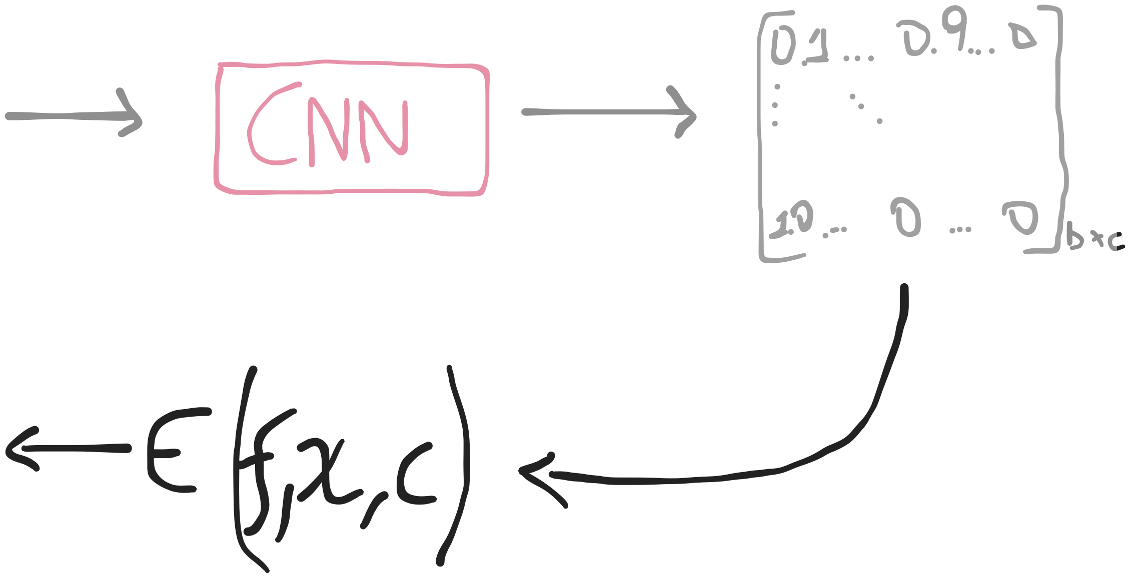


Figure 8: Sensitivity maps produced by Vanilla Gradient[1] (second row) and Smooth-Grad[2] (third row), when employed to explain the predictions made by a Xception model. Source: keras-explainable/methods/saliency/smoothgrad.



[1] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.
[2] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg. *SmoothGrad: removing noise by adding noise*. arXiv preprint arXiv:1706.03825. 2017.

In Computer Vision Explainable AI

Explainability and explainable predictions:

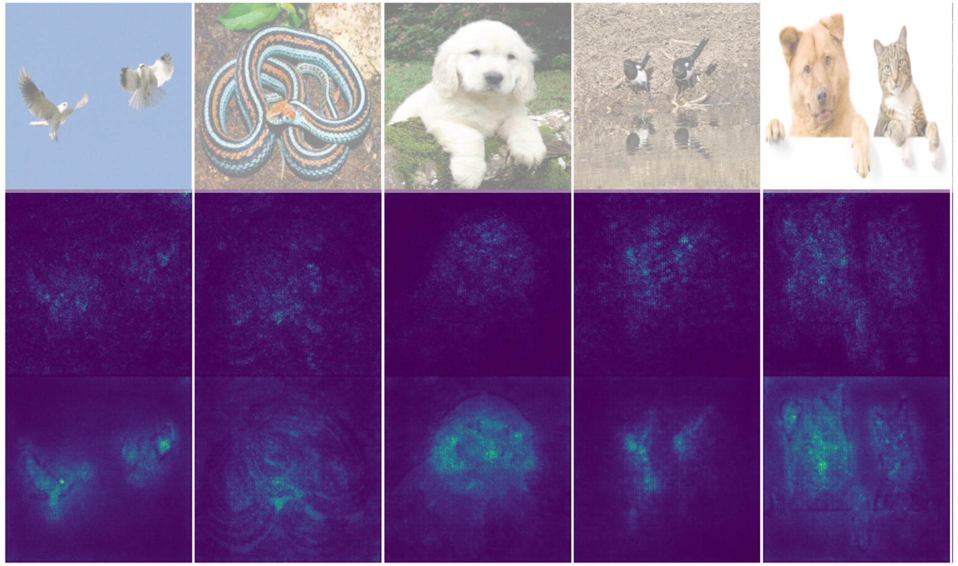


Figure 8: Sensitivity maps produced by Vanilla Gradient[1] (second row) and Smooth-Grad[2] (third row), when employed to explain the predictions made by a Xception model. Source: keras-explainable / methods / saliency / smoothgrad.

$$\longrightarrow \boxed{\text{CNN}} \longrightarrow \begin{bmatrix} 0.1 & \dots & 0.9 & \dots & 0 \\ \vdots & \ddots & & & \\ 1.0 & \dots & 0 & \dots & 0 \end{bmatrix}_{b \times c}$$

$$\longleftarrow E(f, x, c) \longleftarrow$$

Interesting Properties:

1. Completeness
2. Weak dependence
3. Class-specificity

[1] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.
[2] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg. *SmoothGrad: removing noise by adding noise*. arXiv preprint arXiv:1706.03825. 2017.

In Computer Vision Explainable AI

Leveraging internalized knowledge
to solve different tasks:

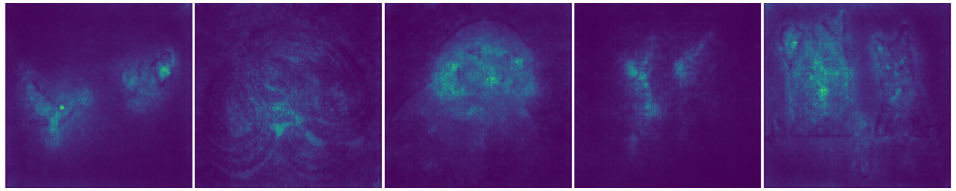
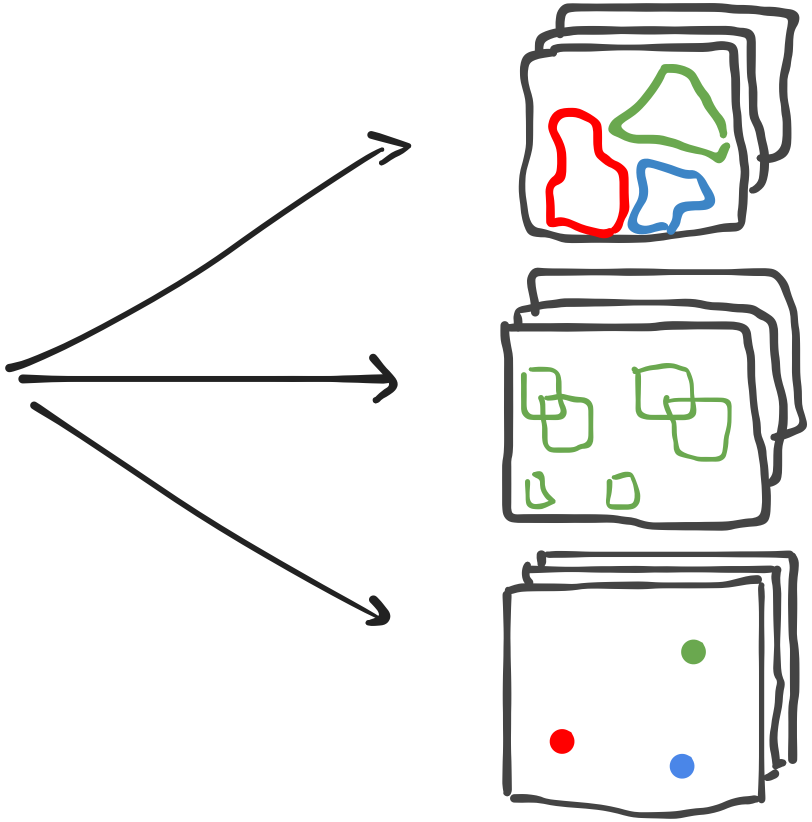


Figure 9: Sensitivity maps produced by Smooth-Grad.

Source: keras-explainable / methods / saliency / smoothgrad.



Schedule

Semantic (and others) Segmentation Introduction

Figure 10: Samples, proposals[1] and ground-truth segmentation annotation from the Pascal VOC 2012 dataset.

Figure 11: Example of samples and ground-truth panoptic segmentation annotation from the MS COCO 2017 dataset. Source: https://cocodataset.org/#panoptic-2020.

Figure 12: Example of semantic segmentation produced by ICNet for a video sample in the Cityscapes dataset. Source: https://gitplanet.com/project/fast-semantic-segmentation.

[1] H. Xiao, D. Li, H. Xu, S. Fu, D. Yan, K. Song, and C. Peng. Semi-Supervised Semantic Segmentation with Cross Teacher Training. *Neurocomputing*, 508, pp.36-46. 2022.
[2] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *European Conference on Computer Vision (ECCV)*, pp. 405-420. 2018.
[3] L. Chan, M.S. Hosseini. and K.N. Plataniotis. A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains. In *International Journal of Computer Vision*, 129, pp.361-384. 2021.

Semantic (and others) Segmentation Introduction

Figure 13: Example of road segmentation in SpaceNet dataset. Source: https://www.v7labs.com/open-datasets/spacenet

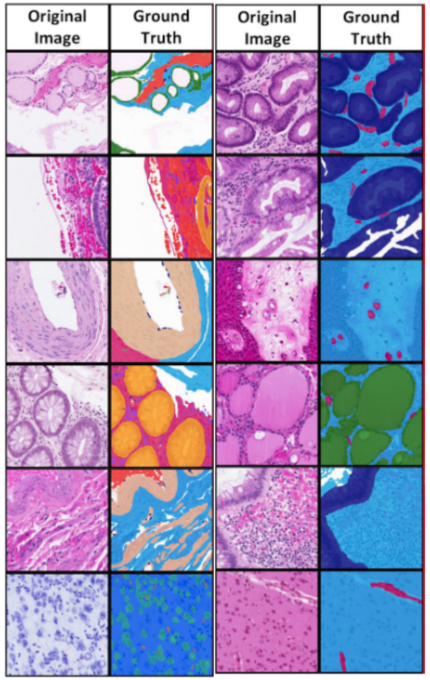


Figure 14: Example of (a) morphological and (b) functional segmentation of samples in the Atlas of Digital Pathology dataset. Source: L. Chan et al.

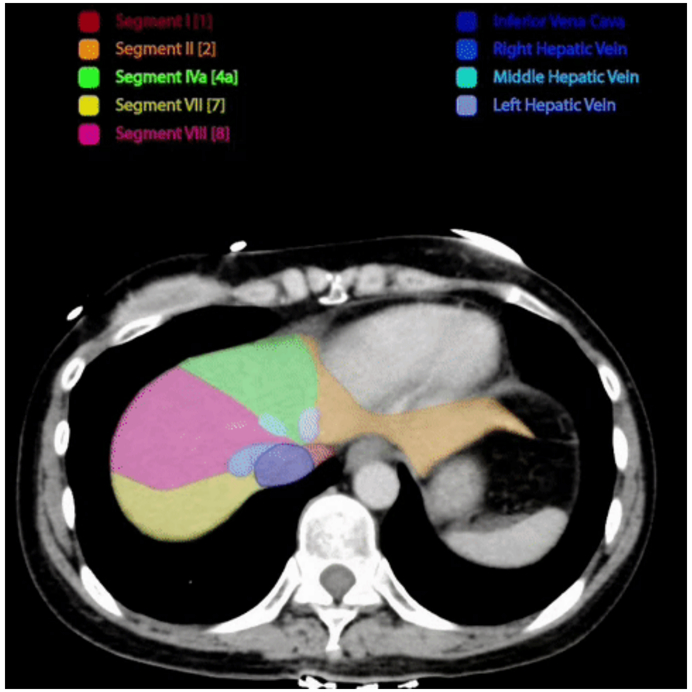


Figure 15: Example of annotated CT Scan image. Source: https://radiopaedia.org/cases/liver-segments-annotated-ct-1

[1] H. Xiao, D. Li, H. Xu, S. Fu, D. Yan, K. Song, and C. Peng. Semi-Supervised Semantic Segmentation with Cross Teacher Training. *Neurocomputing*, 508, pp.36-46. 2022.
[2] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *European Conference on Computer Vision (ECCV)*, pp. 405-420. 2018.
[3] L. Chan, M.S. Hosseini. and K.N. Plataniotis. A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains. In *International Journal of Computer Vision*, 129, pp.361-384. 2021.

How It is Done? Semantic Segmentation

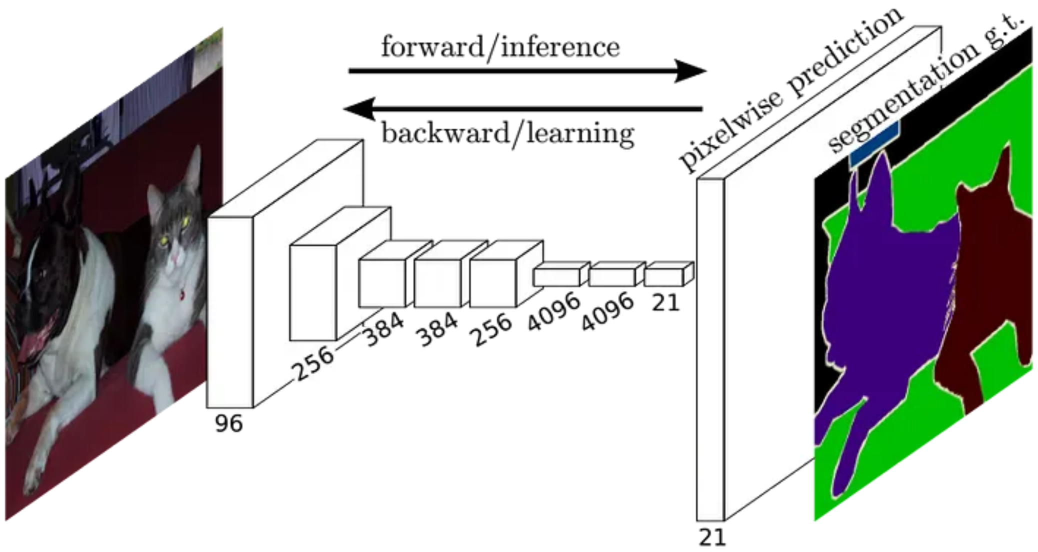


Figure 16: Fully Convolutional Network (FCN) architecture[1], mapping image samples to their respective semantic segmentation maps.

[1] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

How It is Done? Semantic Segmentation

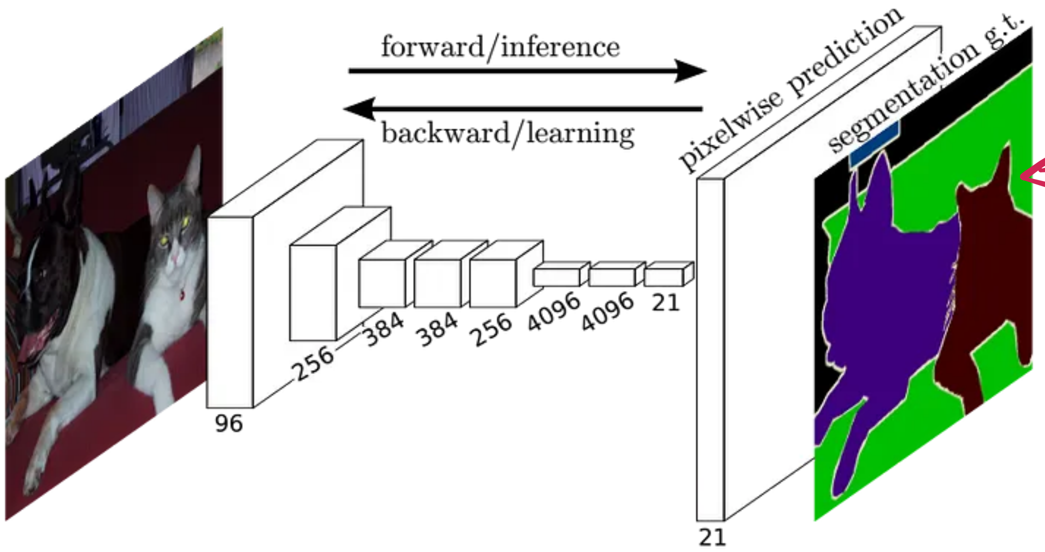


Figure 16: Fully Convolutional Network (FCN) architecture[1], mapping image samples to their respective semantic segmentation maps.

This information needs the be known and available at training time.

$$\mathrm{CE}(p_i, y_i) = -\sum_{c=1}^{M} y_{ic} \log(p_{ic})$$

Equation 1: The (naive) categorical cross-entropy loss function.

[1] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

(Fully) Supervised Learning Semantic Segmentation

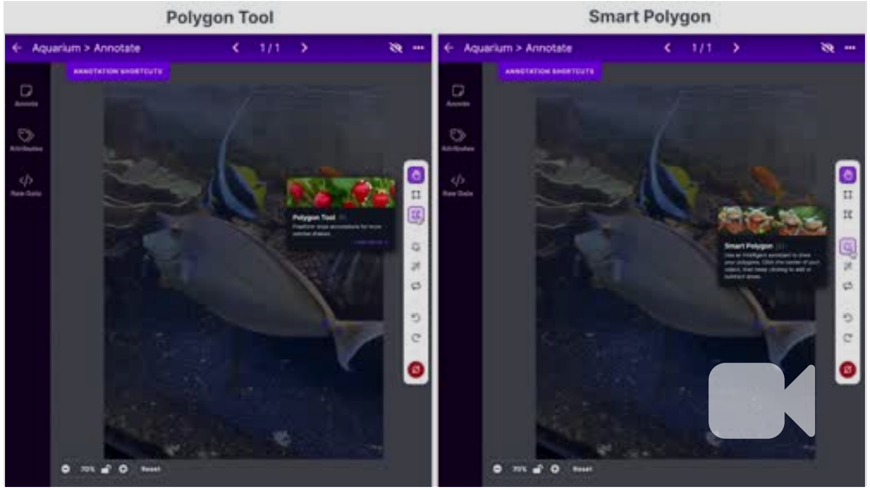


Figure 17: Segmentation annotation example using RoboFlow.
Source: https://blog.roboflow.com/semantic-segmentation-roboflow.

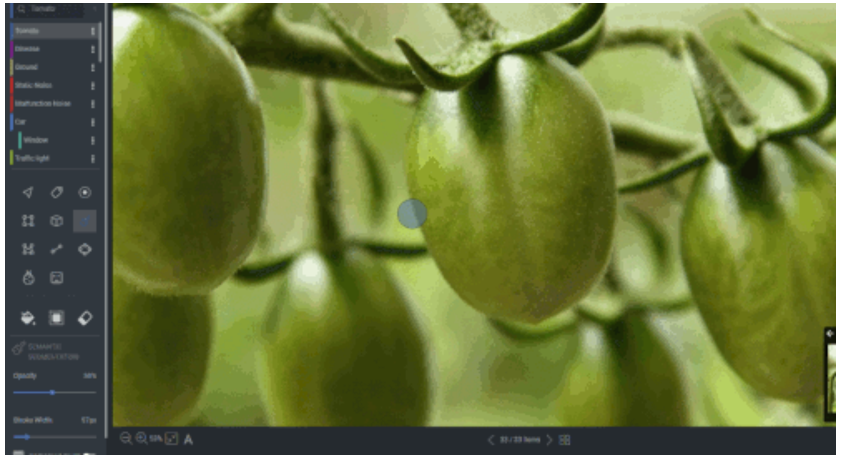


Figure 18: Segmentation annotation example using Dataloop.
Source: https://dataloop.ai/docs.

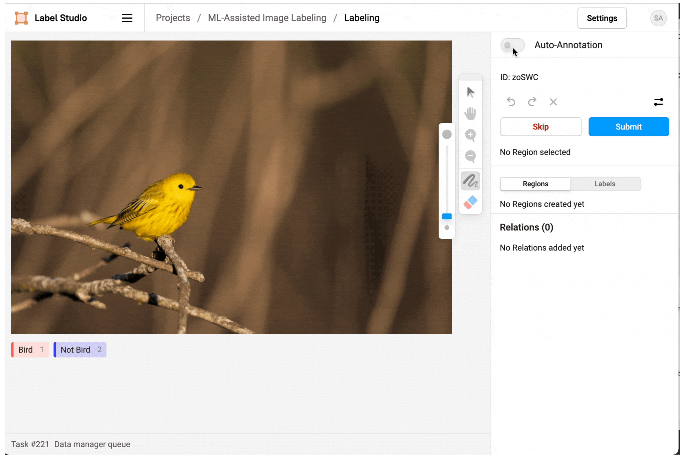
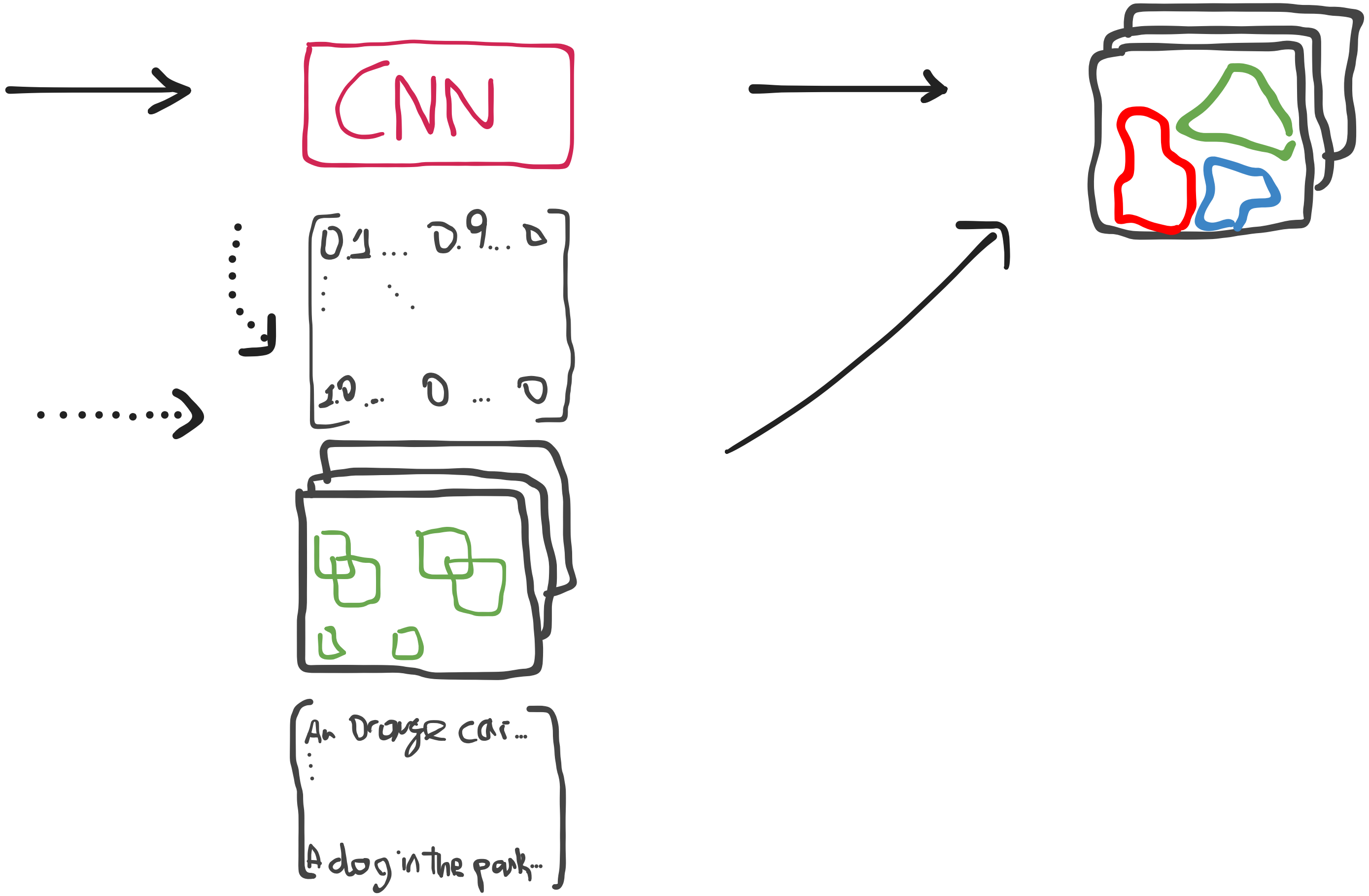


Figure 19: Segmentation annotation example using LabelStudio. Source: https://labelstud.io/blog/perform-interactive-ml-assisted-labeling-with-label-studio-1-3-0.

Coarse annotations are quickly drawn, but lack quality (e.g., precision);
Detailed annotations take time, patience, people and resources;
Assisting labeling tools can speed up this task.

(Weakly) Supervised Learning Semantic Segmentation

Figure 20: Samples in the ImageNet 2012 dataset[1]. Source: cs.stanford.edu/people/karpathy/cnnembed.



[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and A.C. Berg. Imagenet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision*, 115, pp.211-252, 2015.

Schedule

1. **Introduction**
   1.1. Representation Learning
   1.2. Explaining and Interpreting Models
   1.3. Weakly Supervised Semantic Segmentation
   1.4. Research Goals
2. Related Work
3. Research Proposal
4. Preliminary Results
5. Final Considerations

Research Goals

1. To study Class-Specific XAI methods in the multi-label scenarios

2. To study promising weakly supervised strategies and to propose new ones

3. To investigate the behavior of WSSS solutions to more complex boundary cases, such as long-tail and ambiguous functional segmentation problems

Schedule

1. Introduction
2. **Related Work**
3. Research Proposal
4. Preliminary Results
5. Final Considerations

Schedule

Explainable AI Related Work

Figure 21: Sensitivity maps produced by Vanilla Gradient[1] (2nd col) and Full-Grad[2] (3rd col), when employed to explain the predictions made by a ResNet50 model.

Source: keras-explainable.

$$\text{If } f_c \approx w^\intercal I + b,$$

$$S_{f_c}(I_0) = \psi\Big(\frac{\partial f_c}{\partial I}\Big|_{I_0}\Big)$$

Equation 2: Saliency map for the concept c of a model S with respect to an input image x, generated by the (Vanilla) Gradients method[1].

$$S_{f_c}(I_0) = \psi(\nabla_I f(I) \circ I_0) + \textstyle\sum_{l \in L, k \in C_l} \psi(f_b^k(x))$$

Equation 3: Saliency map for the concept c of a model S with respect to an input image x, generated by the Full-Gradient method[2].

[1] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.

[2] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization. In *Advances in neural information processing systems*, 32. 2019.

Explainable AI Related Work

Figure 21: Sensitivity maps produced by Vanilla Gradient[1] (2nd col) and Full-Grad[2] (3rd col), when employed to explain the predictions made by a ResNet50 model.

Source: keras-explainable.

$$\text{If } f_c \approx w^\intercal I + b,$$

$$S_{f_c}(I_0) = \psi\Big(\frac{\partial f_c}{\partial I}\Big|_{I_0}\Big)$$

Equation 2: Saliency map for the concept c of a model S with respect to an input image x, generated by the (Vanilla) Gradients method[1].

$$S_{f_c}(I_0) = \psi(\nabla_I f(I) \circ I_0) + \sum_{l \in L, k \in C_l} \psi(f_b^k(x))$$

Equation 3: Saliency map for the concept c of a model S with respect to an input image x, generated by the Full-Gradient method[2].

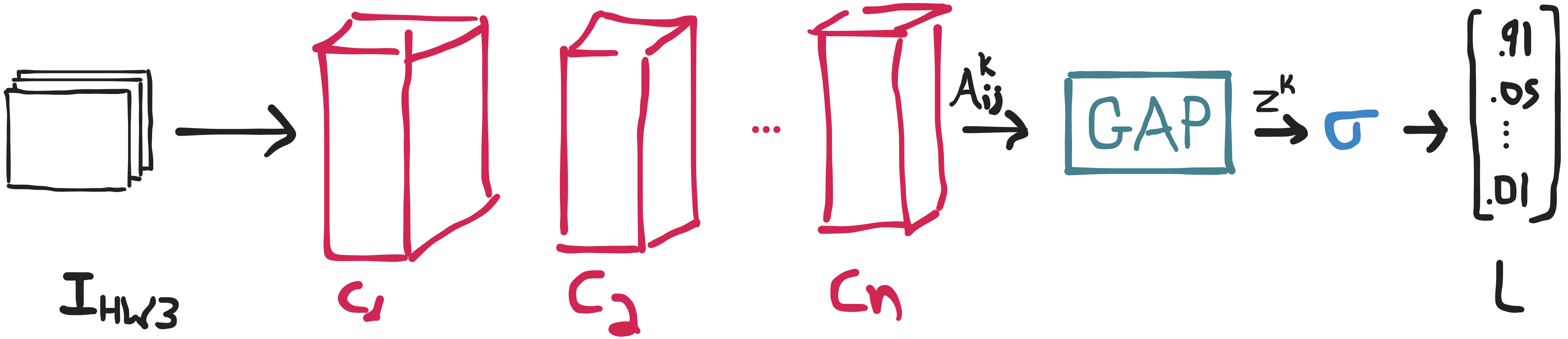
Lack class-sensibility

Expensive to compute

[1] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.

[2] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization. In *Advances in neural information processing systems*, 32. 2019.

Class Activation Mapping Explainable AI



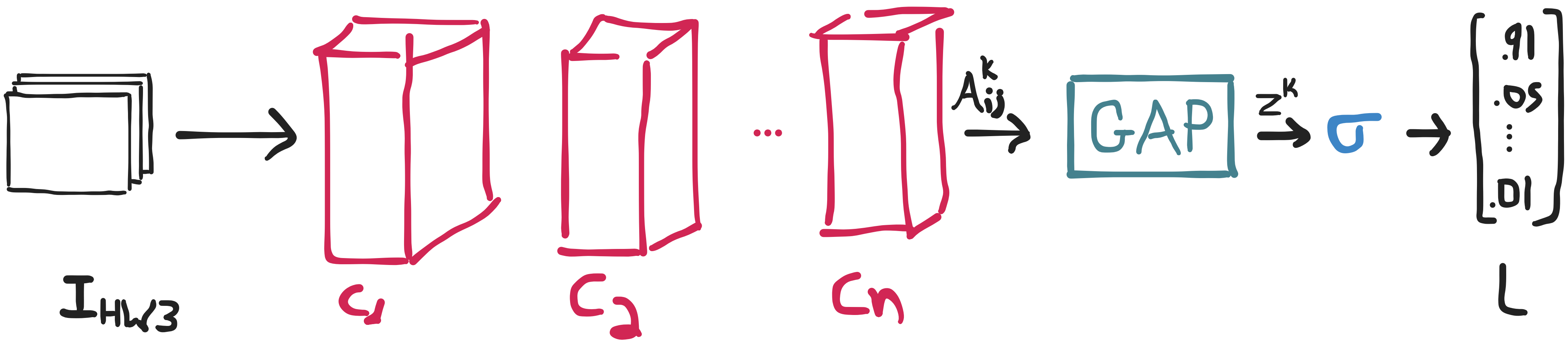
$$f(x) = \sum_k \underline{w_k^c}\, \underline{\mathrm{GAP}}(\underline{A^k}) = \sum_k \underline{w_k^c}\, \underline{\frac{1}{hw} \sum_{ij}}\, \underline{A_{ij}^k}$$

$$f(x) = \frac{1}{hw} \sum_{ij} \sum_k w_k^c A_{ij}^k = \mathrm{GAP}(w^c \cdot A)$$

Equation 4: Feed-Forward for a for Convolutional Networks containing GAP layers and the formulation for CAM[1].

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929. 2016.

Class Activation Mapping Explainable AI



$$f(x) = \sum_k \underline{w_k^c} \underline{\mathrm{GAP}}(\underline{A^k}) = \sum_k \underline{w_k^c} \underline{\frac{1}{hw} \sum_{ij}} \underline{A_{ij}^k}$$

$$f(x) = \frac{1}{hw} \sum_{ij} \sum_k w_k^c A_{ij}^k = \mathrm{GAP}(w^c \cdot A)$$

Equation 4: Feed-Forward for a for Convolutional Networks containing GAP layers and the formulation for CAM[1].

$$\implies L_{\mathrm{CAM}}^c(f, x) = \sum_k w_k^c A^k$$

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929. 2016.

Class Activation Mapping Explainable AI

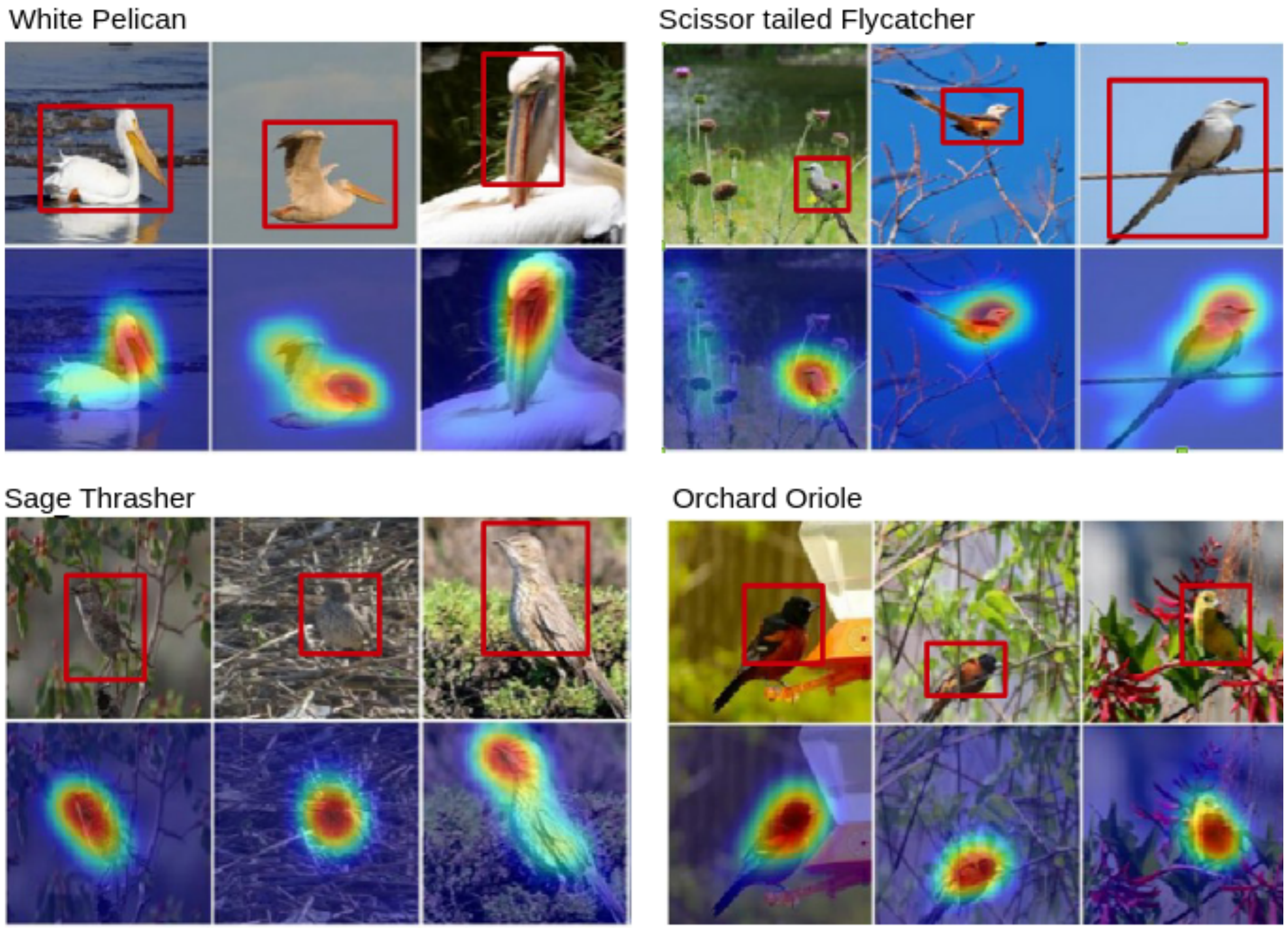


Figure 22: Examples of CAMs and approximate bounding boxes found for different birds in the CUB200 dataset. Source: Zhou et al.[1]

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929. 2016.

Extensions and Alternatives

Grad-CAM

Goal: to explain more complex networks, with non-linear (and yet smooth) operations after the GAP layer.

$$L^c_{\text{Grad-CAM}}(f, x) = \mathrm{ReLU}(\textstyle\sum_k \alpha^c_k A^k)$$
$$\alpha^c_k = \frac{1}{hw} \textstyle\sum_{ij} \frac{\partial f_c(x)}{\partial A^k_{ij}}$$

Equation 5: Definition for Grad-CAM visual explaining method, for an arbitrary convolutional network *f*.

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision*, pp. 618-626. 2017.

Extensions and Alternatives

Grad-CAM

Goal: to explain more complex networks, with non-linear (and yet smooth) operations after the GAP layer.

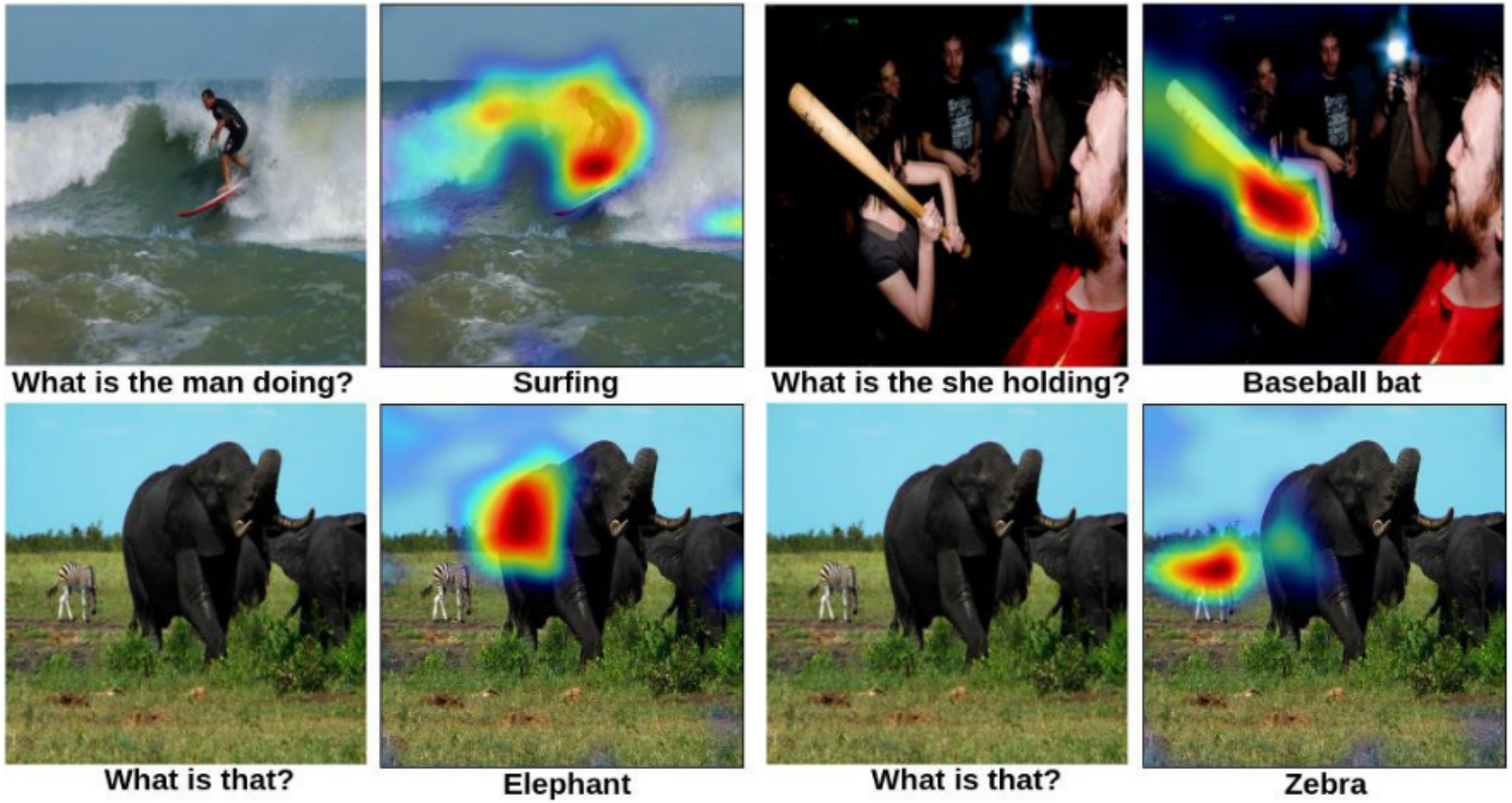


Figure 23: Examples of Grad-CAM being utilized to explaing a Visual Questioning Network based on convolutional layers and LSTM layers. Source: Selvaraju et al.[1]

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision*, pp. 618-626. 2017.

Extensions and Alternatives CAM-Based Explaining Methods

Grad-CAM++

Goal: to activate homogeneously over all instances of the explained concept lying the the visual receptive field.

$$L^c_{\text{Grad-CAM++}}(f, x) = \text{ReLU}\left(\sum_k \sum_{ij} \alpha^{kc}_{ij} \text{ReLU}\left(\frac{\partial S_c}{\partial A^k_{ij}} \right) A^k \right)$$

$$\alpha^{kc}_{ij} = \frac{\frac{\partial^2 S_c}{(\partial A^k_{ij})^2}}{2 \frac{\partial^2 S_c}{(\partial A^k_{ij})^2} + \sum_{ab} A^k_{ab} \frac{\partial^3 S_c}{(\partial A^k_{ij})^3}}$$

Equation 6: Definition of Grad-CAM++ visual explaining method.

[1] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 839-847. IEEE, 2018.

Extensions and Alternatives CAM-Based Explaining Methods

Grad-CAM++

Goal: to activate homogeneously over all instances of the explained concept lying the the visual receptive field.

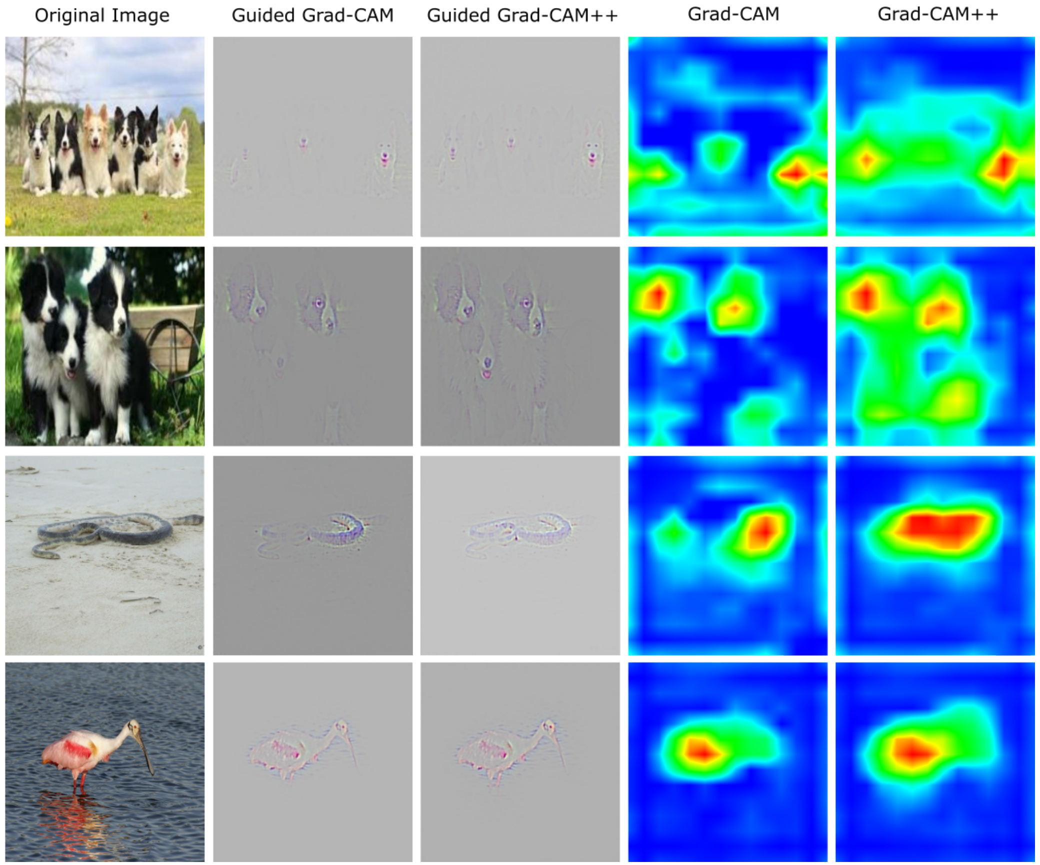


Figure 24: Grad-CAM and Grad-CAM++ being applied to samples in the ImageNet dataset. Source: Chatopadhay et al.[1]

[1] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 839-847. IEEE, 2018.

Score-CAM

Goal: to combine the many activation maps, weighted by their contribution towards the *Average Drop %* metric.

$$L^c_{\text{Score-CAM}}(f, x) = \text{ReLU}\left(\sum_k f_c\left(x \circ \frac{A^k}{\max A^k}\right) A^k\right)$$

Equation 7: Definition of the Score-CAM visual explaining method[1].

[1] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 24-25. 2020.

Extensions and Alternatives

Score-CAM

Goal: to combine the many activation maps, weighted by their contribution towards the *Average Drop %* metric.

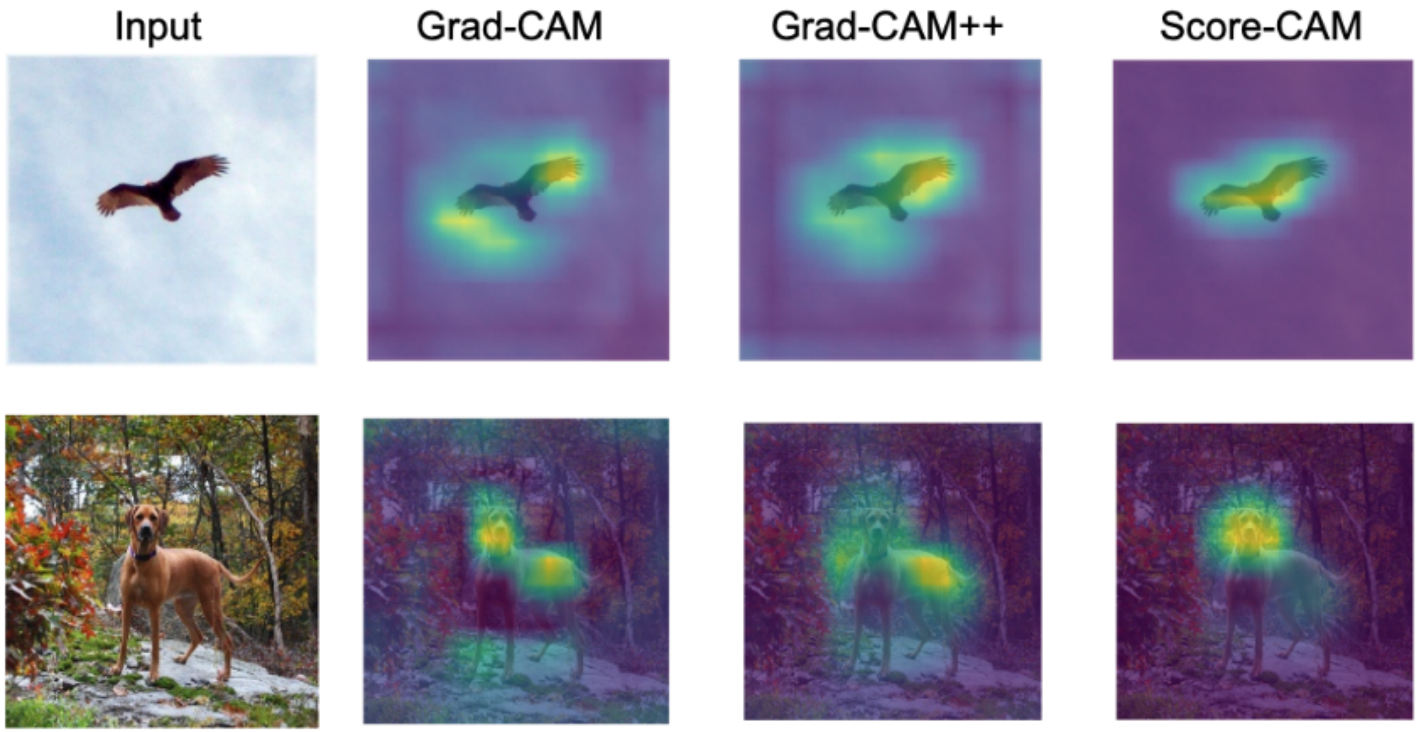


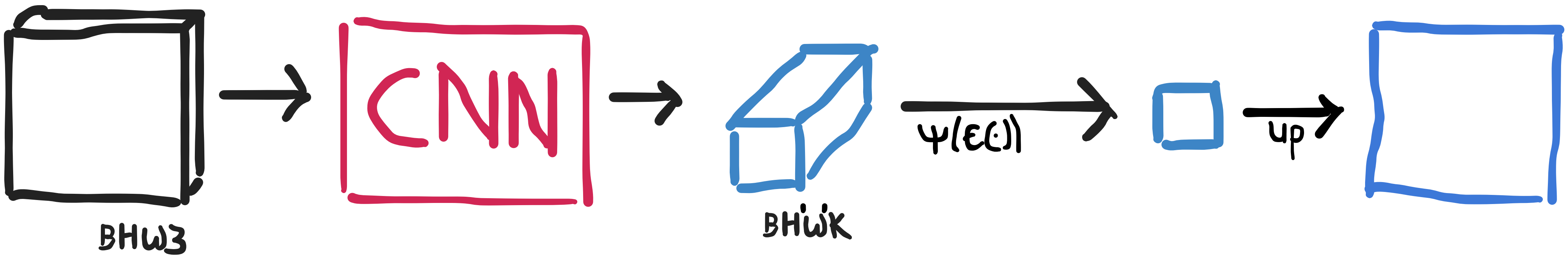
Figure 25: Examples of sensitivity maps obtained from Grad-CAM, Grad-CAM++ and Score-CAM.

Source: Wang et al.[1]

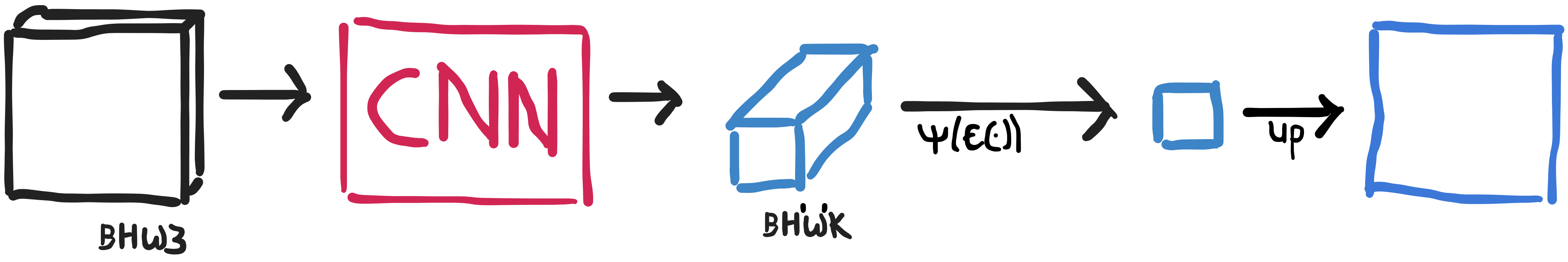
[1] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 24-25. 2020.

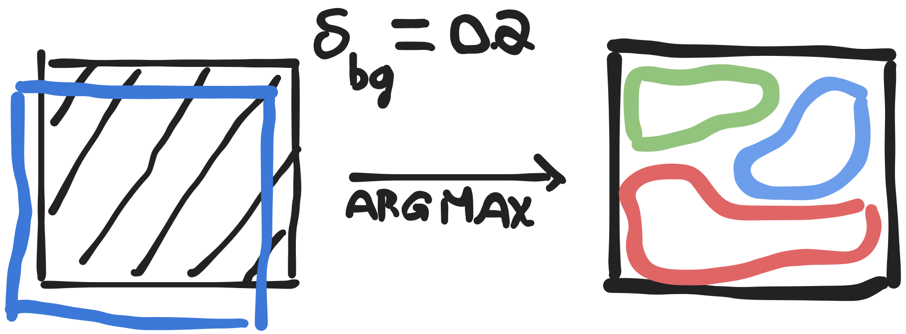
Schedule

Weakly Supervised Semantic Segmentation Related Work



Weakly Supervised Semantic Segmentation





Coarse Semantic Segmentation Priors WSSS

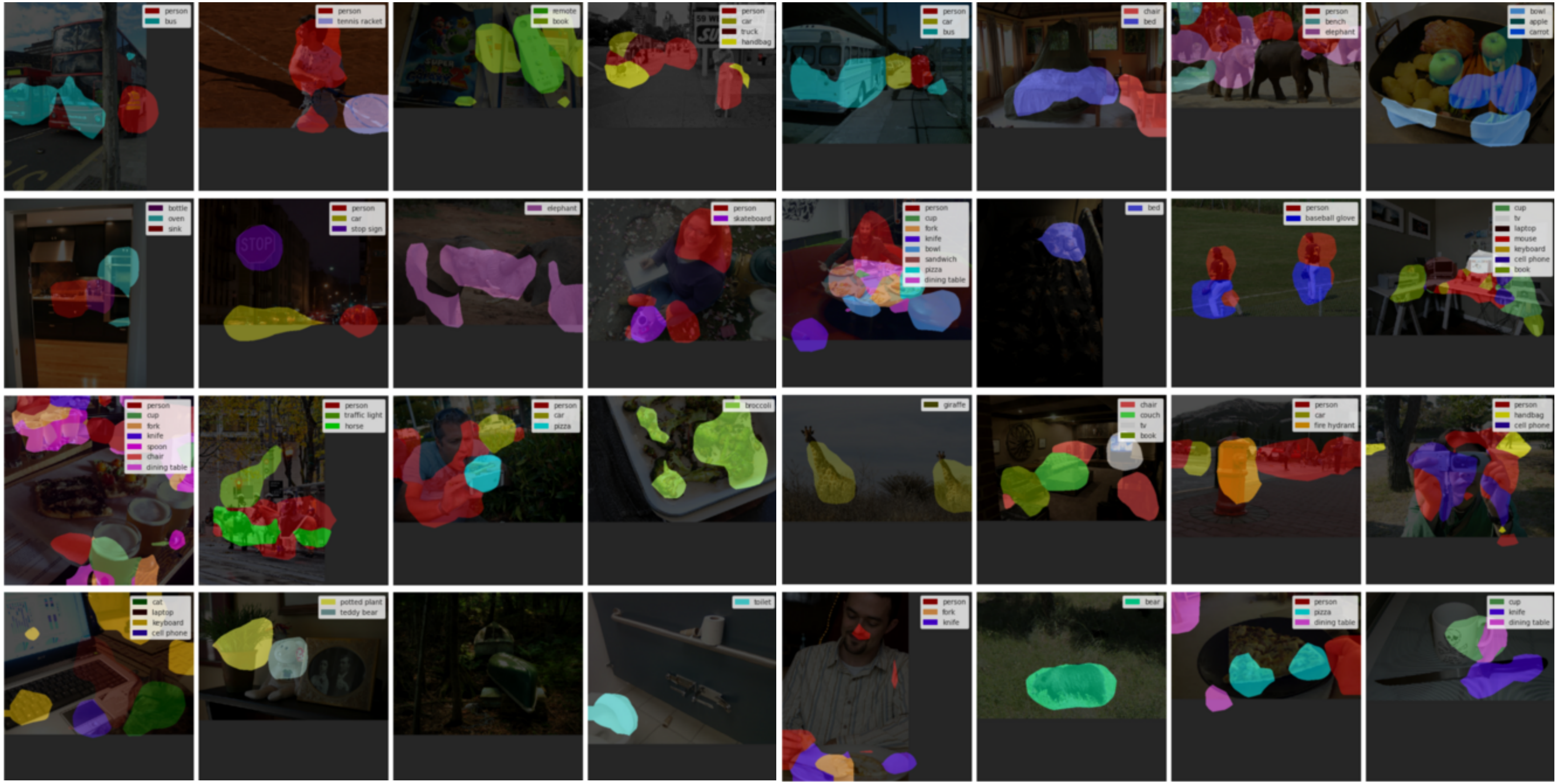


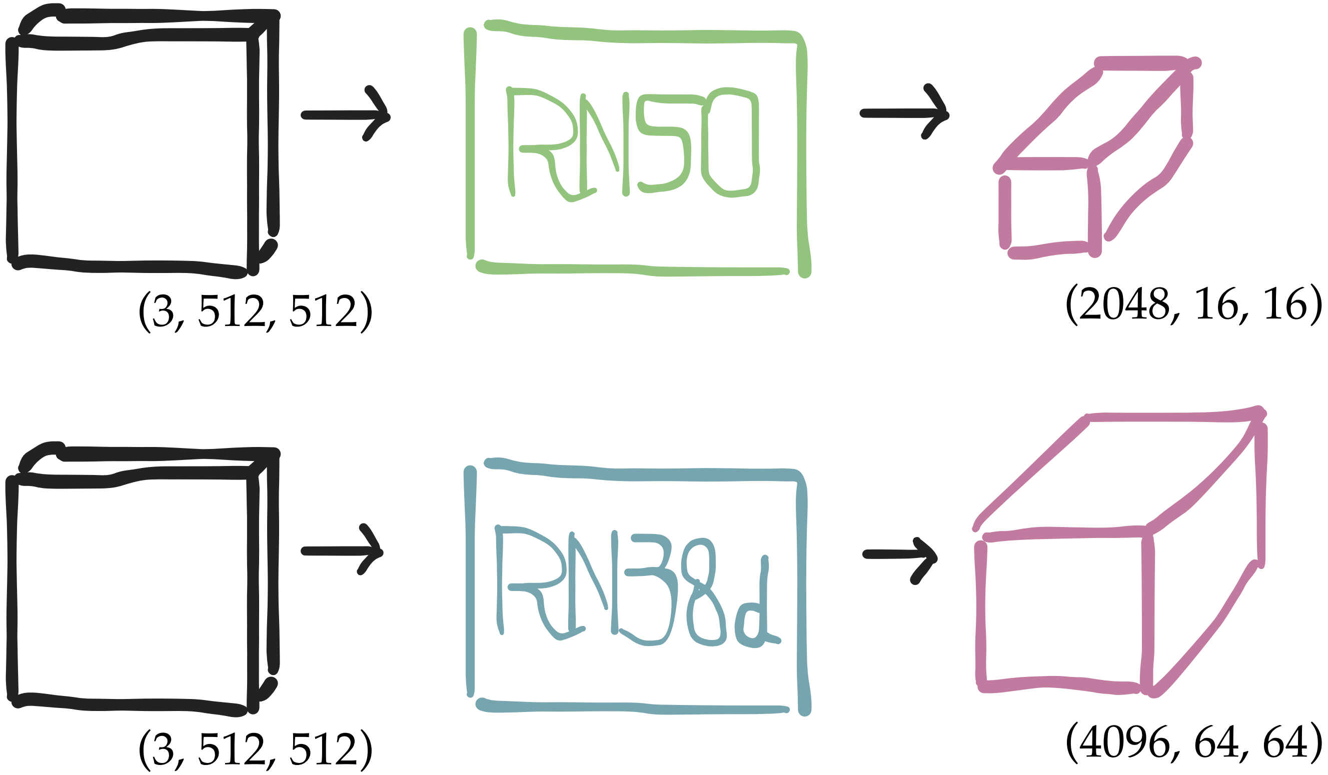
Figure 26: Semantic Segmentation Priors produced by *thresholding* CAMs devised from a ResNet101 model trained over MS COCO 2017 dataset.

Refinement of Segmentation Masks WSSS

1. Architectural
2. Pixel neighborhood affinity and similarity
3. Many other strategies: Seed-Expand-Constrain; region semantic-based clustering; token-based similarity matching, etc.

Refinement of Segmentation Masks WSSS

1. Architectural



[1] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. In *Pattern Recognition*, 90, pp.119-133. 2019.

Refinement of Segmentation Masks WSSS

1. Architectural

- Fewer layers, more units
- "Bottleneck" blocks
- Strong dropout
- Dilation

[1] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. In *Pattern Recognition*, 90, pp.119-133. 2019.

FC Conditional Random Fields Refinement of Segmentation Masks

$$E(x) = \underbrace{\textstyle\sum_i \psi_u(x_i)}_{\text{unary}} + \underbrace{\textstyle\sum_{i<j} \psi_p(x_i, x_j)}_{\text{pairwise}}$$

$$\psi_p(x_i, x_j) = \mu(x_i, x_j)\Big[w^{(1)} \exp\Big(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\Big) + w^{(2)} \exp\Big(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\Big)\Big]$$

label compatibility
function (learnable)

appearance kernel

smoothness kernel

[1] P. Krähenbühl, and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 24. 2011.

FC Conditional Random Fields Refinement of Segmentation Masks

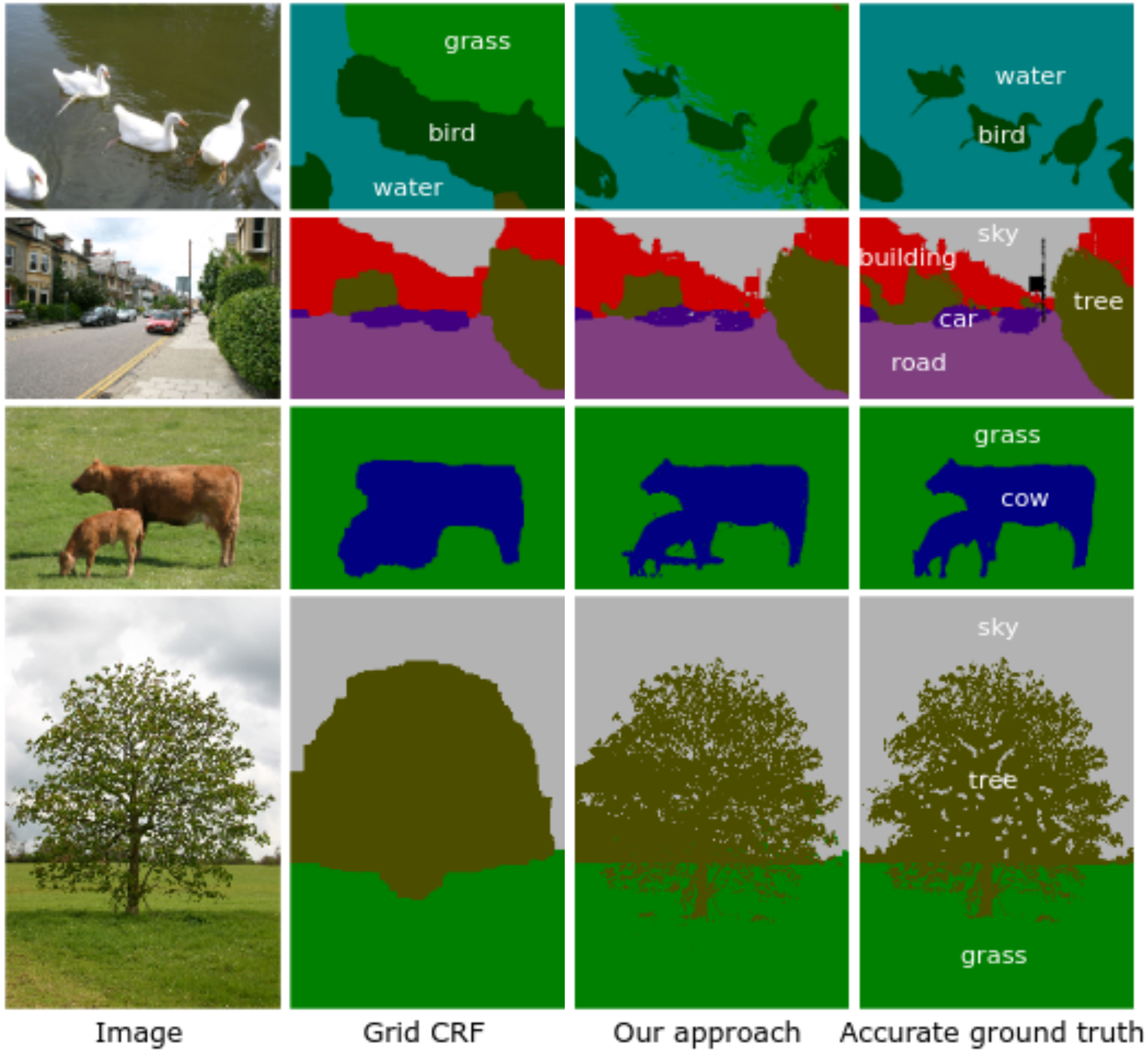


Figure 27: Qualitative results of dCRF. Source: Krähenbühl and Koltun[1].

[1] P. Krähenbühl, and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 24. 2011.

Pixel Semantic Affinity Refinement of Segmentation Masks

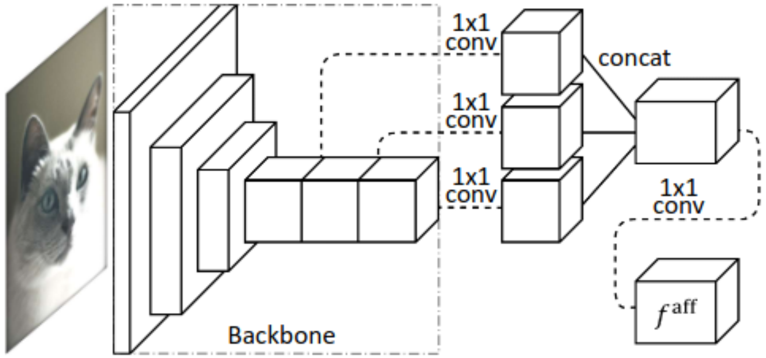


Figure 5: AffinityNet architecture. Source: Ahn and Kwak[1].

Pairs extraction

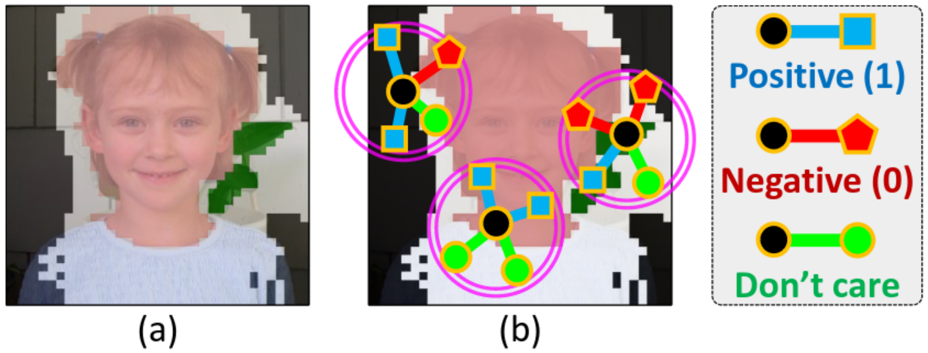


Figure 5: Illustration of pairs of pixels selected for affinity evaluation.
Source: Ahn and Kwak[1].

[1] J. Ahn, and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4981-4990. 2018.

Pixel Semantic Affinity

Refinement of Segmentation Masks

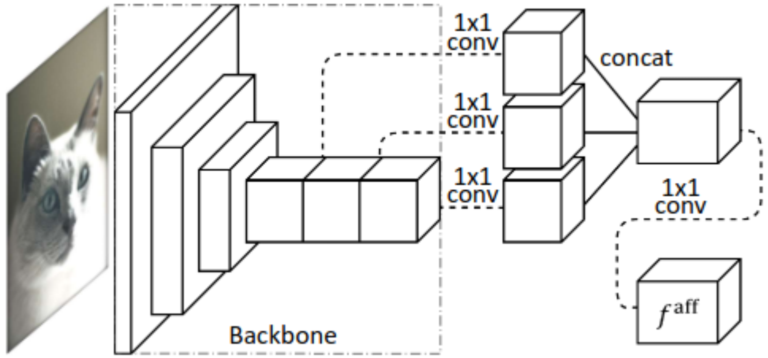


Figure 5: AffinityNet architecture. Source: Ahn and Kwak[1].

Pairs extraction

Training

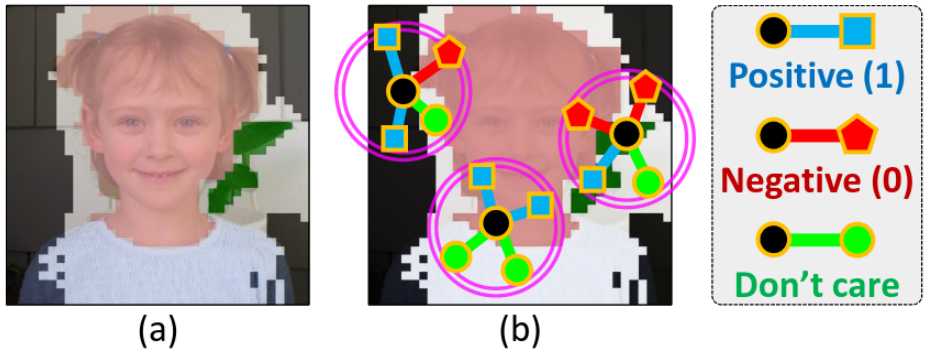


Figure 5: Illustration of pairs of pixels selected for affinity evaluation. Source: Ahn and Kwak[1].

$$W_{ij} = \exp\{-\|f(x_i, y_i) - f(x_j, y_j)\|_1\}$$

$$\mathcal{L} = \mathcal{L}^+_{\text{fg}} + \mathcal{L}^+_{\text{bg}} + 2\mathcal{L}^-$$

$$\begin{aligned}\mathcal{L} = &-\frac{1}{|\mathcal{P}^+_{\text{fg}}|}\sum\nolimits_{ij\in\mathcal{P}^+_{\text{fg}}} \log W_{ij} \\ &-\frac{1}{|\mathcal{P}^+_{\text{bg}}|}\sum\nolimits_{ij\in\mathcal{P}^+_{\text{bg}}} \log W_{ij} \\ &-2\frac{1}{|\mathcal{P}^-|}\sum\nolimits_{ij\in\mathcal{P}^-} \log(1 - W_{ij})\end{aligned}$$

[1] J. Ahn, and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4981-4990. 2018.

Pixel Semantic Affinity Refinement of Segmentation Masks

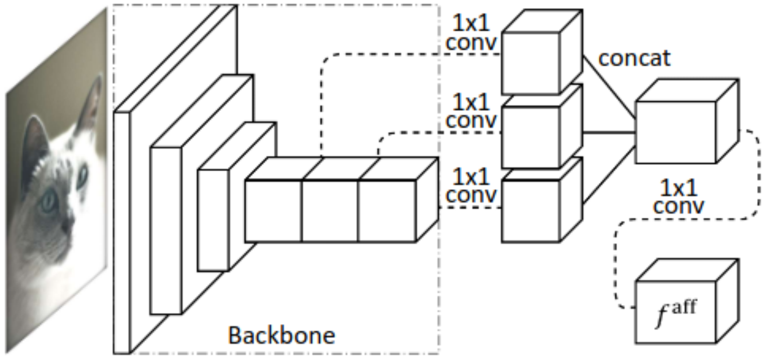


Figure 5: AffinityNet architecture. Source: Ahn and Kwak[1].

Pairs extraction

Training

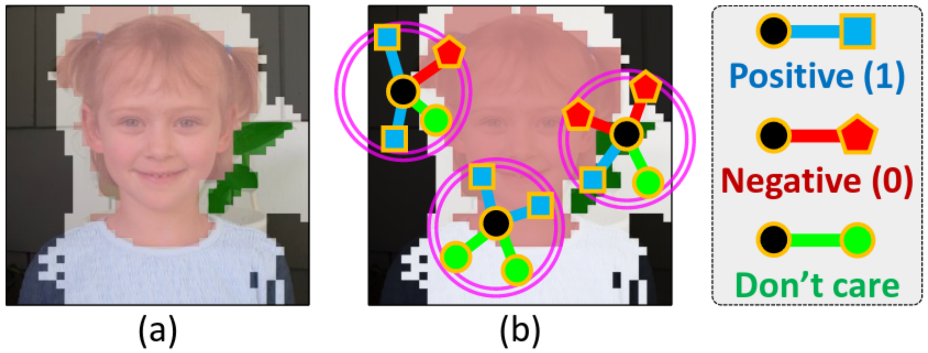


Figure 5: Illustration of pairs of pixels selected for affinity evaluation. Source: Ahn and Kwak[1].

$$W_{ij} = \exp\{-\|f(x_i, y_i) - f(x_j, y_j)\|_1\}$$

$$\mathcal{L} = \mathcal{L}^+_{\text{fg}} + \mathcal{L}^+_{\text{bg}} + 2\mathcal{L}^-$$

$$\mathcal{L} = -\frac{1}{|\mathcal{P}^+_{\text{fg}}|} \sum_{ij\in\mathcal{P}^+_{\text{fg}}} \log W_{ij} - \frac{1}{|\mathcal{P}^+_{\text{bg}}|} \sum_{ij\in\mathcal{P}^+_{\text{bg}}} \log W_{ij} - 2\frac{1}{|\mathcal{P}^-|} \sum_{ij\in\mathcal{P}^-} \log(1 - W_{ij})$$

Inference

$$T = D^{-1} W^{\circ\beta}, D_{ii} = \sum_j W^\beta_{ij}$$

$$\text{vec}(M^*_c) = T^t \cdot \text{vec}(M_c), \forall c \in C \cup \{\text{bg}\}$$

[1] J. Ahn, and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4981-4990. 2018.

Pixel Semantic Affinity Refinement of Segmentation Masks

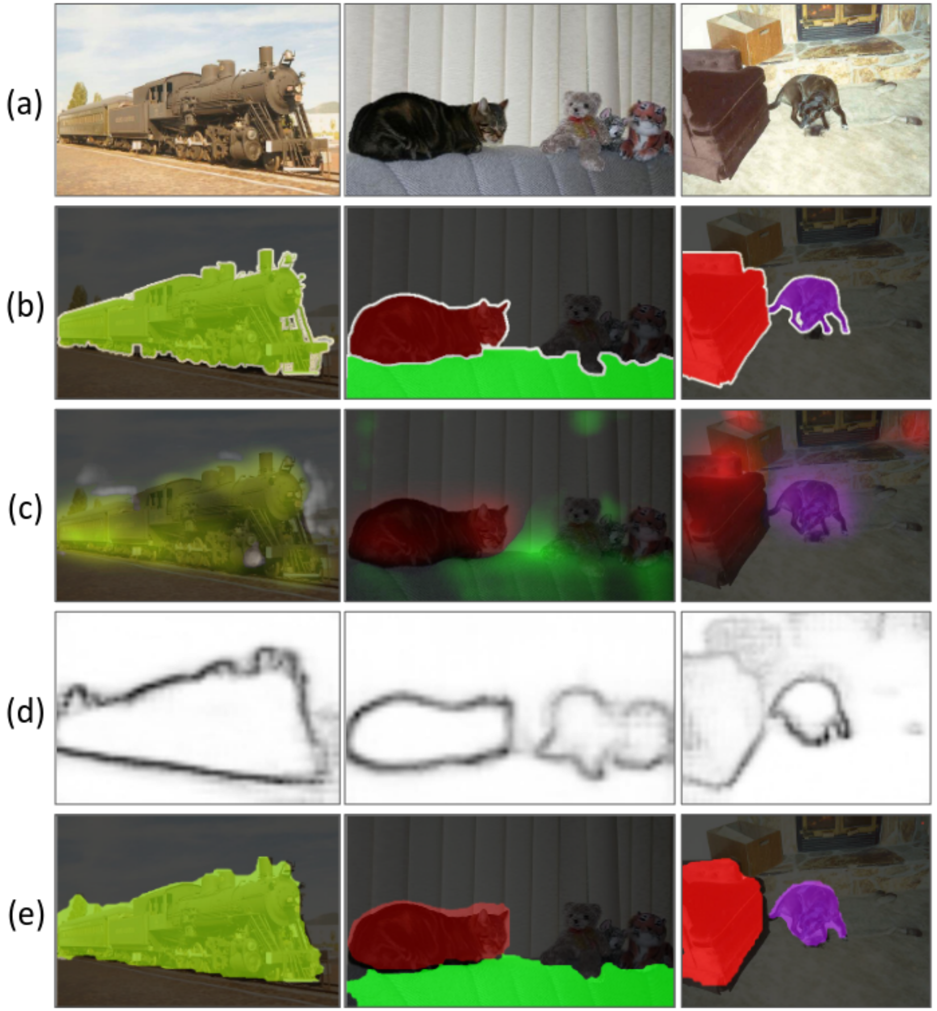


Figure 28: Qualitative results of random walk using Affinity Network.

Source: Ahn and Kwak[1].

[1] J. Ahn, and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4981-4990. 2018.

Puzzle-CAM Better Segmentation Priors

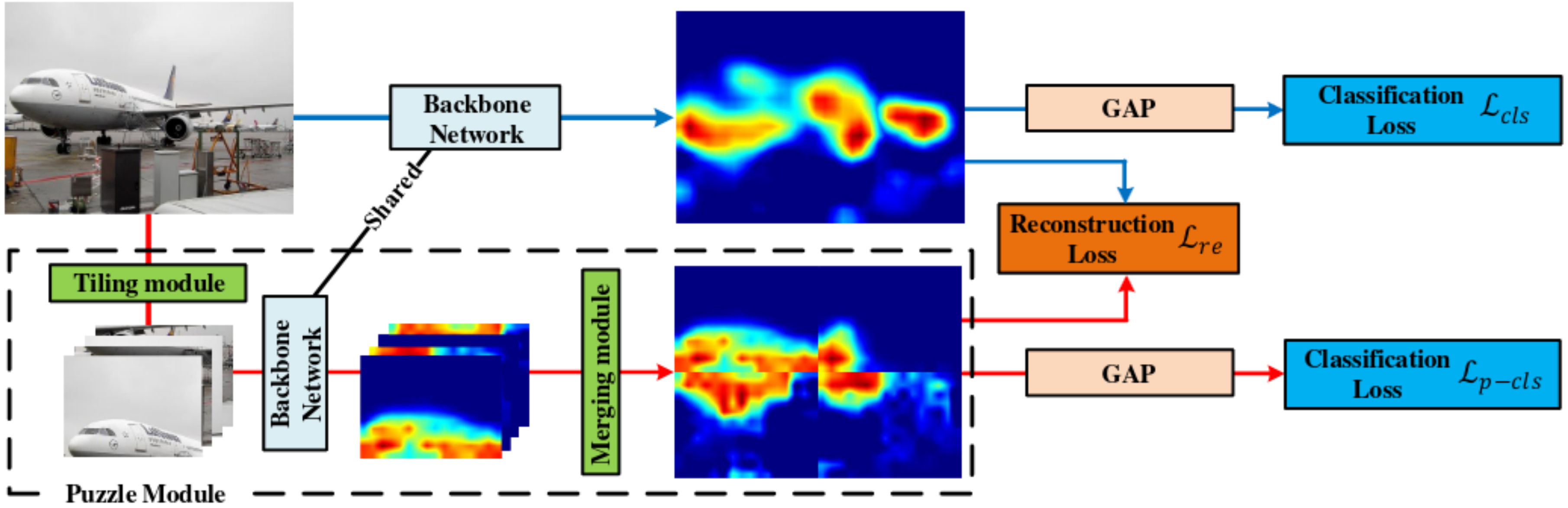


Figure 29: Puzzle-CAM architecture: the input image is forwarded into the model, producing the global stream. Concomitantly, the input is also cut into four "puzzle" pieces and forward separately, which compose the "local" stream when merged. Source: Jo and Yu[1].

[1] S. Jo, and I. Yu. Puzzle-CAM: Improved localization via matching partial and full features. In *IEEE International Conference on Image Processing (ICIP)*, pp. 639-643. IEEE, 2021.

OC-CSE Better Segmentation Priors

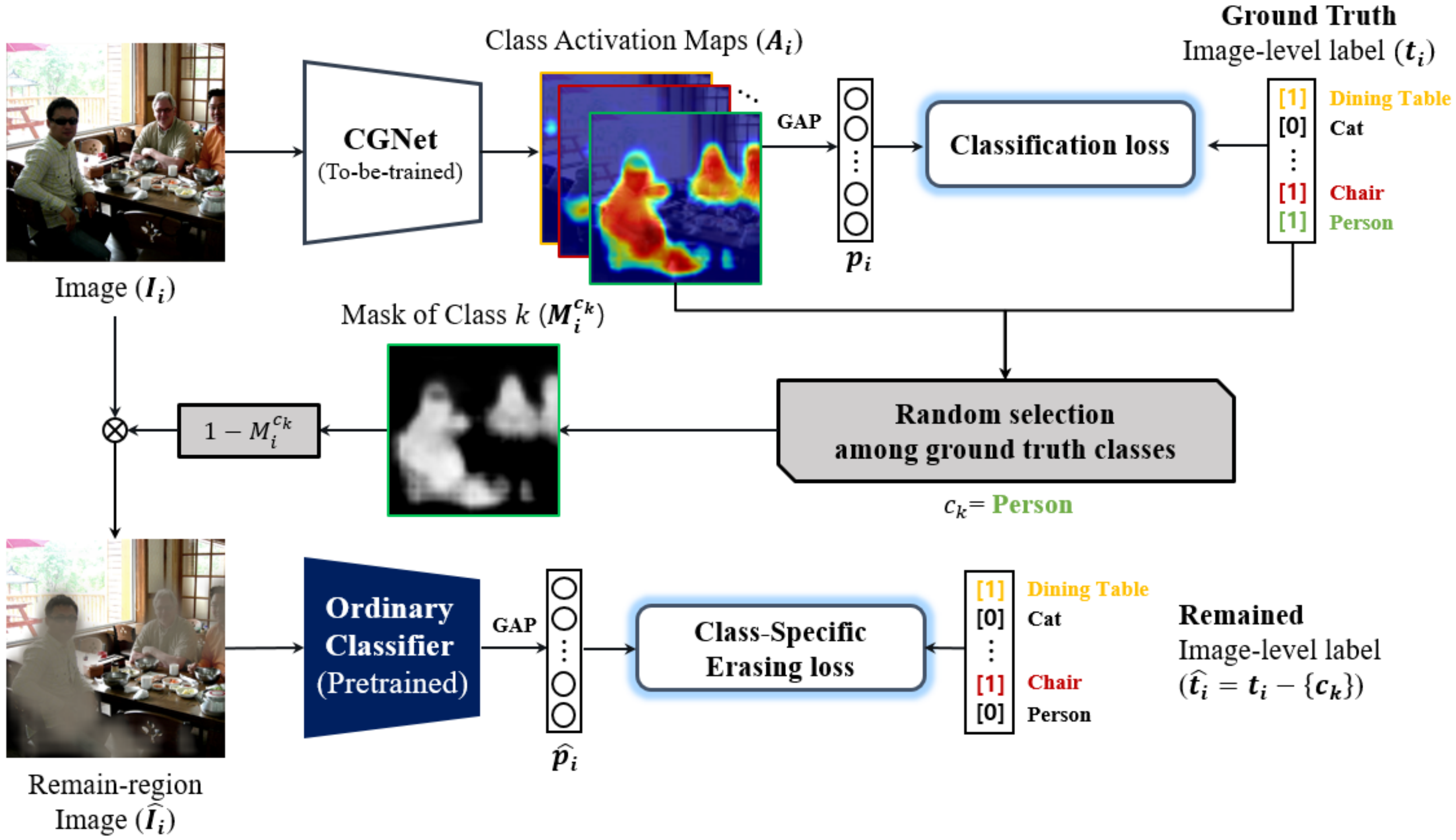
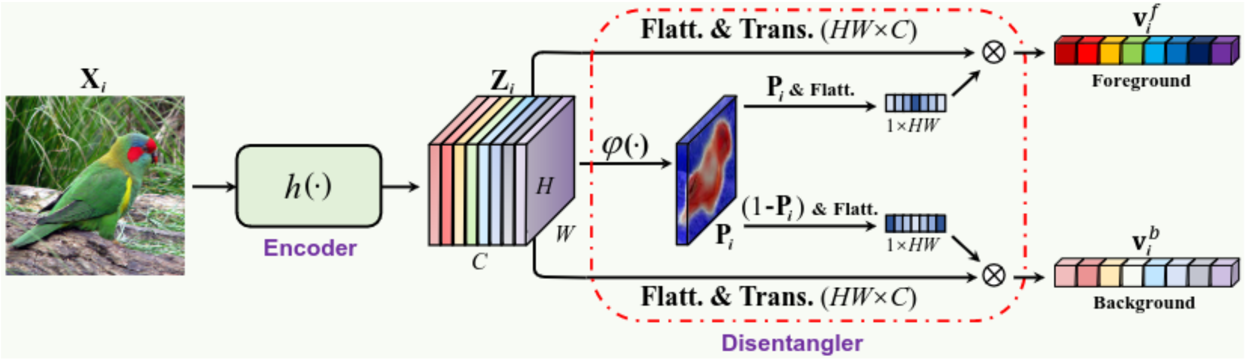
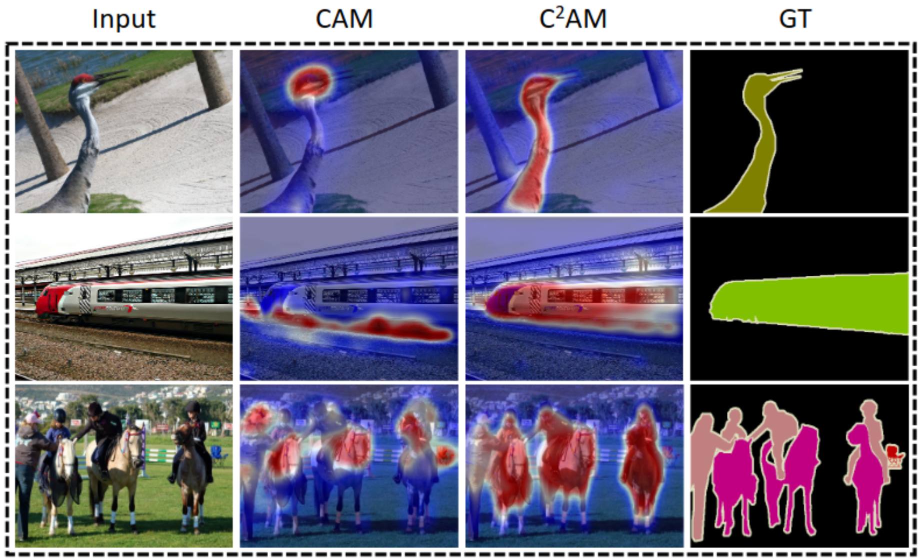


Figure 30: OC-CSE architecture: the input image is forwarded into the CGNet, producing a mask for a random class k. The mask is then used to erase objects of k in the image and fed to a OC (fixed) model. Weights are adjusted so the mask provides a comprehensive erasure of the objects. Source: Jo and Yu[1].

[1] H. Kweon, S. H. Yoon, H. Kim, D. Park, and K. J. Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6994-7003. 2021.

C^2AM Better Segmentation Priors





Training

Refined

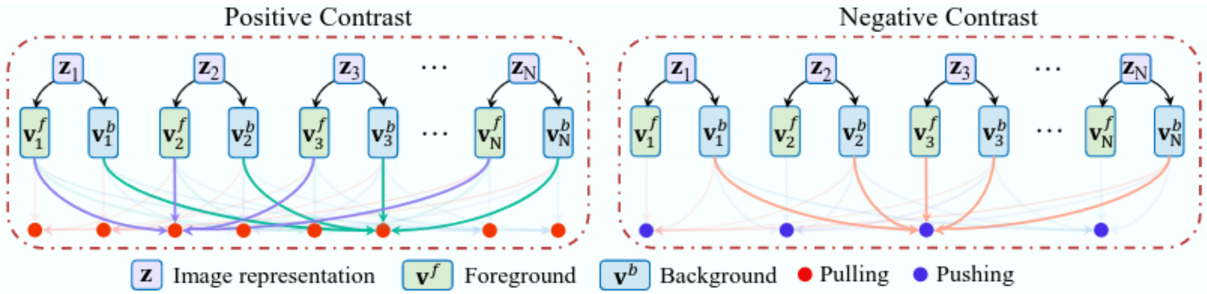
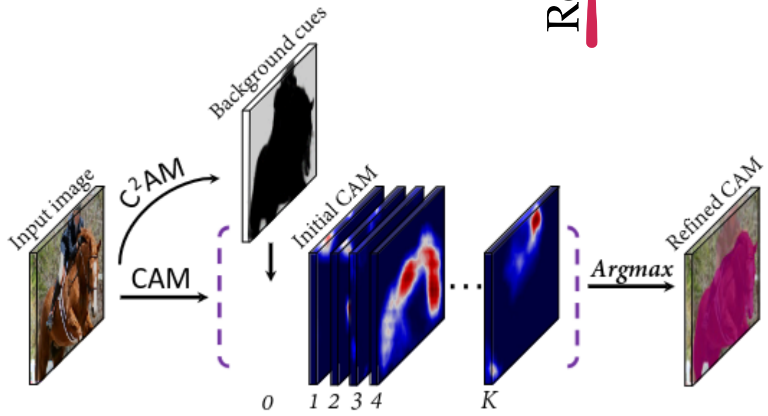


Figure 31: C^2AM processing pipeline. Source: Xie et al.[1]

Inference



[1] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen. Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. arXiv preprint arXiv:2203.13505. 2022.

Schedule

Schedule

Motivation Research Proposal

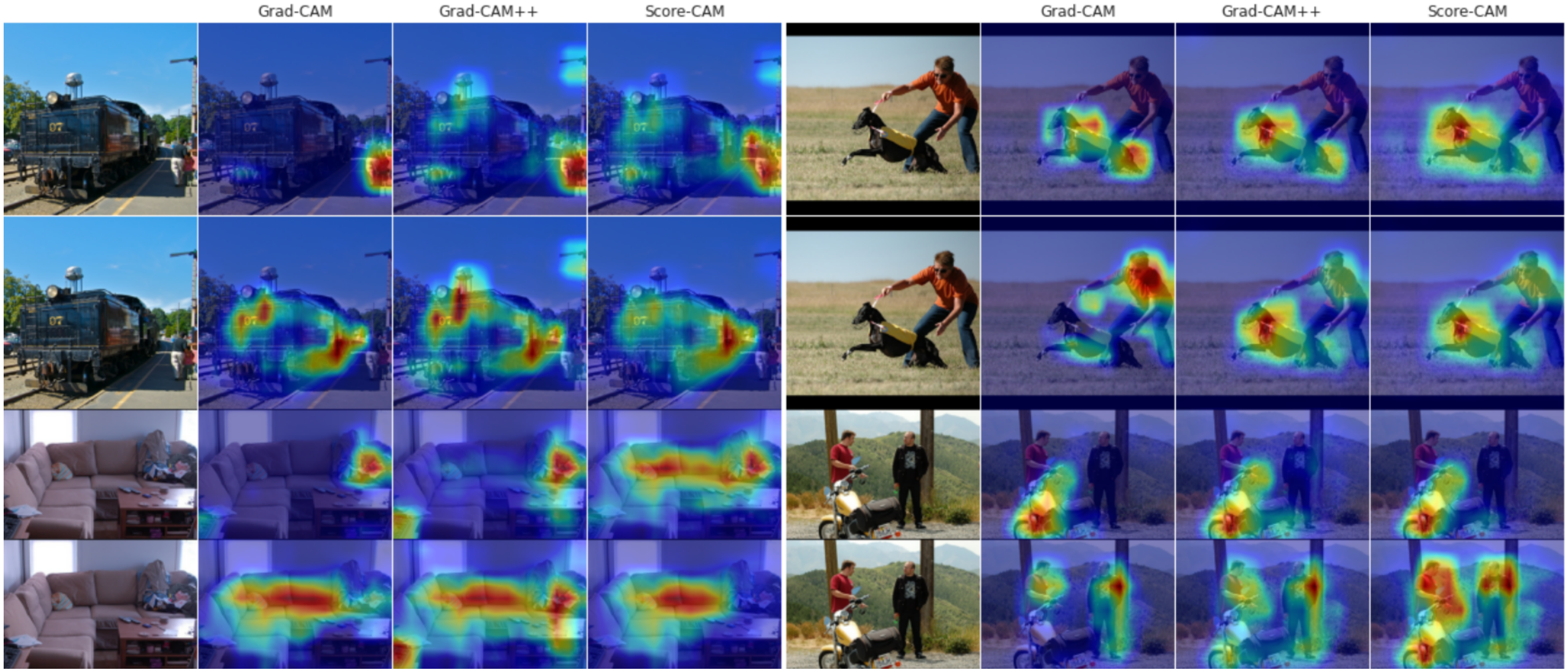


Figure 32: Examples of sensitivity maps obtained from Grad-CAM, Grad-CAM++ and Score-CAM over samples in the Pascal VOC 2007 dataset. Predictions being explained are: *person*, *train*, *person*, *sofa*, *dog*, *person*, *motorcycle*, and *person*. Source: David et al.[1]

[1] L. David., H. Pedrini., and Z. Dias. MinMax-CAM: Improving focus of CAM-based visualization techniques in multi-label problems. In 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, pages 106–117. INSTICC, SciTePress, 2022.

Motivation Research Proposal

Figure 33: Semantic Segmentation priors produced by a ResNet38d model trained with OC-CSE. CAMs were generated using Grad-CAM and Test-Time Augmentation (TTA). Source: keras-explainable/wsol.

[1] W. Sun, J. Zhang, Z. Liu, Y. Zhong, N. Barnes. GETAM: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.02841*. 2021 Dec 6.

Motivation Research Proposal

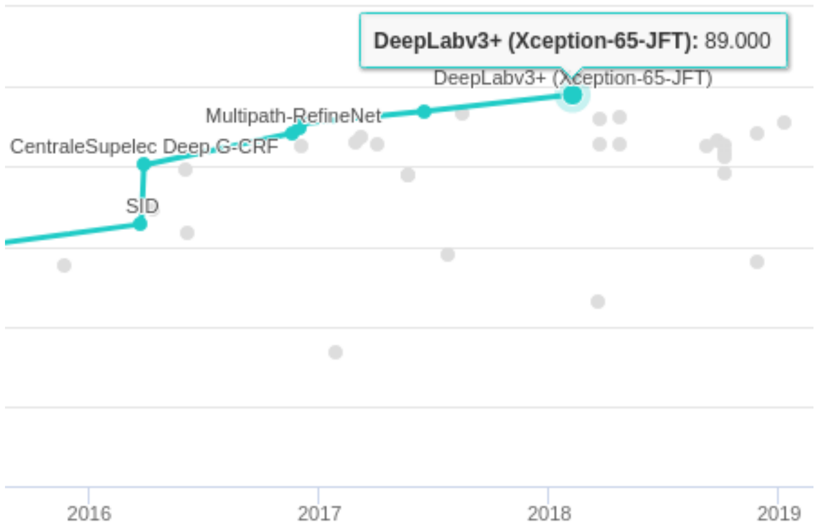


Figure 34: mIoU measured over Pascal VOC 2012 testing dataset. Source: https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012.

	Method	Backbone	Sup.	val	test
Multi-stage	SEAM [66] (CVPR2020)	ResNet38	I	64.5	65.7
	SC-CAM [8] (CVPR2020)	ResNet101	I	66.1	65.9
	CONTA [75] (NeurIPS2020)	ResNet38	I	66.1	66.7
	CDA [56] (ICCV2021)	ResNet101	I	66.1	66.8
	MCS [55] (ECCV2020)	ResNet101	I+S	66.2	66.9
	ECS-Net [56] (ICCV2021)	ResNet38	I+S	66.6	67.6
	EME [20] (ECCV2020)	ResNet101	I+S	67.2	66.7
	ICD [19] (CVPR2020)	ResNet101	I+S	67.8	68.0
	CPN [76] (ICCV2021)	ResNet101	I	67.8	68.5
	CGNet [32] (ICCV2021)	ResNet38	I	68.4	68.2
	AuxSegNet [70] (ICCV2021)	ResNet101	I+S	69.0	68.6
	PMM [39] (ICCV2021)	ResNet101	I	70.0	70.5
	RIB [33](NeurIPS2021)	ResNet101	I+S	70.2	70.0
	NSRM [71] (CVPR2021)	ResNet101	I+S	70.4	70.2
	DRS [30] (AAAI2021)	ResNet101	I	70.4	70.7
	VWL-L [51] (IJCV2022)	ResNet101	I	70.6	70.7
	EDAM [69] (CVPR2021)	ResNet101	I+S	70.9	70.6
	EPS [37](CVPR2021)	ResNet101	I+S	71.0	71.8
	URN [38] (AAAI2022)	ResNet101	I	71.2	71.5
Single-stage	EM [47] (ICCV2015)	VGG16	I	38.2	39.6
	TransferNet [25] (CVPR2016)	VGG16	I+COCO	52.1	51.2
	CRF-RNN [50] (CVPR2017)	VGG16	I	52.8	53.7
	RRM [74] (AAAI2020)	ResNet38	I	62.6	62.9
	1-stage-wseg [3] (CVPR2020)	ResNet38	I	62.7	64.3
	JointSaliency [73] (ICCV2019)	DenseNet169	I+S	63.3	64.3
	AALR [78] (ACMMM2021)	ResNet38	I	63.9	64.8
	GETAM(ours)	ViT-Hybrid	I+S	**71.7**	**72.3**

Table 5. Comparison with the state-of-the-art methods on PASCAL VOC 2012 *val* and *test* sets. Different supervision is used: I: image-level label. COCO: MS-COCO [41], S: saliency. Source: Sun et al.[1]

[1] W. Sun, J. Zhang, Z. Liu, Y. Zhong, N. Barnes. GETAM: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.02841*. 2021 Dec 6.

Schedule

Proposed Approach Research Proposal

1. Exploration of Explainable AI Methods in Multi-Label Problems

1. How do Explainable AI methods behave in multi-label scenarios?
2. Can cross-contributions be erased from the CAMs produced by Grad-CAM?

Proposed Approach Research Proposal

2. Complementary Regularization Strategies in WSSS

- Can complementary strategies be conjointly employed to improve WSSS?
- Is adversarial CAM generation beneficial to WSSS solutions?
- Can context-decoupling help WSSS methods to segment cluttered scenes?

Proposed Approach Research Proposal

3. Exploration of Transformers and Spatial Attention for Highly-Detailed Segmentation

- Can Visual Transformers improve fine-grain WSSS?
- Can WSSS methods be adapted to Vision Transformers?

Proposed Approach Research Proposal

4. Weak Supervision in Boundary and Difficult Scenarios: Class Unbalance, Long-tail and Functional Segmentation

- Can long-tail learning improve WSSS in boundary cases?
- Which features can be drawn from functional segmentation problems to replace visual similarity, a fundamental aspect of WSSS methods?

Proposed Approach Research Proposal

5. Ensemble of Weakly Supervised Semantic Segmentation Systems

- Can WSSS ensembles improve noisy segmentation priors?
- Is contextual information useful when combining predictions?
- Which tasks share mutual information with Semantic Segmentation?
  - Saliency Detection
  - Edge Detection
  - Instance Segmentation

Work Schedule Research Proposal

Activities	1st year				2nd year				3rd year				4th year			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Class attendance and completion of required credits	•	•	•	•												
Exploration of XAI methods in multi-label scenarios	•	•	•	•												
Adversarial and complementary strategies in WSSS					•	•	•	•								
Doctoral Qualifying Exam (EQE)									•							
Participation in "Programa de Estágio Docente" (PED)									•	•						
Exploration of Transformers and Spatial Attention									•	•	•					
Boundary and difficult scenarios											•	•	•	•		
Ensemble of solutions for WSSS														•	•	•
Writing and presentation of Doctoral thesis																•

Schedule

Experimental Setup Research Proposal

Environment

Google Colab

- NVIDIA Tesla K80

SDumont Supercomputer:

- 4x NVIDIA Volta V100 (training)
- 2x NVIDIA K40 (inference)

Tools

- Tensorflow and PyTorch

Experimental Setup Research Proposal

Metrics

XAI

1. Increase in Confidence
2. Average Drop %
3. Average Drop of Others %
4. Average Retention %
5. Average Retention of Others %

Proposed by us.

WSSS

1. mean Intersection over Union (mIoU)
2. Pixel Accuracy
3. F1 Score

Schedule

1. Introduction
2. Related Work
3. Research Proposal
4. **Preliminary Results**
5. Final Considerations

Schedule

MinMax-CAM Contributions for Explainable AI

$$L^c_{\text{CAM}}(f,x) = \sum_k w^c_k A^k$$

$$L^c_{\text{Grad-CAM}}(f,x) = \sum_k \sum_{ij} \frac{\partial f_c(x)}{\partial A^k_{ij}} A^k$$

MinMax-CAM Contributions for Explainable AI

ReLU and GAP omitted for conciseness

$$L^c_{\text{CAM}}(f, x) = \sum_k w^c_k A^k$$

$$L^c_{\text{Grad-CAM}}(f, x) = \sum_k \sum_{ij} \frac{\partial f_c(x)}{\partial A^k_{ij}} A^k$$

Contribution towards the classification of class *c*.

MinMax-CAM Contributions for Explainable AI

ReLU and GAP omitted for conciseness

$$L^c_{\text{CAM}}(f,x) = \sum_k w^c_k A^k$$

$$L^c_{\text{Grad-CAM}}(f,x) = \sum_k \sum_{ij} \frac{\partial f_c(x)}{\partial A^k_{ij}} A^k$$

Contribution towards the classification of class *c*.

$$J_c = S_c - \frac{1}{|N_x|} \sum_{n \in N_x} S_n$$

Regions that contribute t.t.c. of *c*, and do not contribute t.t.c. of the adjacent classes.

MinMax-CAM Contributions for Explainable AI

ReLU and GAP omitted for conciseness

$$L^c_{\text{CAM}}(f,x) = \sum_k w^c_k A^k$$

$$L^c_{\text{Grad-CAM}}(f,x) = \sum_k \sum_{ij} \frac{\partial f_c(x)}{\partial A^k_{ij}} A^k$$

Contribution towards the classification of class *c*.

$$J_c = S_c - \frac{1}{|N_x|} \sum_{n \in N_x} S_n$$

Regions that contribute t.t.c. of *c*, and do not contribute t.t.c. of the adjacent classes.

$$L^c_{\text{MinMax-Grad-CAM}}(f,x) = \sum_k \sum_{ij} \frac{\partial J_c}{\partial A^k_{ij}} A^k$$

$$L^c_{\text{MinMax-CAM}}(f,x) = \sum_k \left[w^c_k - \frac{1}{|N_x|} \sum_{n \in N_x} w^n_k\right]$$

MinMax-CAM Contributions for Explainable AI

$$L^c_{\text{D-MinMax-Grad-CAM}}(f, x) = \mathrm{ReLU}\Big(\textstyle\sum_k \alpha^c_k A^k \Big)$$

$$\alpha^c_k = \textstyle\sum_{ij} \left[\mathrm{ReLU}\left(\frac{\partial S_c}{\partial A^k_{ij}} \right) - \frac{1}{|N_x|} \mathrm{ReLU}\left(\sum_{n \in N_x} \frac{\partial S_n}{\partial A^k_{ij}} \right) + \frac{1}{|C_x|} \min\left(0, \sum_{n \in C_x} \frac{\partial S_n}{\partial A^k_{ij}} \right) \right]$$

MinMax-CAM Contributions for Explainable AI

$$L^c_{\text{D-MinMax-Grad-CAM}}(f, x) = \mathrm{ReLU}\Big(\sum_k \alpha^c_k A^k \Big)$$

$$\alpha^c_k = \sum_{ij} \left[\underbrace{\mathrm{ReLU}\Big(\frac{\partial S_c}{\partial A^k_{ij}} \Big)} - \frac{1}{|N_x|} \mathrm{ReLU}\Big(\sum_{n \in N_x} \frac{\partial S_n}{\partial A^k_{ij}} \Big) + \frac{1}{|C_x|} \min \Big(0, \sum_{n \in C_x} \frac{\partial S_n}{\partial A^k_{ij}} \Big) \right]$$

↑ Positive contributions t.t.c. of *c*

MinMax-CAM Contributions for Explainable AI

$$L^c_{\text{D-MinMax-Grad-CAM}}(f, x) = \mathbf{ReLU}\left(\sum_k \alpha^c_k A^k \right)$$

$$\alpha^c_k = \sum_{ij} \left[\underbrace{\mathbf{ReLU}\left(\frac{\partial S_c}{\partial A^k_{ij}} \right)} - \underbrace{\frac{1}{|N_x|} \mathbf{ReLU}\left(\sum_{n \in N_x} \frac{\partial S_n}{\partial A^k_{ij}} \right)} + \frac{1}{|C_x|} \mathbf{min}\left(0, \sum_{n \in C_x} \frac{\partial S_n}{\partial A^k_{ij}} \right) \right]$$

↑ Positive contributions t.t.c. of *c*

↓ Positive contributions t.t.c. of *n*

MinMax-CAM Contributions for Explainable AI

$$L^c_{\text{D-MinMax-Grad-CAM}}(f, x) = \mathrm{ReLU}\Big(\sum_k \alpha^c_k A^k\Big)$$

$$\alpha^c_k = \sum_{ij}\left[\underbrace{\mathrm{ReLU}\left(\frac{\partial S_c}{\partial A^k_{ij}}\right)} - \underbrace{\frac{1}{|N_x|}\mathrm{ReLU}\left(\sum_{n \in N_x}\frac{\partial S_n}{\partial A^k_{ij}}\right)} + \underbrace{\frac{1}{|C_x|}\min\left(0, \sum_{n \in C_x}\frac{\partial S_n}{\partial A^k_{ij}}\right)}\right]$$

↑ Positive contributions t.t.c. of c

↓ Positive contributions t.t.c. of n

↓ Negative contributions t.t.c. of all.

Qualitative Results over VOC MinMax-CAM

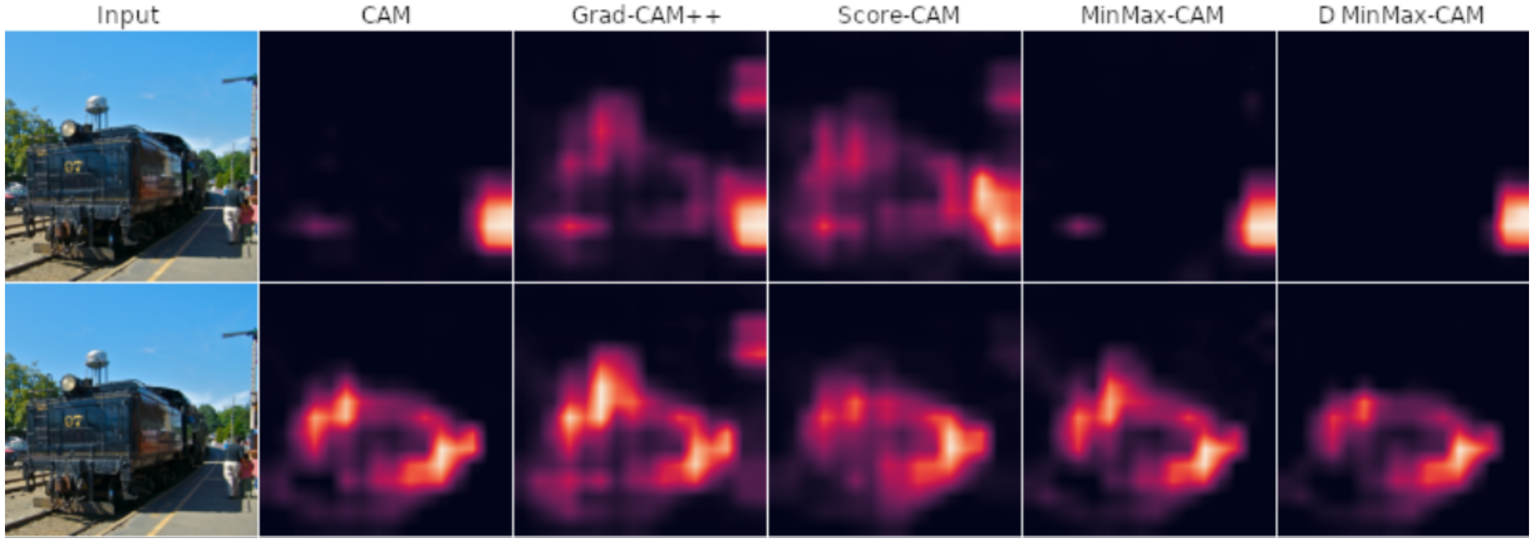


Figure 35: Comparison of CAMs obtained from various XAI methods. Predictions being explained are: *person*, *train*, *motorcycle*, *person*, *chair*, and *table*. Source: David et al.[1]

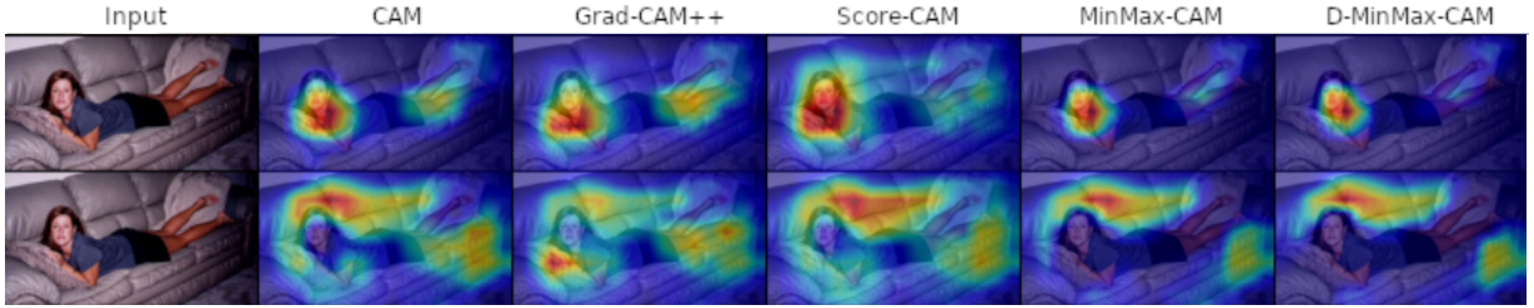


Figure 36: Comparison of sensitivity maps from various XAI methods. Source: David et al.[1]

Qualitative Results over COCO 2017 MinMax-CAM

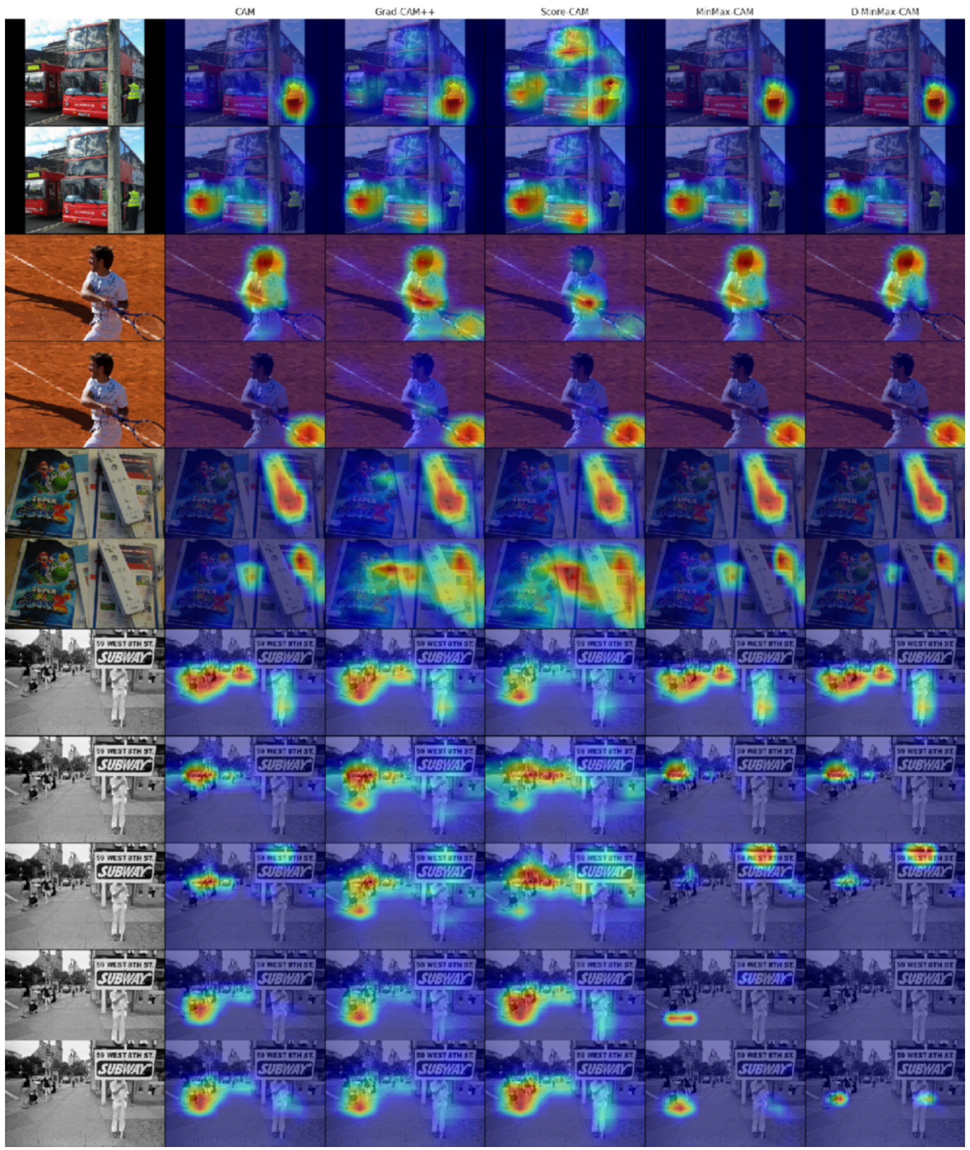


Figure 37: Comparison of sensitivity maps obtained from various XAI methods over the MS COCO 2017 dataset. Source: David et al.[1]

Qualitative Results over HPA MinMax-CAM

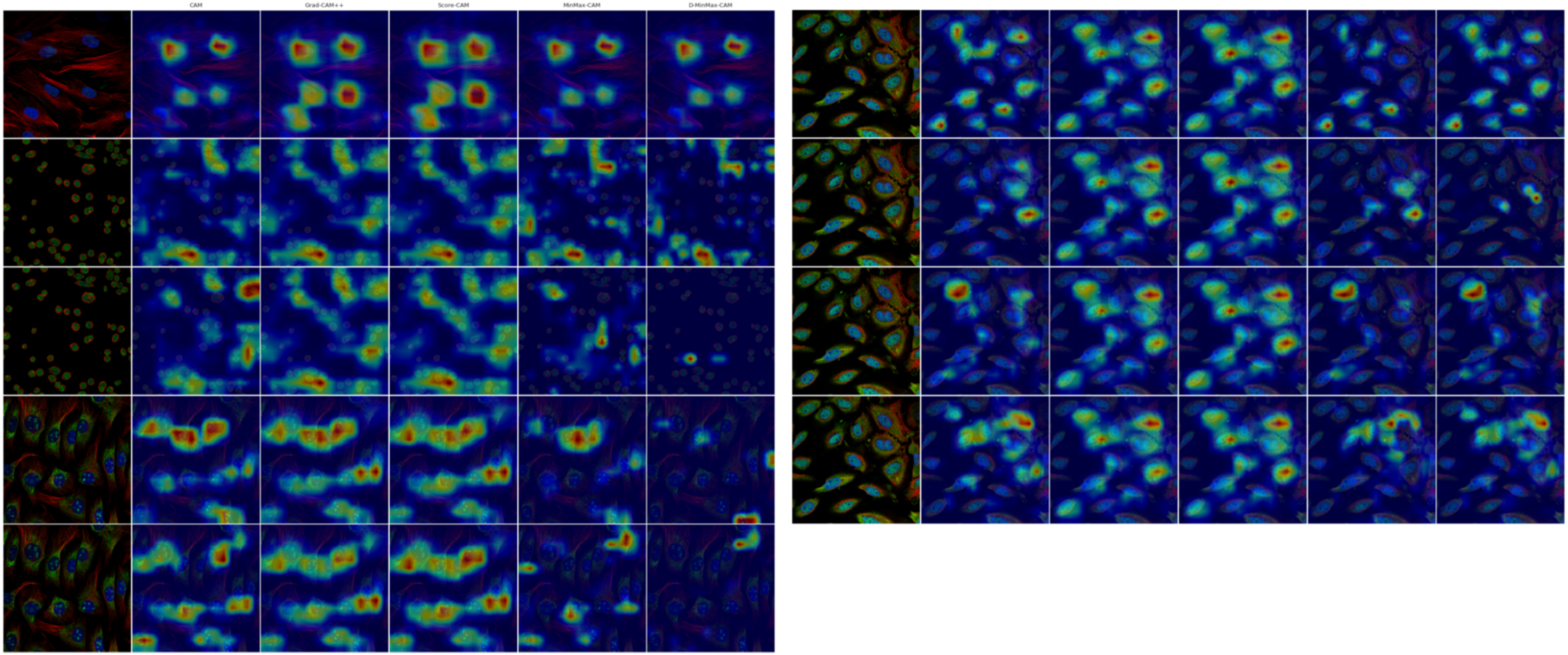


Figure 38: Comparison of sensitivity maps obtained from various XAI methods over the Human Protein Atlas Image Classification dataset. Source: David et al.[1]

Quantitative Results MinMax-CAM

Metric	Dataset	CAM	Grad-CAM++	Score-CAM	MinMax-CAM	D-MinMax-CAM
%IC	P:UAS	6.09%	7.05%	**11.59%**	6.22%	6.27%
	COCO17	30.21%	32.98%	**44.69%**	23.12%	19.20%
	VOC07	27.68%	31.03%	**40.76%**	26.61%	23.83%
	VOC12	27.75%	25.40%	**35.10%**	24.70%	21.66%
	HPA	8.64%	9.29%	**11.27%**	7.63%	5.89%
%AD	P:UAS	55.25%	49.00%	**43.37%**	64.24%	66.88%
	COCO17	27.42%	17.56%	**9.62%**	40.22%	47.43%
	VOC07	25.24%	17.90%	**10.79%**	32.58%	39.25%
	VOC12	24.47%	18.69%	**10.60%**	29.17%	34.22%
	HPA	49.78%	47.02%	**41.50%**	54.16%	64.21%
%ADO	P:UAS	43.61%	33.67%	34.06%	60.04%	**60.62%**
	COCO17	51.49%	20.59%	24.45%	68.04%	**71.90%**
	VOC07	32.73%	12.48%	14.72%	44.03%	**46.49%**
	VOC12	36.44%	14.92%	18.46%	43.65%	**45.02%**
	HPA	24.01%	18.95%	17.07%	29.46%	**39.50%**
%AR	P:UAS	46.42%	**49.45%**	48.01%	37.16%	32.74%
	COCO17	**27.70%**	25.60%	26.64%	24.44%	22.79%
	VOC07	**16.54%**	14.04%	14.94%	14.27%	12.00%
	VOC12	**16.23%**	14.71%	16.22%	14.60%	13.06%
	HPA	29.15%	28.49%	**30.59%**	25.60%	15.44%
%ARO	P:UAS	25.48%	29.46%	28.13%	20.84%	**18.55%**
	COCO17	5.26%	7.92%	7.71%	3.31%	**3.13%**
	VOC07	2.44%	3.94%	3.43%	1.28%	**1.16%**
	VOC12	2.29%	3.76%	3.32%	1.21%	**1.14%**
	HPA	6.69%	9.32%	10.56%	3.60%	**1.32%**
F_1-	P:UAS	30.68%	32.07%	28.46%	28.35%	**26.42%**
	COCO17	8.23%	9.94%	7.39%	5.82%	**5.64%**
	VOC07	4.05%	5.62%	**2.20%**	2.38%	2.21%
	VOC12	3.89%	5.70%	4.30%	2.26%	**2.17%**
	HPA	10.89%	14.26%	15.10%	6.45%	**2.54%**
F_1+	P:UAS	39.54%	35.11%	35.41%	**41.00%**	37.01%
	COCO17	34.05%	21.45%	23.82%	**34.07%**	32.44%
	VOC07	**20.84%**	11.97%	6.89%	19.85%	17.13%
	VOC12	**21.25%**	13.87%	16.39%	20.25%	18.60%
	HPA	**22.85%**	18.30%	18.29%	22.71%	18.79%

Table 2: Report of metric scores over multiple datasets.

Kernel Usage Regularization Contributions for Explainable AI

$$g = [g^k]_K = \mathrm{GAP}_{hw}(A^k)$$
$$W = [w_k^c]_{K\times C}$$
$$b = [b_c]_C$$

$$W_\alpha^r = W \circ \alpha \mathrm{softmax}(W)$$
$$y = \sigma(g \cdot W_\alpha^r + b)$$

Kernel Usage Regularization Contributions for Explainable AI

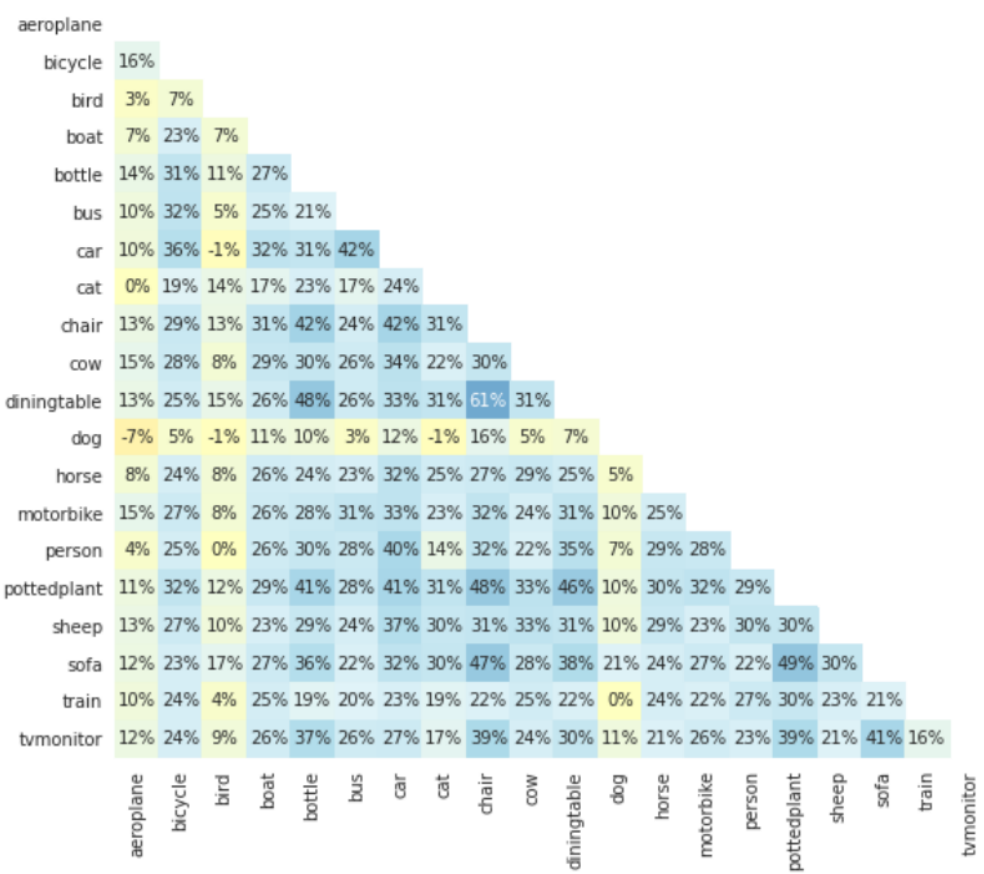


Figure 39: Correlation between different weight vectors in a vanilla (unregularized) *sigmoid* FC layer. Source: David et al.[1]

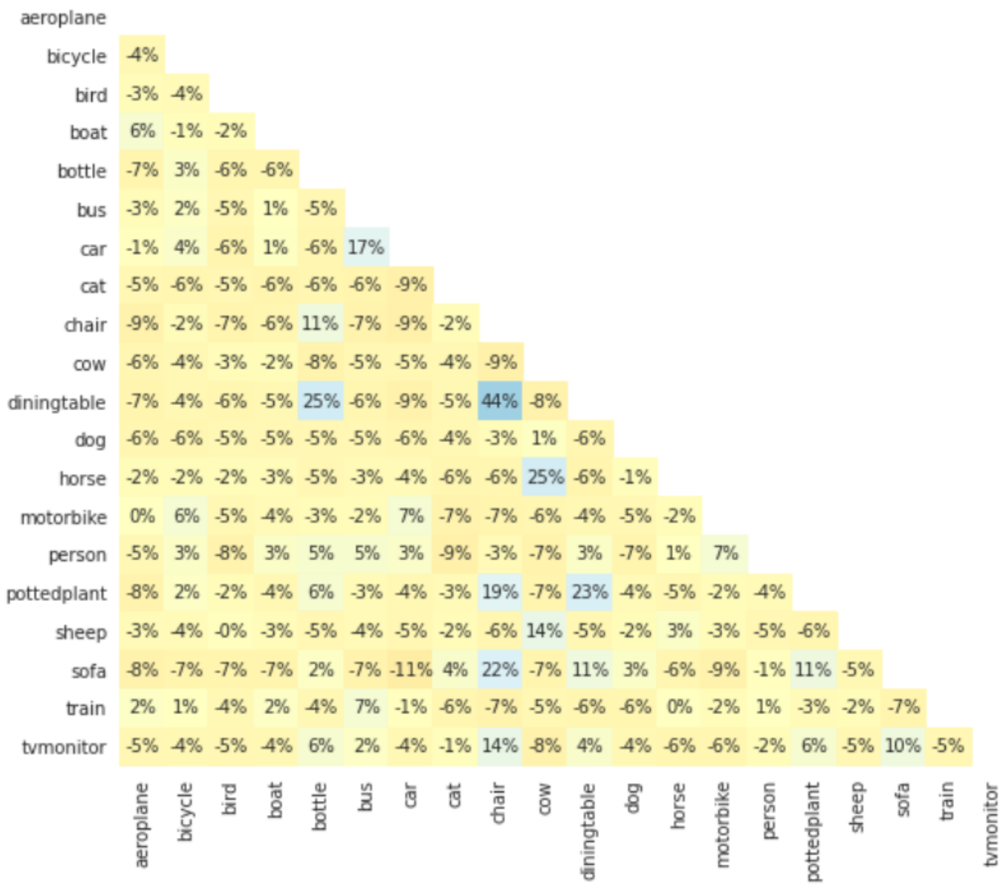


Figure 40: Correlation between different weight vectors in a *sigmoid* FC layer trained with Kernel Usage Regularization. Source: David et al.[1]

Kernel Usage Regularization Contributions for Explainable AI

Metric	Dataset	Baseline	KUR
F_1	VOC07 Test	84.26%	**85.85%**
F_1	VOC12 Val	85.05%	**85.90%**
F_2	P:UAS Val	87.80%	**88.24%**
F_2	P:UAS Private Test	89.22%	**89.81%**
F_2	P:UAS Public Test	89.62%	**90.10%**
F_1	COCO17 Val	**75.64%**	74.23%
F_1	HPA Private Test	**36.05%**	35.54%
F_1	HPA Public Test	**39.72%**	39.46%

Table 3: Report of classification scores over multiple datasets, considering a baseline classifier the model trained with Kernel Usage Regularization (KUR).

Schedule

Exploration of Complementary WSSS Strategies

Contributions for WSSS

$$\begin{aligned}
\mathcal{L}_{\text{P-OC}} &= \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{re-cls}} + \mathcal{L}_{\text{re}} + \lambda_{\text{cse}}\mathcal{L}_{\text{cse}} \\
&= \ell_{\text{bce}}(p_i, t_i) + \ell_{\text{bce}}(p_i^{\text{re}}, t_i) \\
&\quad + \lambda_{\text{re}} \| A_i - A_i^{\text{re}} \|_1 + \lambda_{\text{cse}}\ell_{\text{bce}}(\hat{p}_i, \hat{t}_i)
\end{aligned}$$

Exploration of Complementary WSSS Strategies

Contributions for WSSS

Figure 41: Priors obtained by (from left to right): Vanilla (RandAugment), OC-CSE, Puzzle, P-OC.

P-NOC Contributions for WSSS

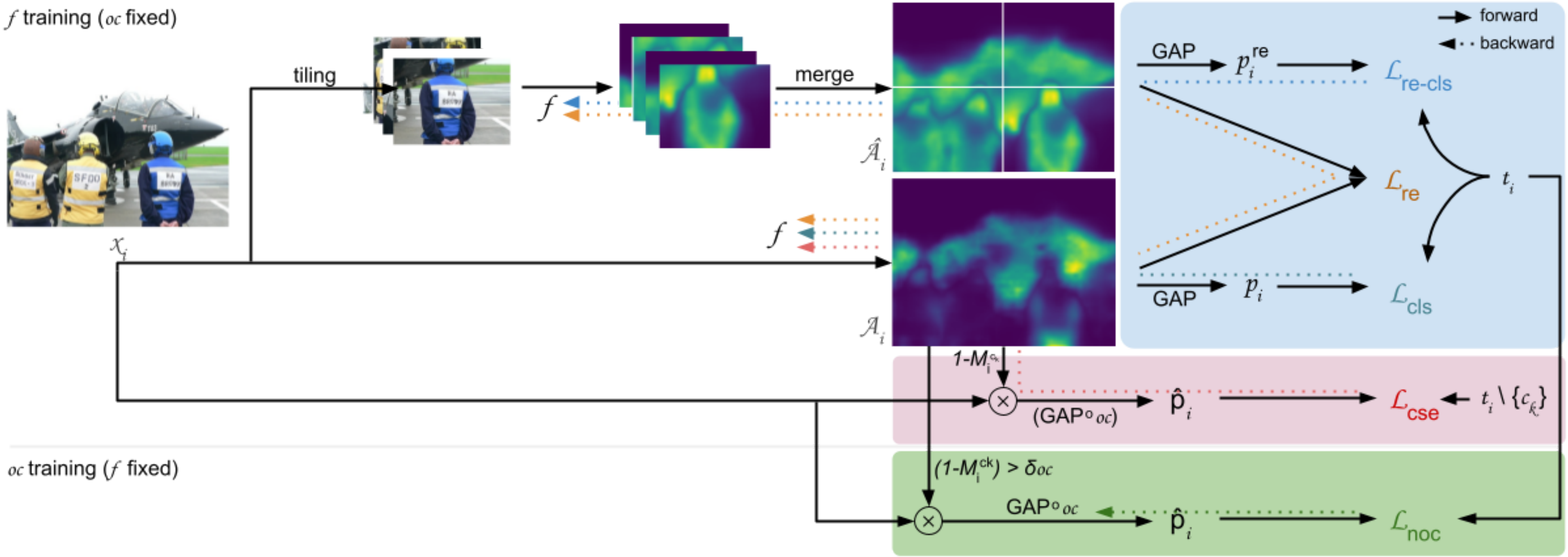


Figure 42: Overview of our adversarial training setup, in which *f* is optimized considering both Puzzle module and the ordinary classifier *oc*. *f* is sub-sequentially fixed and *oc* is updated to shift its attention towards regions currently ignored by *f*.

P-NOC Contributions for WSSS

Algorithm 1 Proposed P-NOC algorithm

Require: Training set $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, CAM networks f, oc; $k_{noc} \in \mathbb{N}$, $\delta_{\text{noc}} \in [0, 1]$

$i \leftarrow 0$
while not done **do**
    Sample a batch (x, y) from $\mathcal{D}$
    // Fix oc and train f
    Compute $A_i^c = f(x_i)$, $\hat{A}_i^c = \text{merge}(f(\text{tile}(x_i)))$
    Compute $\mathcal{L}_{\text{P-OC}}$ loss from Eq. (26)
    Update weights of f by $\nabla \mathcal{L}_{\text{P-OC}}$
    $i \leftarrow i + 1$
    if $i \mod k_{noc} = 0$ **then**
        // Fix f and train oc
        $\hat{x} = x \circ (M < \delta_{\text{noc}})$
        Compute $\mathcal{L}_{noc}$ from Eq. (27)
        Update weights of oc by $\nabla \mathcal{L}_{\text{noc}}$
    end if
end while

$$\mathcal{L}_{\text{P-OC}} = \ell_{\text{bce}}(p_i, t_i) + \ell_{\text{bce}}(p_i^{\text{re}}, t_i) + \lambda_{\text{re}} \| A_i - A_i^{\text{re}} \|_1 + \lambda_{\text{cse}} \ell_{\text{bce}}(\hat{p}_i, \hat{t}_i)$$

$$\mathcal{L}_{\text{noc}} = \lambda_{\text{noc}} \ell_{\text{bce}}(oc(x_i \circ (M_i^{c_k} < \delta_{\text{noc}})), t_i)$$

C^2AM-H Contributions for WSSS

Figure 43: CAMs produced by a network trained with P-OC, when presented with samples from the Pascal VOC 2012 *train* set.

Figure 44: Hints obtained by binarizing the CAMs, using a threshold of 0.4.

$$\mathcal{L}^{\mathcal{B}}_{\mathrm{C^2AM\text{-}H}} = \mathcal{L}^{\mathcal{B}}_{\text{pos-f}} + \mathcal{L}^{\mathcal{B}}_{\text{pos-b}} + \mathcal{L}^{\mathcal{B}}_{\text{neg}} + \lambda_h \sum_{i \in b} \underline{\sum_{h,w} 1_{[A_i^{hw} > \delta_{\text{fg}}]} \ell_{\text{bce}}}(\hat{y}_i^{hw}, p_i^{hw})$$

C²AM-H Contributions for WSSS

Figure 45: Saliency proposals obtained from a PoolNet model, after being trained with C²AM-H pseudo saliency maps.

C^2AM-H Contributions for WSSS

Figure 45: Saliency proposals obtained from a PoolNet model, after being trained with C^2AM-H pseudo saliency maps.

Figure 46: Affinity labels. From left to right: (a) ground-truth maps, (b) coarse priors, (c) priors +dCRF, and (d) priors +C^2AM-H +dCRF.

Ablation Studies Contributions for WSSS

Method	+LS	+C²AM-H	+NOC	*train* (%)	*val* (%)
P				73.74	72.31
P^f				71.35	70.67
P-OC				73.50	72.08
P-OC	✓			71.45	70.15
P-OC		✓		**73.90**	72.53
P-OC	✓	✓		73.07	72.14
P-OC	✓		✓	73.31	72.83
P-OC	✓	✓	✓	73.59	**73.37**

Table 4: Ablation studies of pseudo segmentation masks, measured in mIoU (%) over Pascal VOC 2012 training and validation sets.

(Refined) Pseudo Segmentation Maps P-NOC +C²AM-H

Figure 47: Pseudo segmentation maps obtained by random walking over segmentation priors generated by a model trained with P-NOC proposals. The Affinity Network was trained over labels refined with saliency maps devised from C²AM-H.

Qualitative Results over VOC 2012 P-NOC +C^2AM-H

Figure 48: Qualitative results over Pascal VOC 2012 datasets. Segmentation proposals obtained by a DeepLabV3+ model trained with pseudo labels devised from P-NOC +C^2AM-H.

Quantitative Results over VOC 2012 P-NOC +C²AM-H

Method	Backbone	Val	Test
AffinityNet [3]	Wide-ResNet-38	61.7	63.7
IRNet [2]	ResNet-50	63.5	64.8
ICD [24]	ResNet-101	64.1	64.3
SEAM [80]	Wide-ResNet-38	64.5	65.7
OC-CSE [37]	Wide-ResNet-38	68.4	68.2
Puzzle-CAM [34]	ResNeSt-269	71.9	72.2
RIB [39]	ResNet-101	68.3	68.6
EPS [43]	ResNet-101	70.9	70.8
AMN [42]	ResNet-101	69.5	69.6
ViT-PCM [59]	ViT-B/16	70.3	70.9
MCTformer [86]	Wide-ResNet-38	**71.9**	71.6
P-$OC_{+C^2AM\text{-}H}$ (ours)	ResNeSt-269	71.4	72.4
P-$NOC_{+LS+C^2AM\text{-}H}$ (ours)	ResNeSt-269	71.5	**72.7**

Table 5: Comparison with other methods in literature. mIoU (%) scores are reported for both Pascal VOC 2012 validation and testing sets.

Quantitative Results over COCO 2014 P-NOC +C²AM-H

Method	Backbone	Val
IRNet [Ahn *et al.*, 2019]	ResNet-50	32.6
IRN+CONTA [Zhang *et al.*, 2020]	ResNet-50	33.4
OC-CSE [Kweon *et al.*, 2021]	Wide-ResNet-38	36.4
PPM [Li *et al.*, 2021]	ScaleNet	40.2
RIB [Lee *et al.*, 2021a]	ResNet-101	43.8
EPS† [Lee *et al.*, 2021d]	ResNet-101	35.7
URN [Li *et al.*, 2022]	ResNet-101	40.7
IRN+AMN [Lee *et al.*, 2022]	ResNet-101	44.7
ViT-PCM [Rossetti *et al.*, 2022]	ViT-B/16	45.0
MCTformer [Xu *et al.*, 2022]	Wide-ResNet-38	42.0
P-OC+C²AM-H (ours)‡	ResNeSt-269	39.8
P-NOC+LS+C²AM-H (ours)‡	ResNeSt-269	41.2

Table 5: Comparison with other methods in literature. mIoU (%) scores are reported for MS COCO 2014 validation set. P-NOC and OC-CSE: priors employed, no refinement conducted.

Schedule

1. Introduction
2. Related Work
3. Research Proposal
4. Preliminary Results
5. **Final Considerations**

Final Considerations

We conducted studies over:

- XAI in broader (multi-label) scenarios
  - MinMax-CAM
- Complementary Regularization Strategies in WSSS
  - Adversarial CAM generation for more robust priors

As future work, we propose to:

- Transformers in WSSS
- WSSS in Boundary and Difficult Scenarios
- Ensemble and meta-learning strategies in WSSS

Scientific Production

1. L. David, H. Pedrini, and Z. Dias. MinMax-CAM: Improving focus of CAM-based visualization techniques in multi-label problems. In *17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 106–117. INSTICC, SciTePress, 2022.

2. L. David, H. Pedrini, and Z. Dias. MinMax-CAM: Increasing Precision of Explaining Maps by Contrasting Gradient Signals and Regularizing Kernel Usage (Springer). In *17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), CCIS Series*, 2023.

3. L. David, H. Pedrini, and Z. Dias. Not so Ordinary Classifier: Revisiting Complementary Regularizing Strategies for More Robust Priors in Weakly Supervised Semantic Segmentation.

Technical Contributions Final Considerations

1. Implement pixel ignoring functionality in the cross-entropy loss in Keras, for semantic segmentation problems[2].

2. Ported the Wide ResNet38-d and ResNeSt architectures, originally trained in PyTorch, to TensorFlow.

3. Created the keras-explainable library, containing out-of-the box implementations of many Explainable AI algorithms.

4. Various fixes in Keras and TensorFlow-Addons, often related to the optimizer, mixed-precision when training in a Multi-Worker-Mirrored-Strategy environment.

Acknowledgements

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

Laboratório
Nacional de
Computação
Científica
UNIDADE DE PESQUISA DO MCTI