
# PUZZLE-CAM: IMPROVED LOCALIZATION VIA MATCHING PARTIAL AND FULL FEATURES

*Sanghyun Jo*

GYNetworks

josanghyeokn@gynetworks.com

*In-Jae Yu**

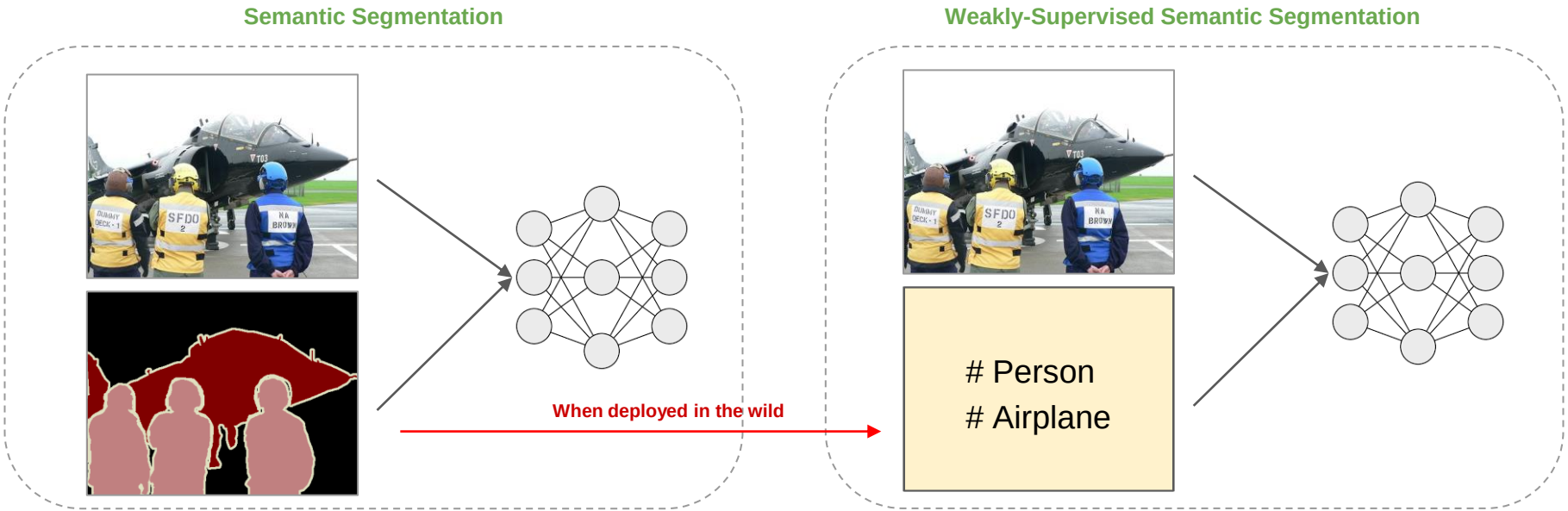
KAIST
School of Computing

myhome98304@gmail.com

*Corresponding Author

# Motivation [1/3]

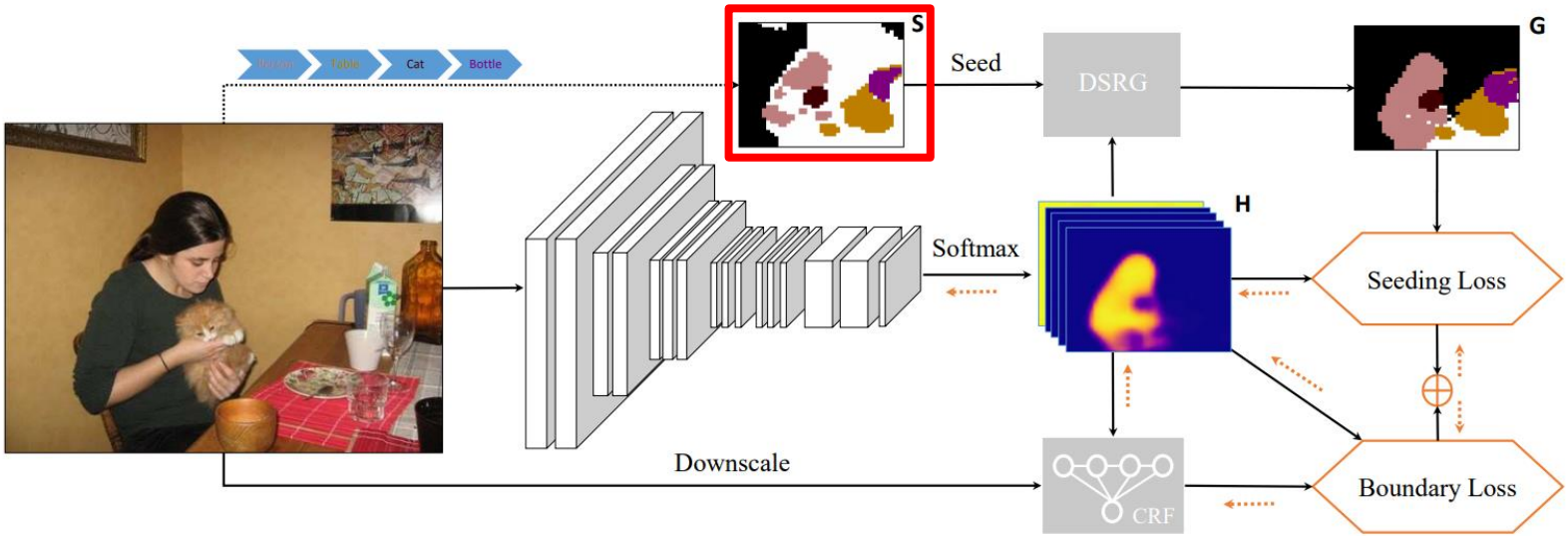
**Producing large-scale dataset to train segmentation model is labor-intensive and time-consuming.**



# Motivation [2/3]

**Most of previous methods are based on the class activation maps (CAMs).**

- Deep Seeded Region Growing (DSRG[1]) takes the seed cues and segmentation map as input produces latent pixel-wise supervision which is more accurate and more complete than seed cues.

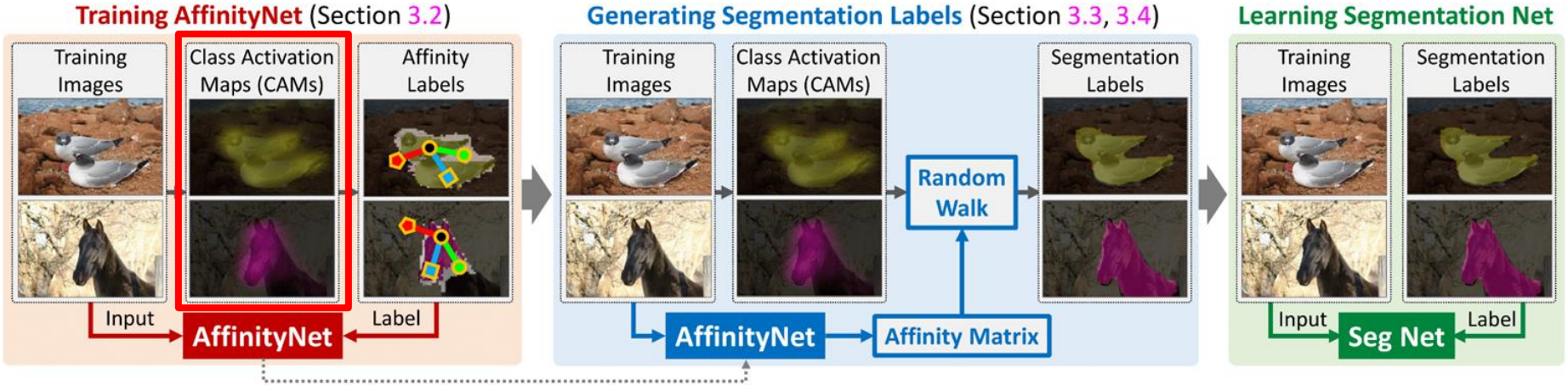



[1] Huang et al., Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing

# Motivation [3/3]

**Most of previous methods are based on the class activation maps (CAMs).**

- The conventional CAMs activates part of objects for classification.
- AffinityNet[1] and IRNet[2] have multiple stages to generate segmentation labels using the conventional CAMs.



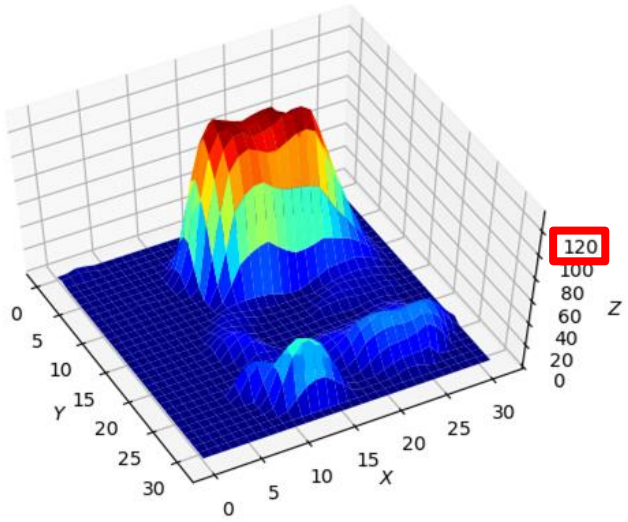

[1] Ahn et al., Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation (CVPR 2018)
[2] Ahn et al., Weakly supervised learning of instance segmentation with inter-pixel relations (CVPR 2019)

# Proposed Method [1/5]

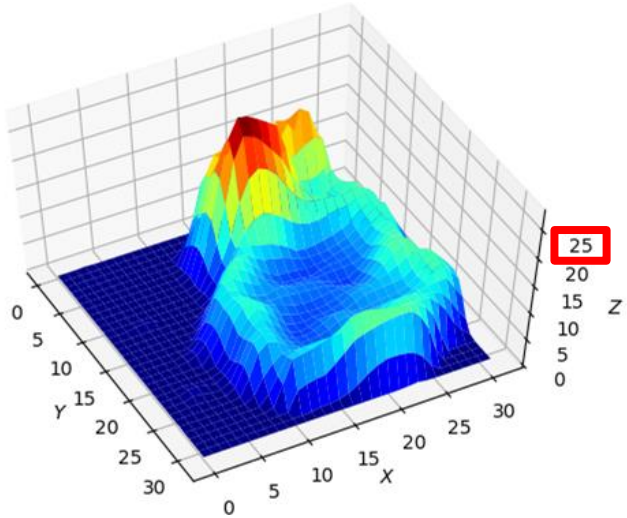
**How to avoid detecting discriminative region of the object?**

- Puzzle-CAM suppresses the attention on discriminative region of the object.

CAM



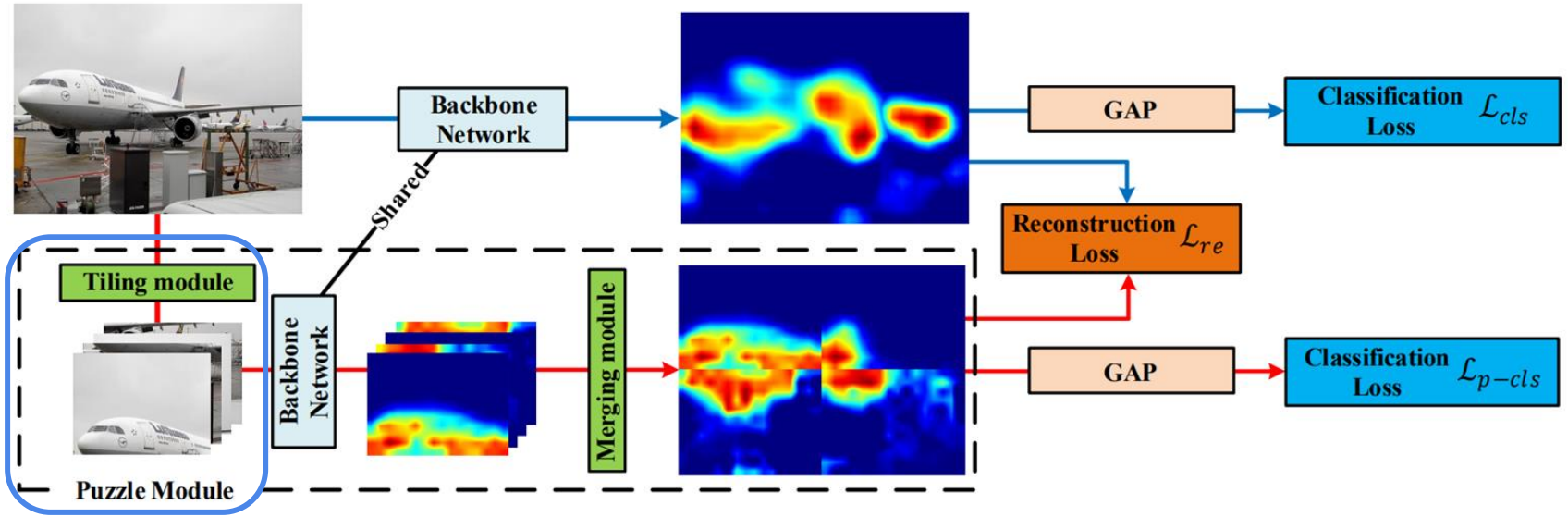

Puzzle-CAM (Ours)



# Proposed Method [2/5]

**How to learn the integral region of the object by using the tags?**

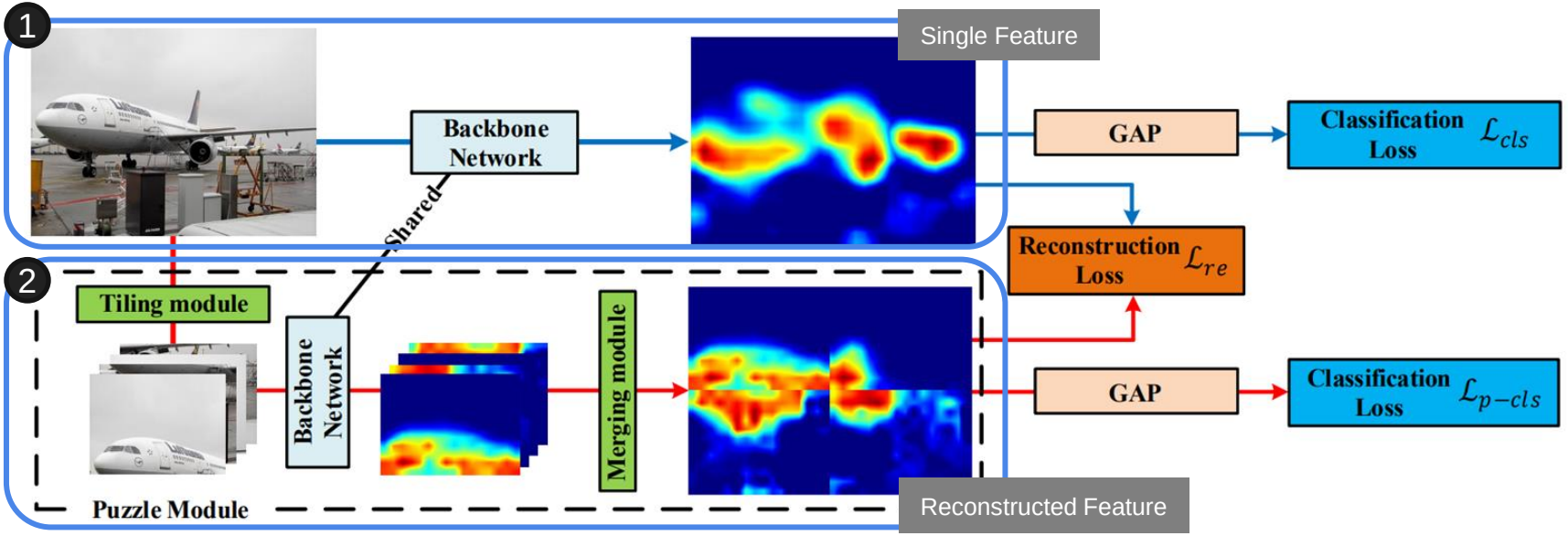
1. Tilling an image to image patches to divide into the attention.
2. Merging the feature maps from the network to produce the reconstructed features.
3. Matching partial and full features with reconstructing regularization.



# Proposed Method [2/5]

**How to learn the integral region of the object by using the tags?**

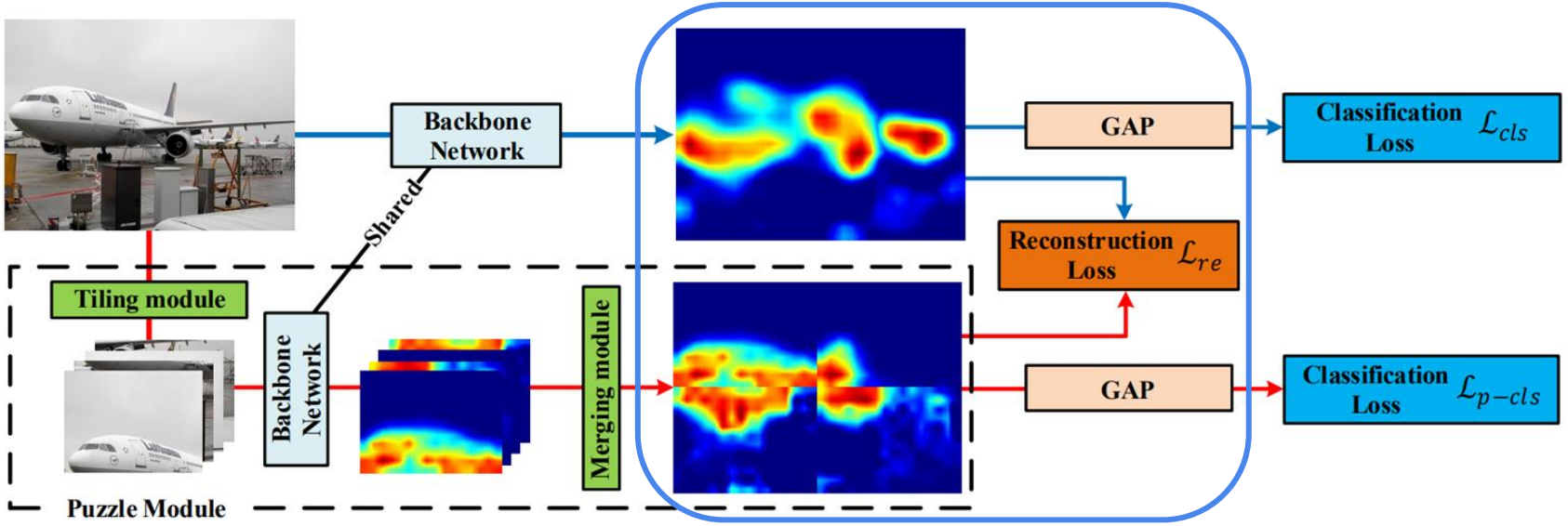
1. Tilling an image to image patches to divide into the attention.
2. Merging the feature maps from the network to produce the reconstructed features.
3. Matching partial and full features with reconstructing regularization.



# Proposed Method [2/5]

**How to learn the integral region of the object by using the tags?**

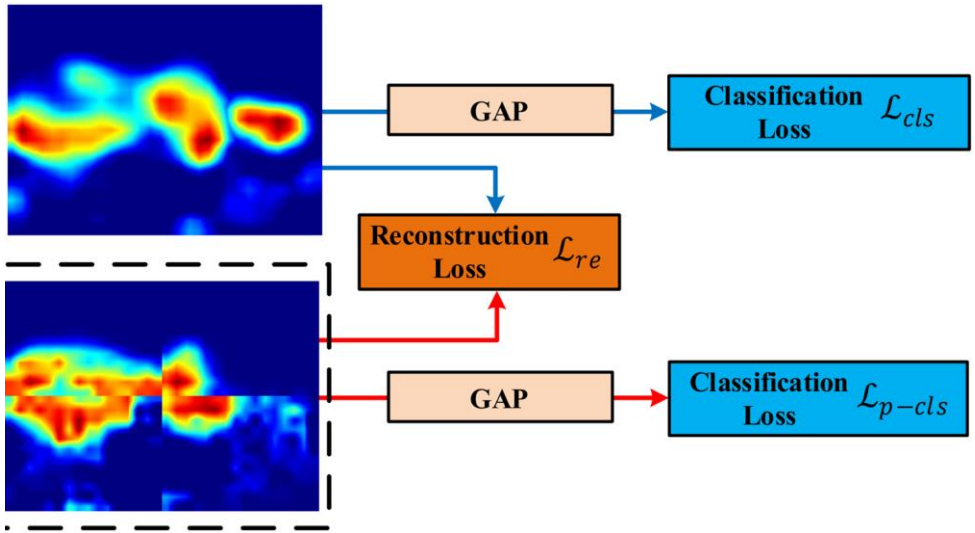
1. Tilling an image to image patches to divide into the attention.
2. Merging the feature maps from the network to produce the reconstructed features.
3. Matching partial and full features with reconstructing regularization.



# Proposed Method [3/5]

**How to narrow the gaps between the single and reconstructed features?**

- The CAMs of the original $A^s$ and tiled $A^{re}$ images are converted using the GAP layer with prediction vectors.

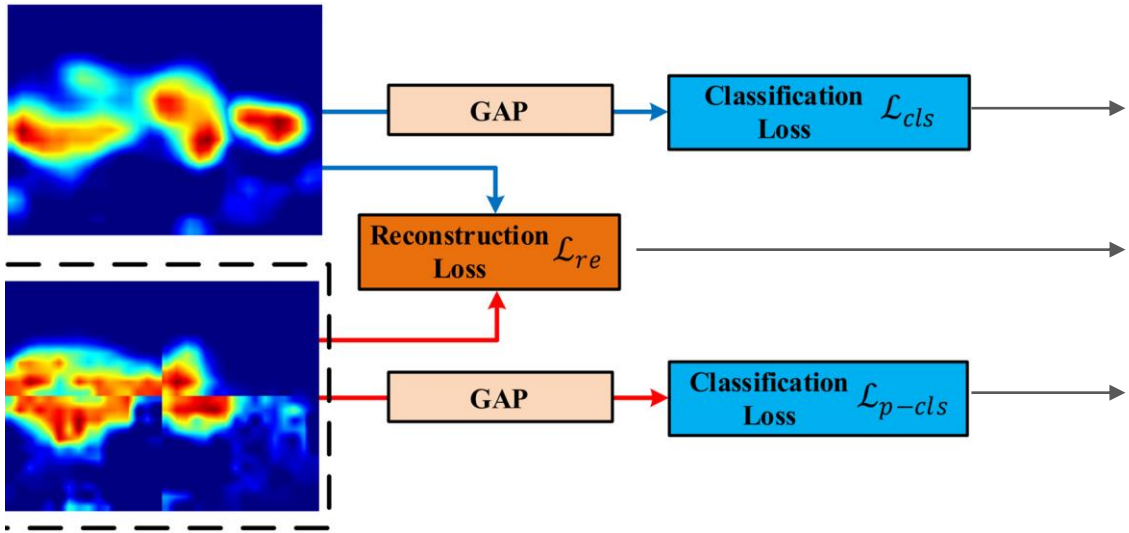



$$f = F(I)$$
$$A_c = \theta_c^\top f$$
$$\hat{Y}^s = G(A^s)$$
$$\hat{Y}^{re} = G(A^{re})$$

# Proposed Method [4/5]

**How to narrow the gaps between the single and reconstructed features?**

- $\alpha$ is the balance of the weights for the different losses.




$$\mathcal{L}_{cls} = \ell_{cls}(\hat{Y}^s, Y)$$

$$\mathcal{L}_{re} = \|A^s - A^{re}\|_1$$

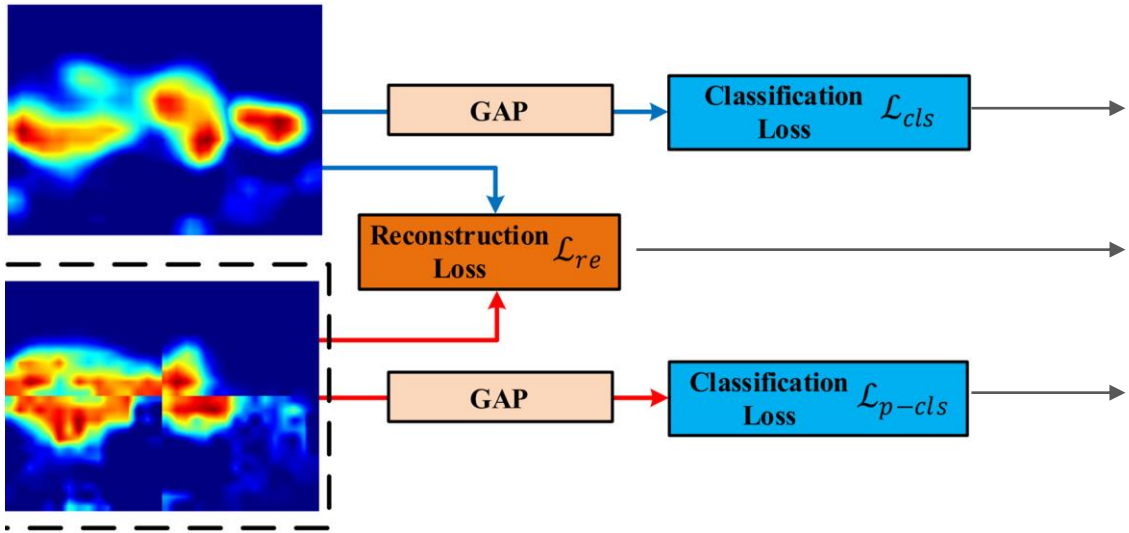
$$\mathcal{L}_{p-cls} = \ell_{cls}(\hat{Y}^{re}, Y)$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{p-cls} + \alpha\mathcal{L}_{re}$$

# Proposed Method [4/5]

**How to narrow the gaps between the single and reconstructed features?**

- $\alpha$ is the balance of the weights for the different losses.




$$\mathcal{L}_{cls} = \ell_{cls}(\hat{Y}^s, Y)$$

$$\mathcal{L}_{re} = \|A^s - A^{re}\|_1$$

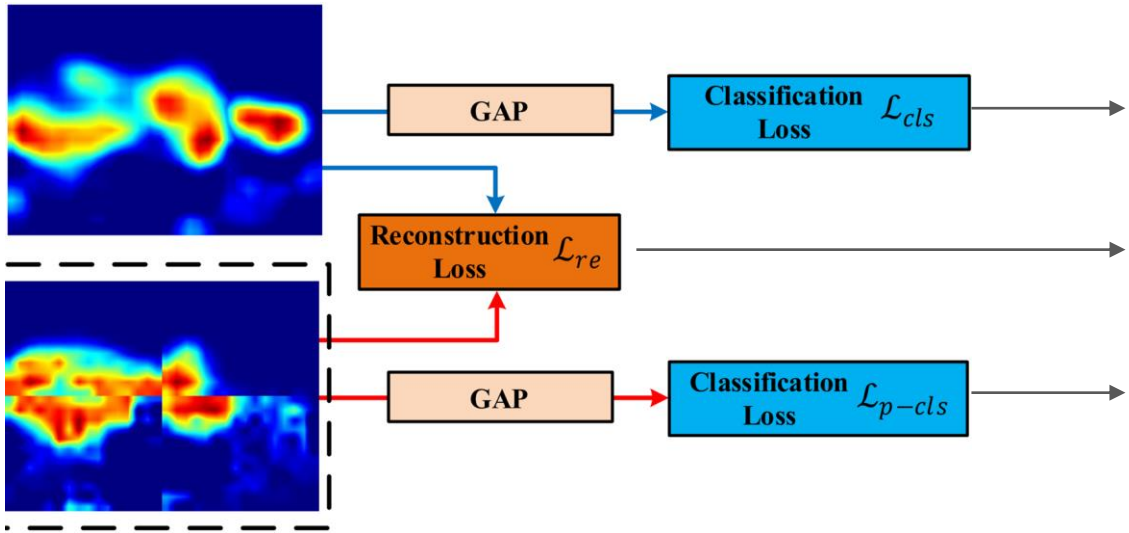
$$\mathcal{L}_{p-cls} = \ell_{cls}(\hat{Y}^{re}, Y)$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{p-cls} + \alpha\mathcal{L}_{re}$$

# Proposed Method [4/5]

**How to narrow the gaps between the single and reconstructed features?**

- $\alpha$ is the balance of the weights for the different losses.




$$\mathcal{L}_{cls} = \ell_{cls}(\hat{Y}^s, Y)$$

$$\mathcal{L}_{re} = \|A^s - A^{re}\|_1$$

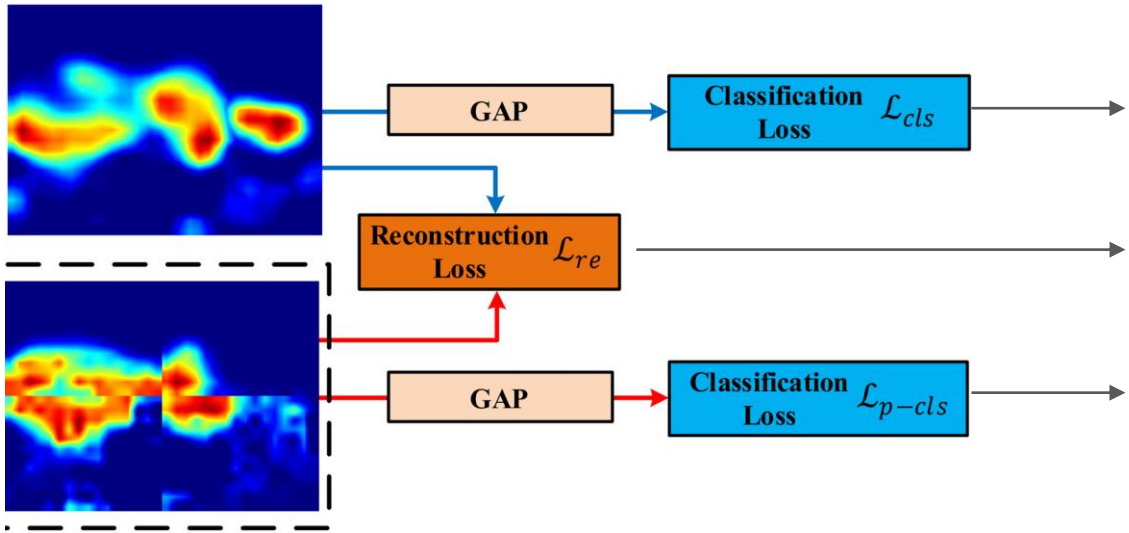
$$\mathcal{L}_{p-cls} = \ell_{cls}(\hat{Y}^{re}, Y)$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{p-cls} + \alpha\mathcal{L}_{re}$$

# Proposed Method [4/5]

**How to narrow the gaps between the single and reconstructed features?**

- $\alpha$ is the balance of the weights for the different losses.




$$\mathcal{L}_{cls} = \ell_{cls}(\hat{Y}^s, Y)$$

$$\mathcal{L}_{re} = \|A^s - A^{re}\|_1$$

$$\mathcal{L}_{p-cls} = \ell_{cls}(\hat{Y}^{re}, Y)$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{p-cls} + \alpha\mathcal{L}_{re}$$

# Proposed Method [5/5]

## Using different loss functions

- Proposed loss functions consistently improved the baseline by a 3.71%.
- In figure (d), the final CAMs not only suppressed over-activation but also expanded the CAMs into complete object activation coverage.


(a) $L_{cls}$


(b) $L_{cls} + L_{p-cls}$


(c) $L_{cls} + L_{re}$


(d) $L_{cls} + L_{p-cls} + L_{re}$

**Table 1**: Ablation study of the Puzzle-CAM loss functions using ResNet-50 as the backbone.

| $L_{cls}$ | $L_{p-cls}$ | $L_{re}$ | mIoU (%) |
|---|---|---|---|
| ✓ | | | 47.82 |
| ✓ | ✓ | | 47.70 |
| ✓ | | ✓ | 49.21 |
| ✓ | ✓ | ✓ | **51.53** |

# Experimental Results [1/3]

## Quantitative results

- Our approach outperforms existing state-of-the-art methods without additional supervisions on PASCAL VOC 2012 *val* and *test* sets.
- Validation set: 61.7% → 71.9%, Test set: 63.7% → 72.2%

**Table 2**: Quality of the pseudo semantic segmentation labels in mIoU, evaluated on the PASCAL VOC 2012 training set [14]. RW, random walk with AffinityNet [4]; dCRF, dense conditional random field [16].

| Method | Backbone | CAM (%) | CAM +RW (%) | CAM+RW +dCRF (%) |
|---|---|---|---|---|
| AffinityNet [4] | ResNet-50 | 47.82 | 58.10 | 59.70 |
| Puzzle-CAM | ResNet-50 | 51.53 | 64.16 | 64.70 |
| Puzzle-CAM | ResNeSt-50 | 57.59 | 69.48 | 69.91 |
| Puzzle-CAM | ResNeSt-101 | 61.85 | 71.92 | 72.46 |
| Puzzle-CAM | ResNeSt-269 | **62.45** | **74.14** | **74.67** |

**Table 3**: Comparison of Puzzle-CAM and existing state-of-the-art methods on the PASCAL VOC 2012 *val* and *test* datasets. $\mathcal{I}$, image-level labels; $\mathcal{S}$, external saliency models.

| Method | Backbone | Supervision | val | test |
|---|---|---|---|---|
| AffinityNet [4] | Wide-ResNet-38 | $\mathcal{I}$ | 61.7 | 63.7 |
| DSRG [12] | ResNet-101 | $\mathcal{I}+\mathcal{S}$ | 61.4 | 63.2 |
| SeeNet [13] | ResNet-101 | $\mathcal{I}+\mathcal{S}$ | 63.1 | 62.8 |
| IRNet [4] | ResNet-50 | $\mathcal{I}$ | 63.5 | 64.8 |
| FickleNet [6] | ResNet-101 | $\mathcal{I}+\mathcal{S}$ | 64.9 | 65.3 |
| ICD [17] | ResNet-101 | $\mathcal{I}$ | 64.1 | 64.3 |
| SEAM [5] | Wide-ResNet-38 | $\mathcal{I}$ | 64.5 | 65.7 |
| Ours (Puzzle-CAM) | ResNeSt-101 | $\mathcal{I}$ | 66.9 | 67.7 |
| Ours (Puzzle-CAM) | ResNeSt-269 | $\mathcal{I}$ | **71.9** | **72.2** |

[4] Ahn et al., Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation (CVPR 2018)
[5] Wang et al., Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation (CVPR 2020)
[17] Fan et al., Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation (CVPR 2020)

# Experimental Results [2/3]

## Qualitative results

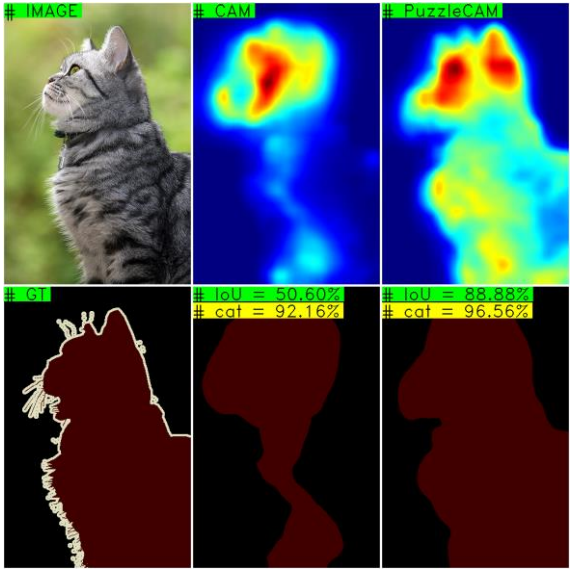
- Comparison of (a) original images. (b) ground truth. (c) segmentation results predicted by AffinityNet. (d) segmentation results predicted by Puzzle-CAM.

# Experimental Results [3/3]

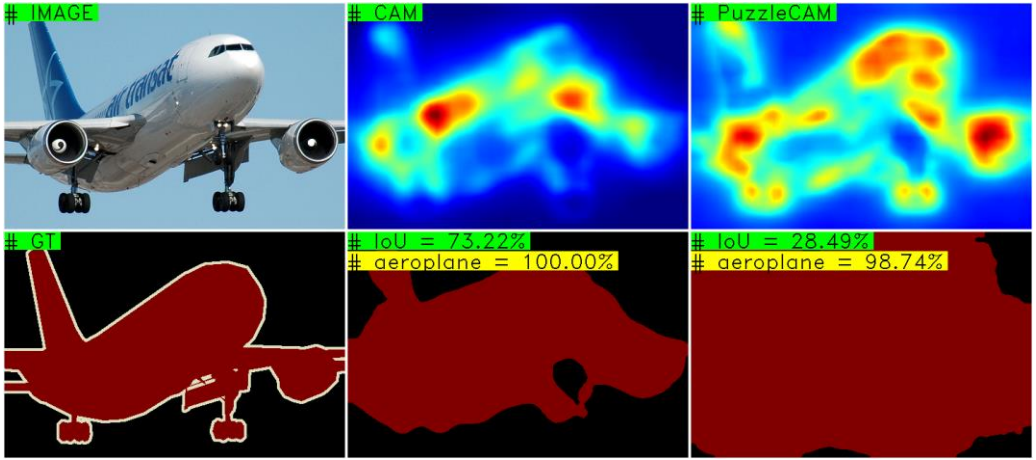
## Additional qualitative results

- Our method can produce high quality semantic segmentation results. Sometimes, our method produces over-activated CAMs on similar background or small objects.

Successful case




Typical failure case



# Conclusions

**Resulting in enhanced CAMs detecting the integral region of the object.**

- Designing a puzzle module and reconstructing regularization to effectively enhance the quality of CAMs without adding layers and pixel-level annotations.

- Our method significantly outperforms existing state-of-the-art methods with the same level of supervision on the PASCAL VOC 2012 dataset.

https://github.com/OFRIN/PuzzleCAM

# Thank you !

*Sanghyun Jo*

GYNetworks

josanghyeokn@gynetworks.com

*In-Jae Yu**

KAIST
School of Computing

myhome98304@gmail.com

*Corresponding Author