# HRDA Loss Explained

(원 논문 리뷰)

(+ Quick overview of mean teacher semi supervised learning)

**(+ 저번에 물어보셨던 것에 대한 더 나은 답변 드립니다!)**

✔️ High Resolution vs. Low Resolution 관련

High resolution 이미지에 bilinear downsampling $\zeta$ 을 활용해서 Low Resolution을 만들었다고 합니다 :)

**[Low resolution 수식]**

$x^T_{LR} = \zeta(x^T_{HR}, 1/s_T) \in \mathbb{R}^{\frac{H_T}{s_T} \times \frac{W_T}{s_T} \times 3}$

($s_T$: dataset specific factor로 1 이상의 scalar 값)

## Preliminary

📎 **Basic Notations**

$f_\theta$**: neural network**

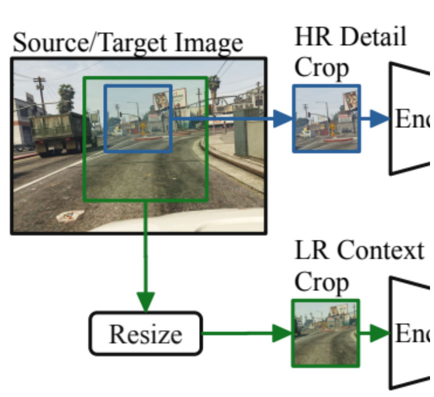
$m$**: index** $H$**: height** $W$**: width** $HR$**: High resolution** $LR$**: Low resolution**

$\mathcal{X}^S = \{x^{S,m}_{HR}\}^{N_S}_{m=1}$**: source domain images (**$x^{S,m}_{HR} \in \mathbb{R}^{H_S \times W_S \times 3}$**)**

$\mathcal{X}^T = \{x^{T,m}_{HR}\}^{N_T}_{m=1}$**: target domain images (**$x^{T,m}_{HR} \in \mathbb{R}^{H_T \times W_T \times 3}$**)**

$\mathcal{Y}^S = \{y^{S,m}_{HR}\}^{N_S}_{m=1}$**: labels for the source domain (**$\{y^{S,m}_{HR}\}^{N_S}_{m=1} \in \{0,1\}^{H_S \times W_S \times C}$**)**

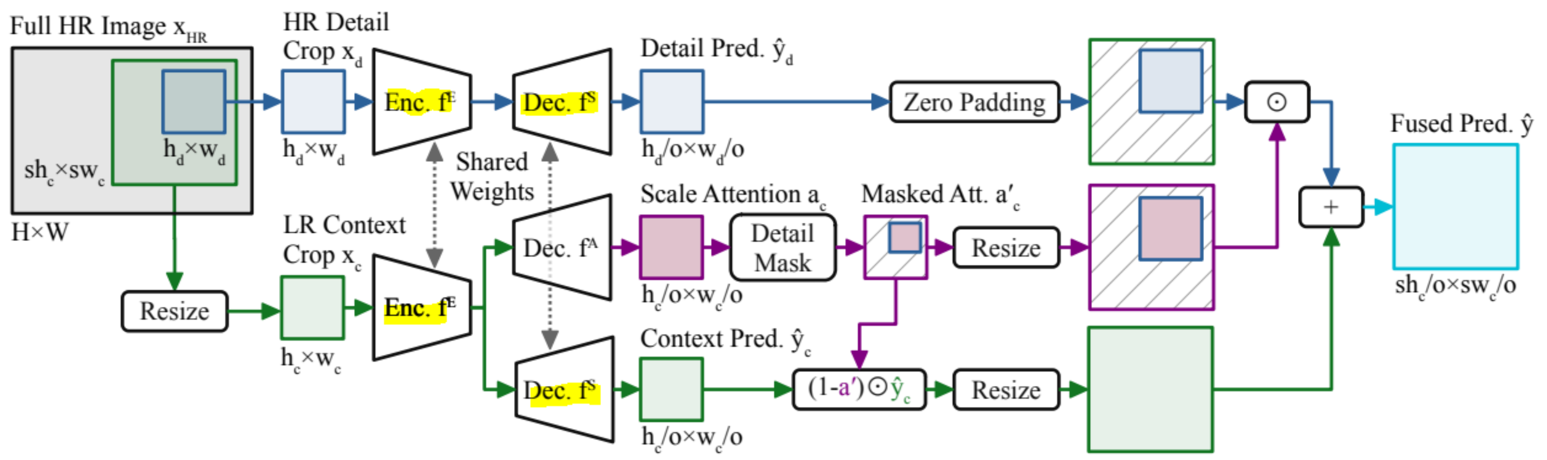
## Notations about the proposed method




$x_c$**: context crop (**$\in \mathbb{R}^{h_c \times w_c \times 3}$**)**

$x_d$: **detail crop** ($\in \mathbb{R}^{h_d \times w_d \times 3}$)

( $h_c$ = $h_d$ , $w_c$ = $w_d$ )




$f^E$: feature encoder

$f^S$: semantic decoder

$f^A$: scale attention decoder

$\hat{y}_c = f^S(f^E(x_c)) \in \mathbb{R}^{\frac{h_c}{o} \times \frac{w_c}{o} \times C}$: the context semantic segmentation

$\hat{y}_d = f^S(f^E(x_d)) \in \mathbb{R}^{\frac{h_d}{o} \times \frac{w_d}{o} \times C}$: the detail semantic segmentation

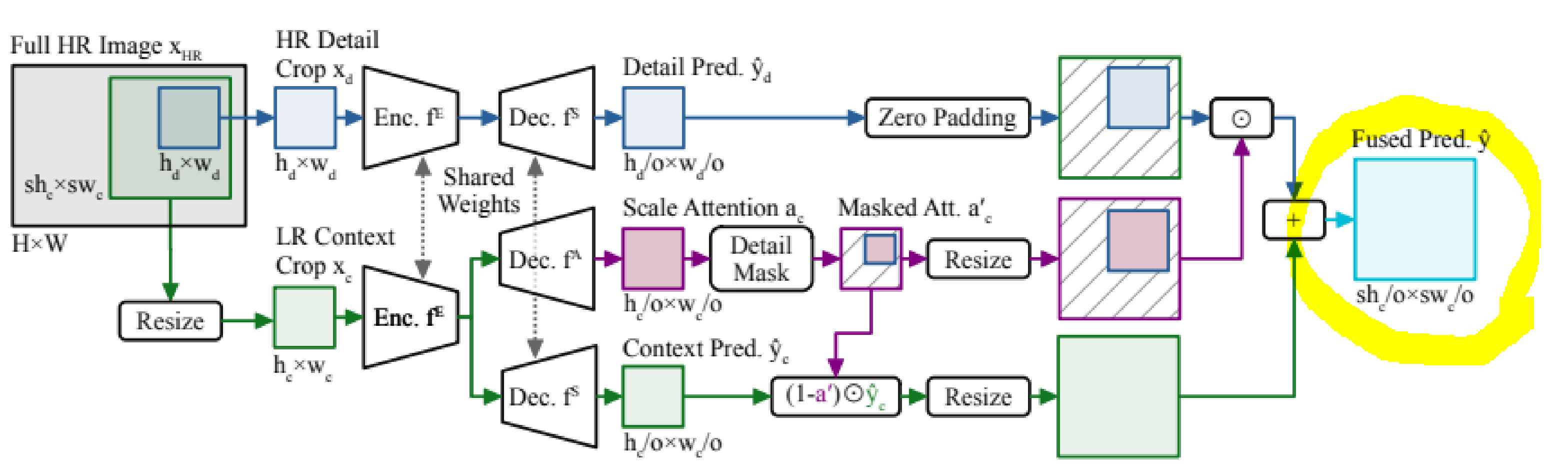
$a_c = \sigma(f^A(f^E(w_c))) \in [0,1]^{\frac{h_c}{o} \times \frac{w_c}{o} \times C}$: the scale attention to weigh the trustworthiness of LR context and HR detail predictions (1: focus on the HR detail crop)

$a'_c \in \mathbb{R}^{\frac{h_d}{o} \times \frac{w_d}{o}}$

$$a'_c(i,j) = \begin{cases} a_c(i,j) & \text{if } \frac{b_{d,1}}{s \cdot o} \le i < \frac{b_{d,2}}{s \cdot o} \wedge \frac{b_{d,3}}{s \cdot o} \le j < \frac{b_{d,4}}{s \cdot o} \\ 0 & \text{otherwise} \end{cases}$$

( = detail crop 외의 부분은 다 0으로 처리)

$\hat{y}'_d$: $\hat{y}_d$에 0으로 테두리 패딩을 두른 mask




**the predictions from multiple scales fused by the attention-weighted sum (노란 동그라미 부분)**

$\hat{y}_{c,F} = \zeta((1 - a'_c) \odot \hat{y}_c, s) + \zeta(a'_c, s) \odot \hat{y}'_d$

# Loss

> In this work, we mainly evaluate HRDA with the self-training method DAFormer [29], as it is the current state-of-the-art method for UDA semantic segmentation.

**Preliminary**

$$\mathcal{L}^S = L_{ce}(\hat{y}^S_{LR}, y^S_{LR}, 1)$$
$$L_{ce} = -\sum_{i=1}^{H(y)} \sum_{j=1}^{W(y)} \sum_{c=1}^{C} q_{ij} y_{ijc} \log\zeta(\hat{y}, \frac{H(y)}{H(\hat{y})})_{ijc}$$

**[Loss for the source domain $\mathcal{L}^S$]**

$$\mathcal{L}^S_{HRDA} = (1-\lambda_d)\mathcal{L}_{ce}(\hat{y}^S_{c,F}, y^S_{c,HR}, 1) + \lambda_d \mathcal{L}_{ce}(\hat{y}^S_d, y^S_d, 1),$$

설명 (논문에는 안 나온 내용이긴 합니다..!)

- $L^S_{HRDA}$: context를 위한 loss(첫항)과 detail를 위한 loss (두번째 항)의 조합
- $\lambda_d$: context를 위한 Loss와 detail을 위한 loss를 조절하는 값

**[Loss for the target domain $\mathcal{L}^T$]**

**Preliminary**

**Pseudo-label**
$p^T_{LR,ijc} = [c = argmax_{c'} g_\phi(x^T_{LR})_{ijc'}]$
$g_\phi$: teacher network

$$\mathcal{L}^T_{HRDA} = (1-\lambda_d)\mathcal{L}_{ce}(\hat{y}^T_{c,F}, p^T_{c,F}, q^T_{c,F}) + \lambda_d \mathcal{L}_{ce}(\hat{y}^T_d, p^T_d, q^T_d)$$

설명 (논문에는 안 나온 내용이긴 합니다..!)

- $L^T_{HRDA}$: context를 위한 loss(첫항)과 detail를 위한 loss (두번째 항)의 조합
- $\lambda_d$: context를 위한 Loss와 detail을 위한 loss를 조절하는 값
- target domain에선 label이 없기 때문에 network에 넣어서 얻은 pseudo label인 $p^T$로 대체
- confidence estimate $q^T$ 계산 방법:

pseudo-labels. Here, we use the ratio of pixels exceeding a threshold $\tau$ of the maximum softmax probability [71]

$$q_T^{(i)} = \frac{\sum_{j=1}^{H\times W}[\max_{c'} h_\phi(x_T^{(i)})^{(j,c')} > \tau]}{H \cdot W}. \quad (3)$$

Daformer에서 발췌

Pseudo-label $p$ 생성시 HRDA에선 teacher network가 없기 때문에 network $f_\phi$를 활용

**Final Loss** $\mathcal{L} = \mathcal{L}^S + \mathcal{L}^T + \lambda_{FD}\mathcal{L}_{FD}$

(DAFormer 관련)

> Therefore, we assume that the useful features from ImageNet pretraining are corrupted by $L_S$ and the model overfits to the synthetic source data. In order to prevent this issue, we regularize the model based on the Feature Distance (FD) of the bottleneck features $F_\theta$ of the semantic

> segmentation UDA model $g_\theta$ and the bottleneck feature $F_{ImageNet}$ of the ImageNet model. (DAFormer에서 발췌)

$$d^{(i,j)} = ||F_{ImageNet}(x_S^{(i)})^{(j)} - F_\theta(x_S^{(i)})^{(j)}||_2 \,.$$

> However, the ImageNet model is mostly trained on thing-classes (objects with a well-defined shape such as car or zebra) instead of stuff-classes (amorphous background regions such as road or sky). Therefore, we calculate the FD loss only for image regions containing thing-classes $C_{things}$ described by the binary mask $M_{things}$ (DAFormer에서 발췌)

$$\mathcal{L}_{FD}^{(i)} = \frac{\sum_{j=1}^{H_F \times W_F} d^{(i,j)} \cdot M_{things}^{(i,j)}}{\sum_j M_{things}^{(i,j)}}$$

L_FD (ImageNet 지식을 유지하기 위함!)

This mask is obtained from the downscaled label $y_{S,small}$

$$M_{things}^{(i,j)} = \sum_{c'=1}^{C} y_{S,small}^{i,j,c'} \cdot [c' \in \mathcal{C}_{things}] \,. \qquad (10)$$

M_things 계산공식