Semantic Image Matting

Yanan Sun
HKUST
now.syn@gmail.com

Chi-Keung Tang
HKUST
cktang@cs.ust.hk

Yu-Wing Tai
Kuaishou Technology
yuwing@gmail.com

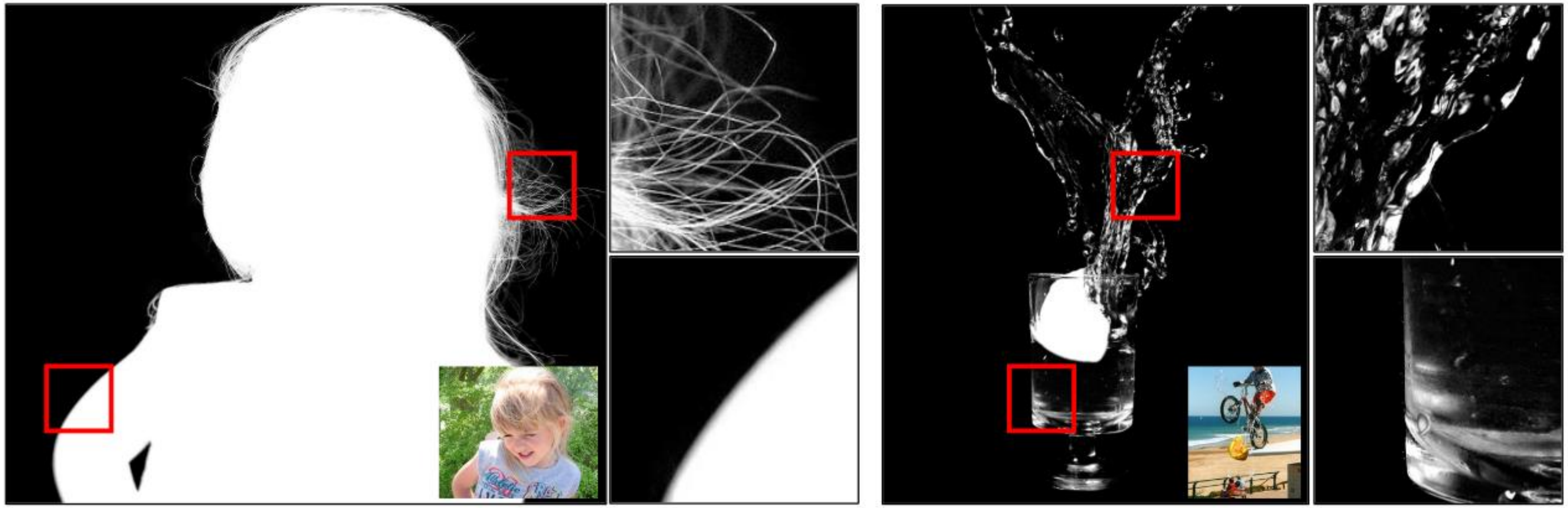


Image Matting Review

- Matting
  - Alpha matte： shape of 1xHxW, each pixel represents the transparency of that pixel’s corresponding foreground
  - An ill-posed problem： $I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad \alpha_i \in [0, 1].$
  - General guidance： Trimap

  - Input: RGB image and trimap(1 channel or 3 channel)
  - Output: 1 channel alpha matte
  - Model： U-Net like structure

Main Idea & Contributions

1. First introduce semantics into the matting task

2. Purpose The first large-scale class-balanced Semantic Image Matting Dataset (SIMD)

3. Main technical contributions include:
    1. the introduction of semantic trimap
    2. the proposal of learnable content-sensitive weights
    3. the usage of multi-class discriminator to regularize the matting results.

Semantic Image Matting Dataset

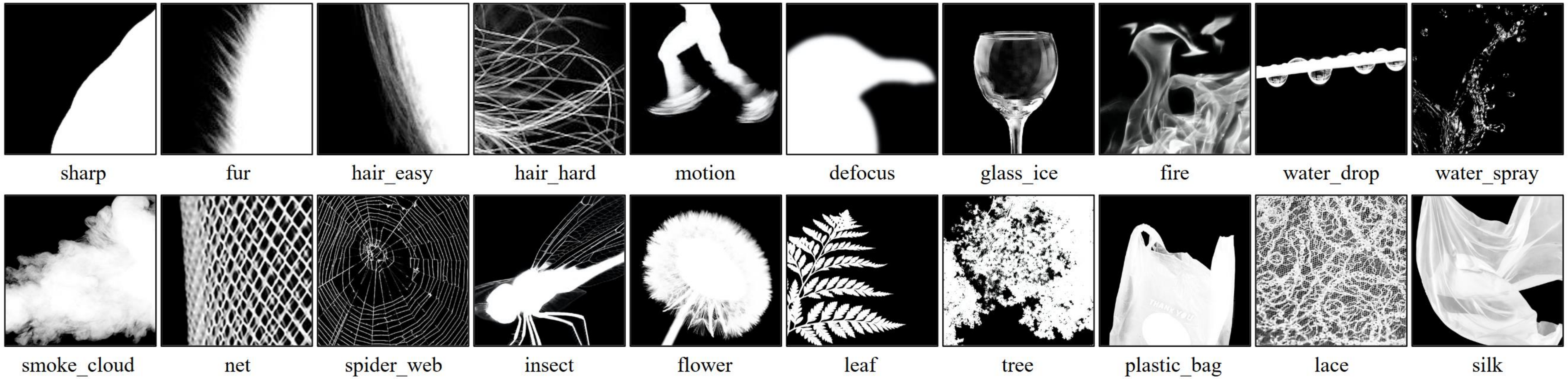


Figure 2. The 20 matting classes with high diversity in appearance across different classes.

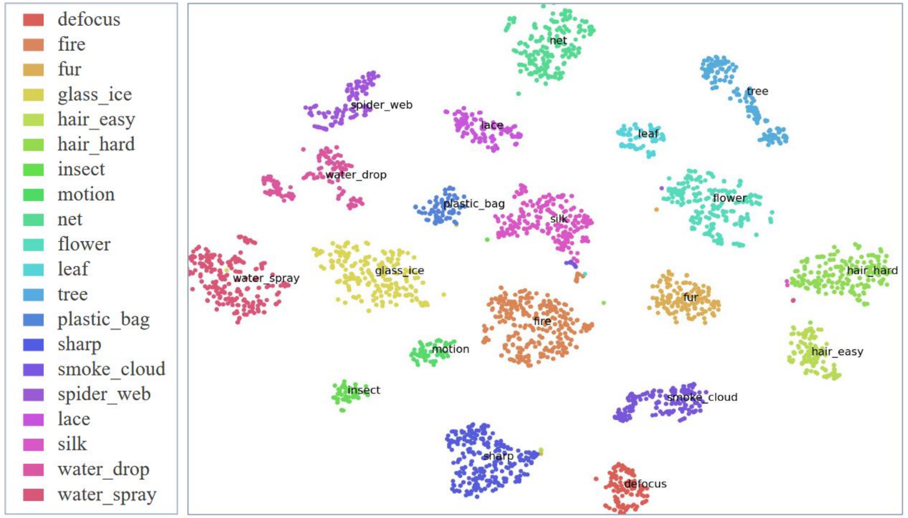


Figure 3. t-SNE visualizations of the class-specific features extracted from our discriminator (for its design see Method section).

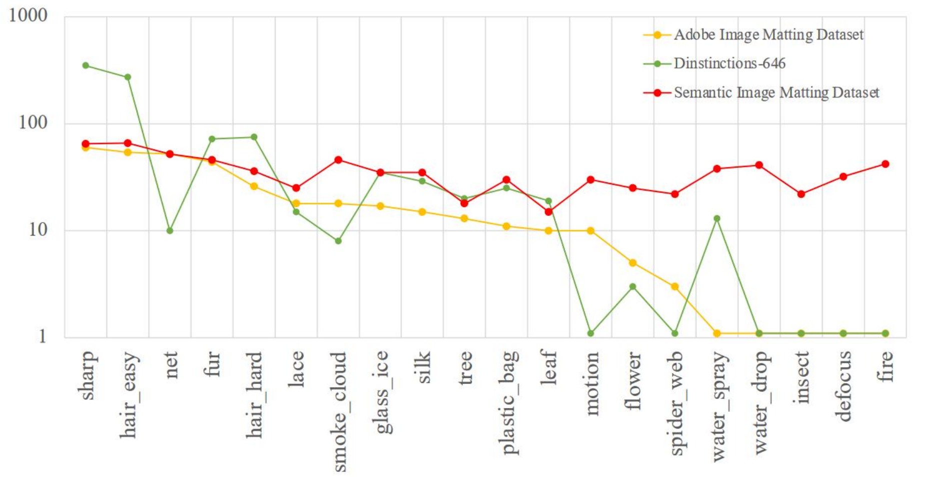
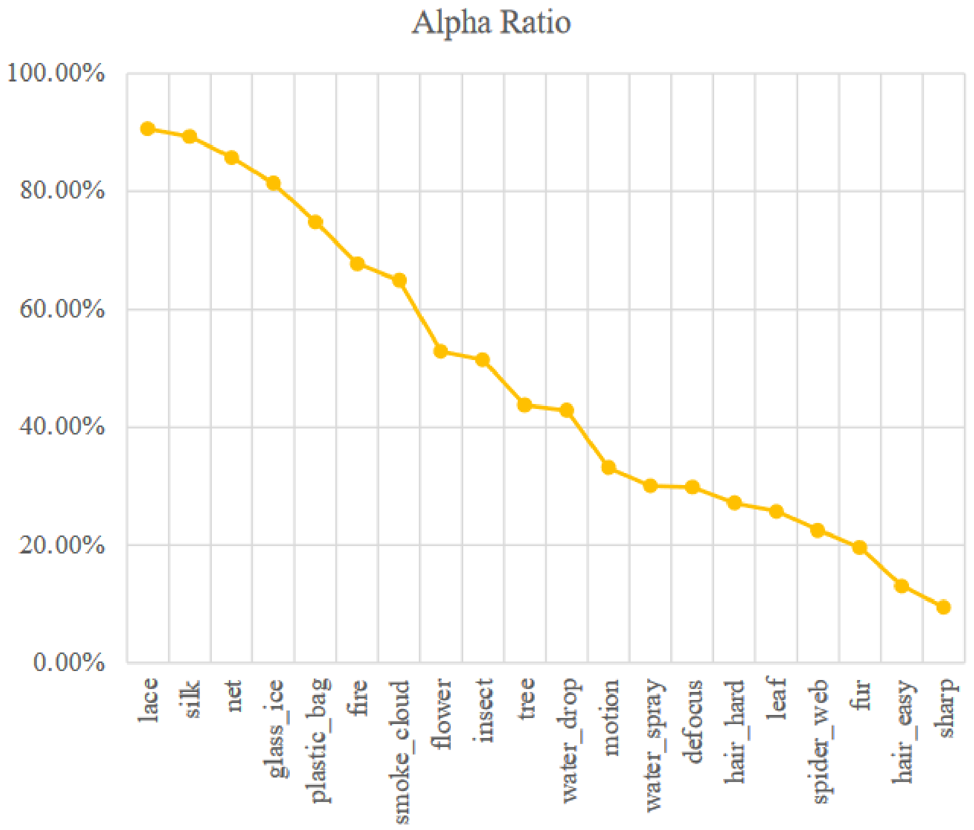
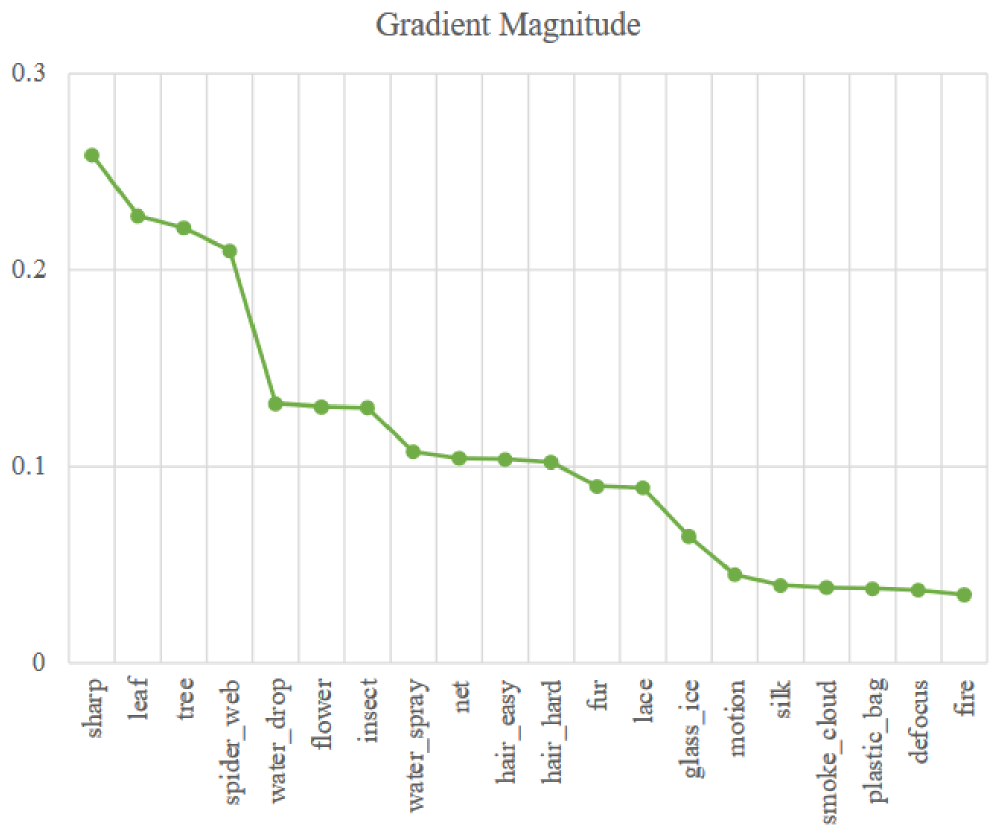


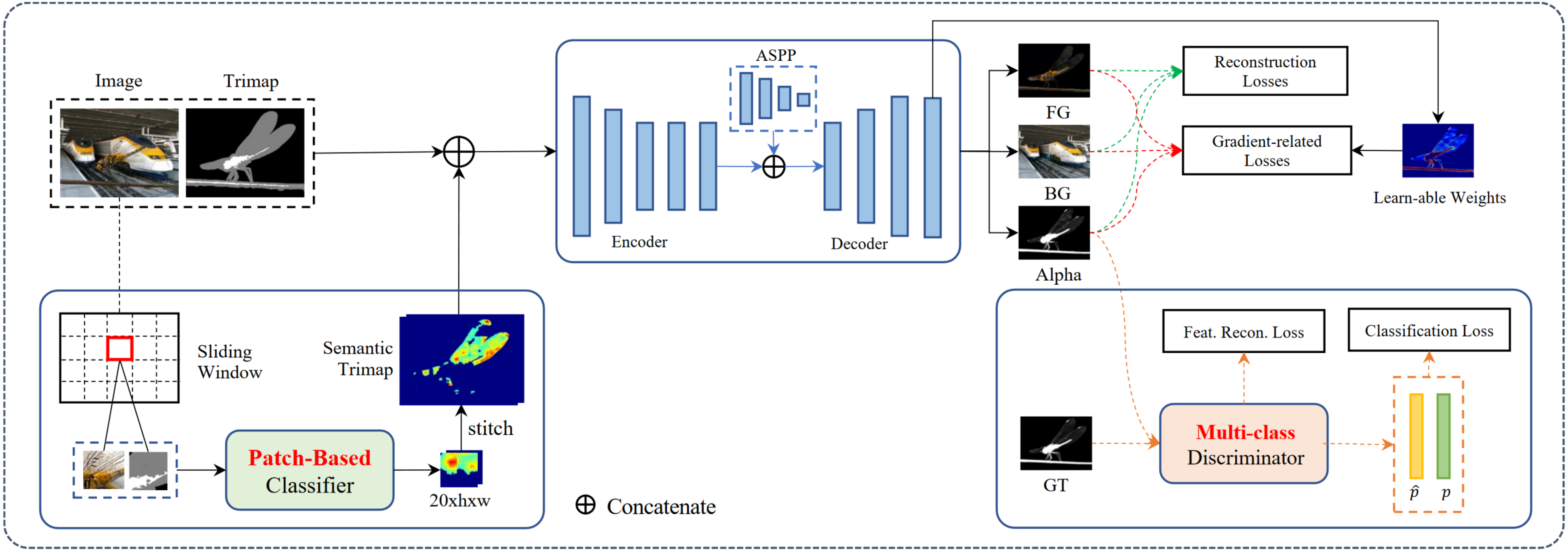
Figure 4. Class distribution of three matting datasets: Adobe Image Matting Dataset [43], Distinctions-646 [33] and our Semantic Image Matting Dataset.

Analysis and Motivation

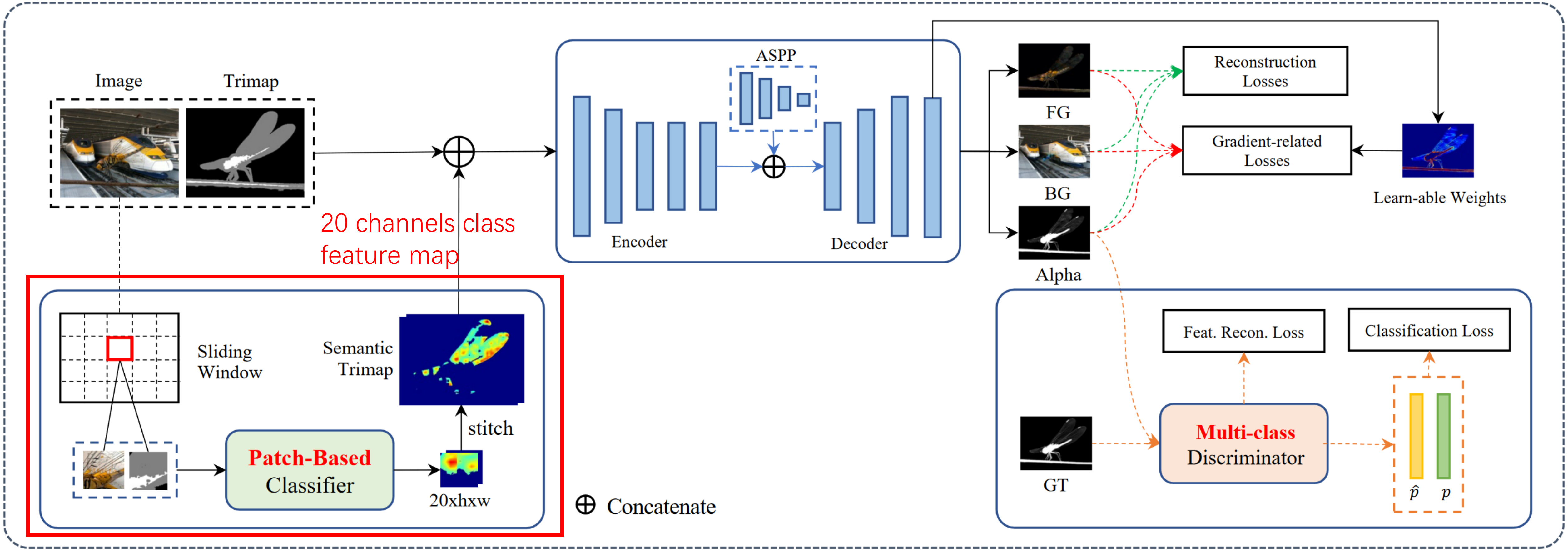




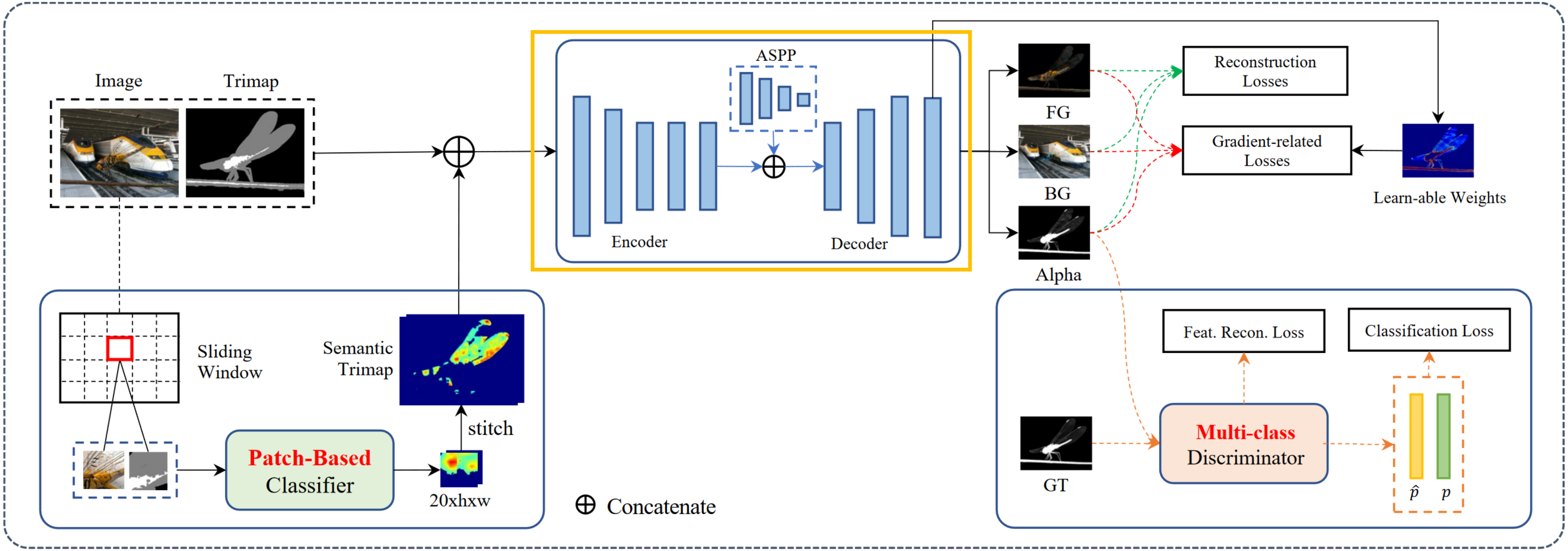
Overview Framework



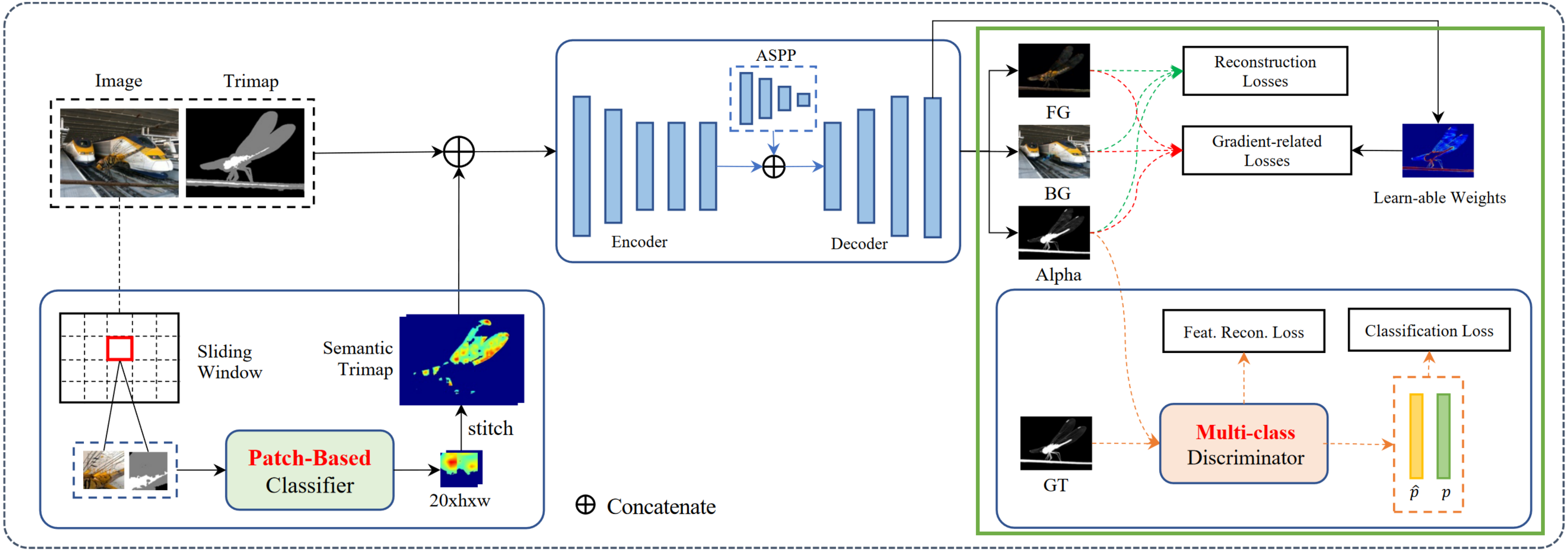
Overview Framework



Overview Framework



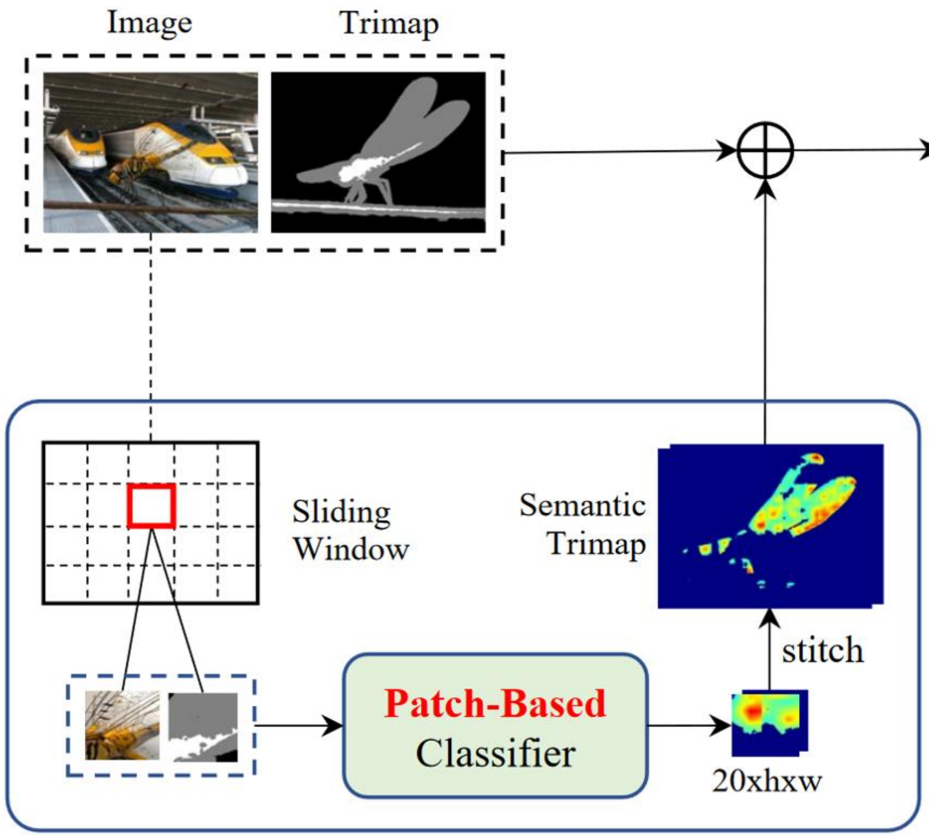
Overview Framework



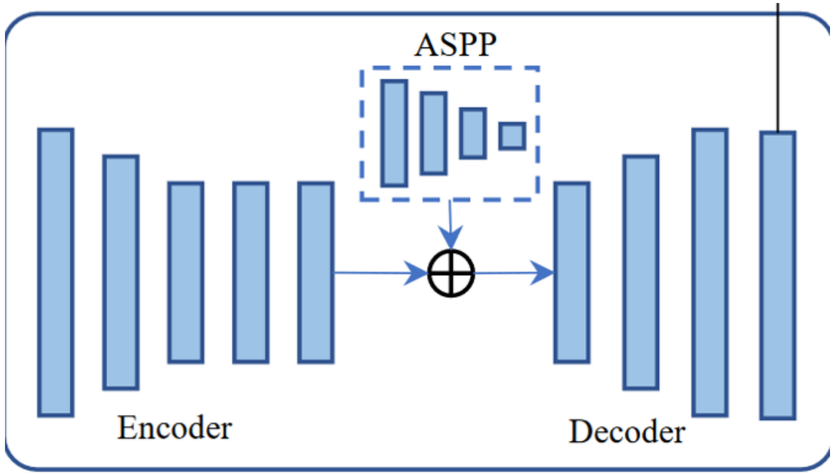
Semantic Trimap

- An extra classifier
  - Input: Image concats original trimap
  - Random crop the unknown area of input to different size patches and resize for training
  - Output: 20 classes classify result

- Semantic trimap
  - After the classifier is well trained, change the softmax layer to fc to get 20xhxw classes feature map
  - Stitch patches to a whole guidance map



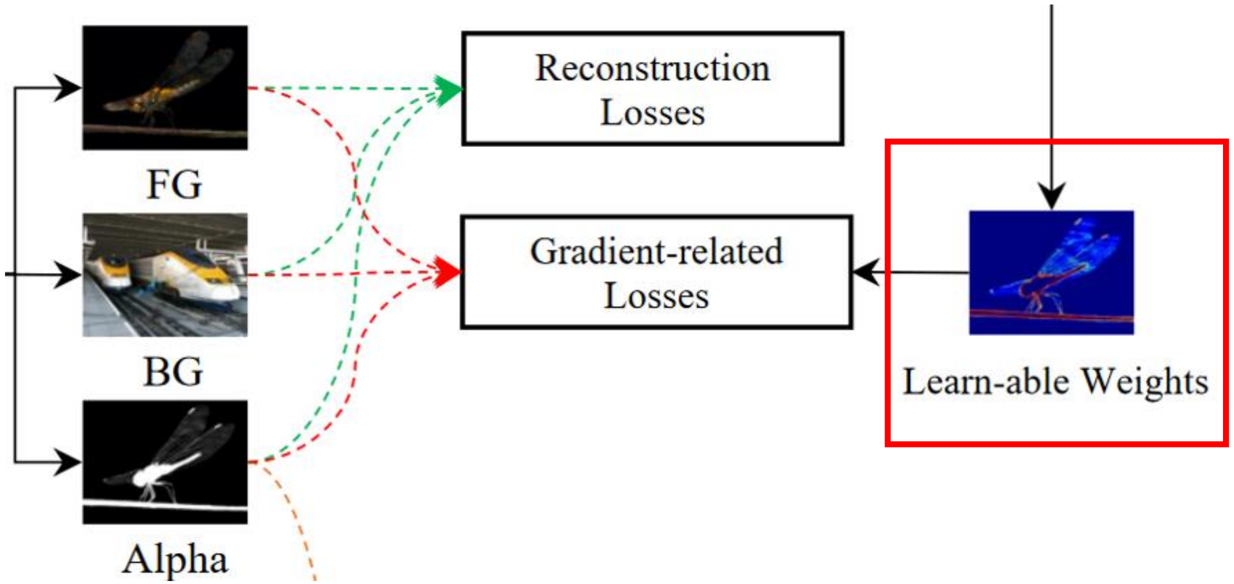
Encoder-Decoder Structure



- Encoder
  - Resnet 50 backbone
  - Changing layer3 and layer4’s downsample layer to dilated convolution
  - Be used to enlarge receptive fields.
- Extra ASPP
  - Applied to aggregate features of different receptive fields in order to enhance the feature representation capability.
- Decoder
  - Sample FPN structure
  - Output 7 channel (Alpha matte prediction, Foreground prediction, Background prediction)

Learnable Content-Sensitive Weights

- Observation:
  - Each matting class represents a distinct appearance and structure and thus its respective color and alpha exhibit different gradient distributions from others
- Example:
  - Hair consists of fine structures with large gradients along hair boundaries
  - Fire exhibits smooth transition across its foreground region.
- Make the model aware of gradient changes
  - Derivation of the original formula:
  - Introduction of content sensitive weights:



$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad \alpha_i \in [0, 1].$$

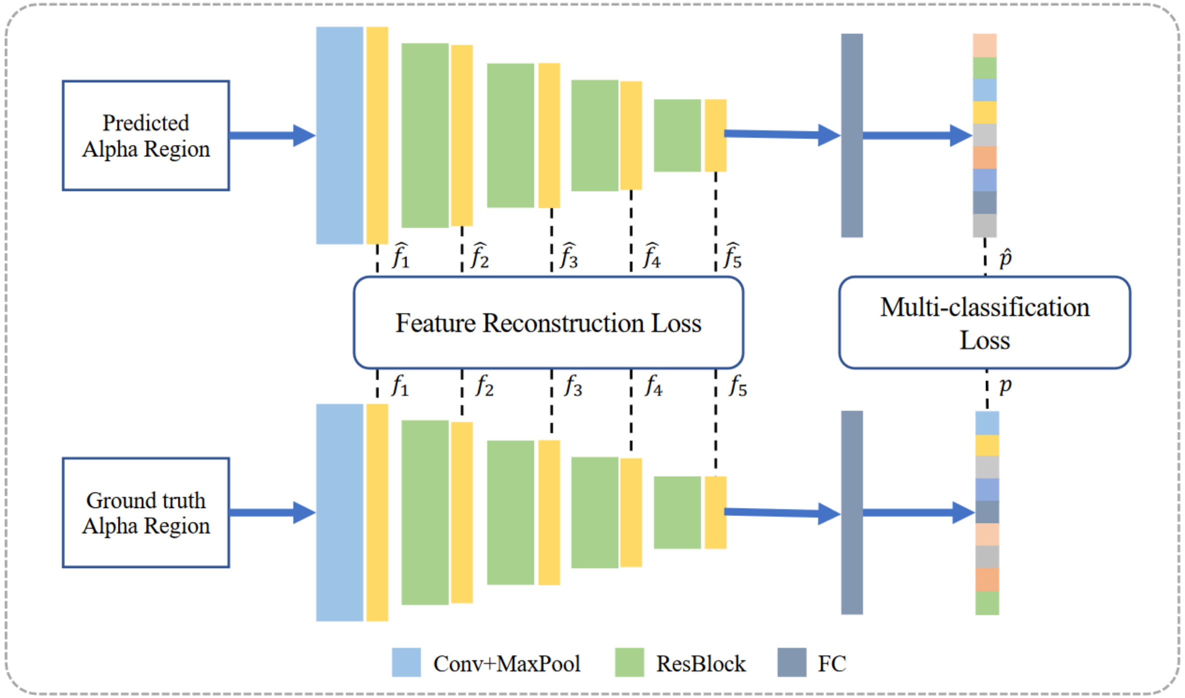
$$\boxed{\nabla I = (F - B)\nabla\alpha + \alpha\nabla F + (1 - \alpha)\nabla B}$$

$$\nabla I = \lambda_1 \nabla\alpha + (1 - \lambda_2)\nabla F + \lambda_2 \nabla B$$

$\lambda 1, \lambda 2 \in \mathrm{R}^{3*H*W}$

Multi-Class Discriminator

- Another extra classifier
  - Similar structure with trimap classifier, but the input is only 1-channel alpha matte
  - Output include classification results and intermediate features



Loss Function

- Reconstruction Losses

$$L_\alpha = \frac{1}{|U|}\sum_{i\in U}\|\hat{\alpha}_i - \alpha_i\|_1 + \frac{1}{|U|}\sum_{i\in U}\|\hat{I}_i - I_i\|_1 + L_{lap}$$

$$L_{FB} = \frac{1}{|\widetilde{F}|}\sum_{i\in\widetilde{F}}\|\hat{F}_i - F_i\|_1 + \frac{1}{|\widetilde{B}|}\sum_{i\in\widetilde{B}}\|\hat{B}_i - B_i\|_1$$

- Classification and Feature Reconstruction Loss

$$L_c = -\sum_j \hat{p}_j \log p_j$$

$$L_f = \sum_k \frac{1}{|f_k|}\|\hat{f}_k - f_k\|_2$$

- Gradient-related Loss

$$L_g = \frac{1}{|U|}\sum_{i\in U}\|\nabla\hat{I}_i - \nabla I_i\|_1$$

$$\nabla\hat{I}_i = \lambda_1\nabla\hat{\alpha}_i + (1-\lambda_2)\nabla\hat{F}_i + \lambda_2\nabla\hat{B}_i$$

- Total

$$L_e = \frac{1}{|U|}\sum_{i\in U}\|\nabla\hat{F}_i\|_1\|\nabla\hat{B}_i\|_1 + \|\nabla\hat{\alpha}_i\|_1\|\nabla\hat{B}_i\|_1$$

$$L = L_\alpha + 0.2(L_{FB} + L_f + L_g + L_e) + 0.1L_c$$

Results

Method	SAD	MSE	Grad	Conn	# Params
CF [15]	168.1	0.091	126.9	167.9	-
KNN [2]	175.4	0.103	124.1	176.4	-
DIM [34]	50.4	0.014	31.0	50.8	$>$ 130.55M
IndexNet [22]	45.8	0.013	25.9	43.7	8.15M
CA [14]	35.8	0.0082	17.3	33.2	107.5M
CA+$\mathfrak{DA}$ [14]	71.3	0.0236	38.8	72.0	107.5M
GCA [18]	35.28	0.0091	16.9	32.5	25.27M
A^2U [4]	32.15	0.0082	16.39	29.25	8.09M
SIM [28]	28.0	0.0058	10.8	24.8	70.16M
FBA [11]	26.4	0.0054	10.6	21.5	34.69M
FBA+TTA [11]	25.8	0.0052	10.6	20.8	34.69M

	Distinction-646 [24]				SIMD_{our} [28]			
Method	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn
IndexNet* [22]	42.64	0.0256	40.17	42.76	92.45	0.0388	45.85	93.14
CA* [14]	49.07	0.0557	114.77	48.27	79.46	0.0291	51.03	77.88
CA+$\mathfrak{DA}$* [14]	46.03	0.0356	55.45	46.18	102.97	0.0469	74.39	103.52
GCA* [18]	31.00	0.0171	21.19	29.62	75.81	0.0271	40.57	74.45
A^2U* [4]	28.74	0.0143	17.42	27.62	68.70	0.0268	39.00	66.76
SIM* [28]	**22.68**	0.0137	20.11	21.03	**37.07**	0.0099	22.29	33.30
FBA* [11]	30.70	0.0150	18.89	29.65	41.55	0.0109	23.21	35.07

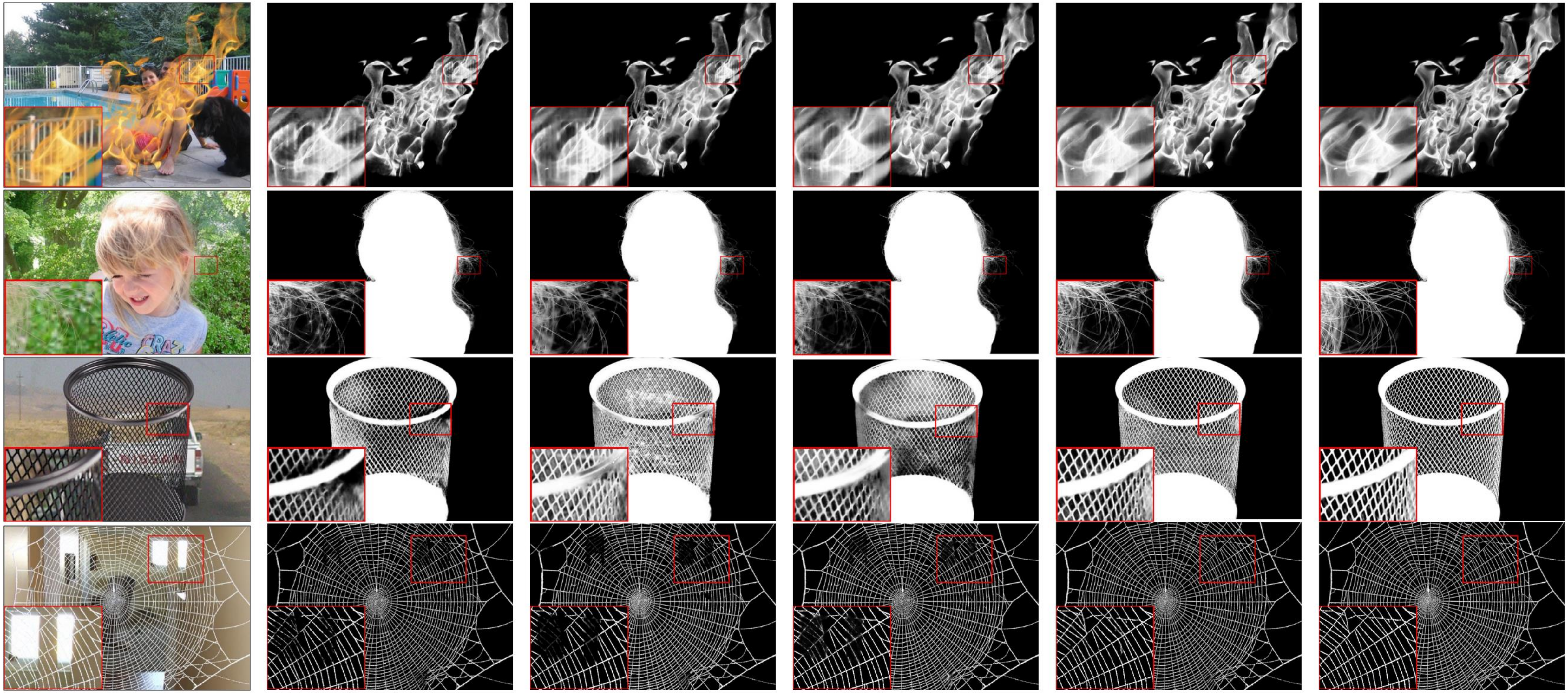


Image DIM IndexNet GCA SIM (Ours) GT

Robustness to different classes

Classes	defocus	fire	fur	glass_ice	hair_easy	hair_hard	insect	motion	net	flower
DIM [43]	25.91	60.53	9.88	91.36	11.23	13.01	111.21	6.78	87.09	65.40
IndexNet [30]	22.86	97.85	9.99	91.95	8.33	13.24	130.52	6.68	91.43	59.60
GCA [26]	18.33	46.29	8.12	76.20	8.24	11.31	99.11	6.08	83.71	44.86
SIM (Ours)	**13.49**	**35.44**	**5.90**	**49.19**	**5.68**	**7.72**	**96.85**	**4.04**	**50.35**	**37.10**

Classes	leaf	tree	plastic_bag	sharp	smoke_cloud	spider_web	lace	silk	water_drop	water_spray
DIM [43]	45.43	91.71	65.44	2.96	48.21	145.57	101.78	51.89	32.48	41.96
IndexNet [30]	43.85	99.26	89.70	3.32	35.31	145.62	114.47	62.81	33.90	34.92
GCA [26]	41.12	87.61	47.40	3.35	41.18	107.14	80.51	51.93	25.83	31.12
SIM (Ours)	**20.98**	**34.14**	**36.70**	**1.39**	**27.42**	**63.79**	**51.08**	**41.78**	**16.94**	**20.53**

Robustness to trimap

Methods	SAD				MSE	Grad
	Overall	S	L	U	Overall	Overall
AdaMatting [6]	7.6	6.9	6.5	9.4	8.5	8.1
SampleNet [40]	8.2	6.5	7.6	10.5	9.2	9.5
Background [37]	7.9	5.9	5.4	12.4	7.4	6.9
GCA [26]	9	10	6.4	10.8	9.9	8.2
SIM (Ours)	**2.5**	**2.6**	**1.8**	**3**	**2.9**	**3.1**

Table 4. Quantitative results of our method and several representative state-of-the-art methods on alphamatting.com [1] benchmark. “S”, “L”, “U” denote three trimap sizes and scores denote average rank across 8 test samples. Best results are shown in bold.

Ablation

Methods	SAD	MSE(10^3)	Grad	Conn
Basic	32.04	5.9	12.05	26.20
Basic + S	30.24	5.4	11.60	23.83
Basic + S + D	29.84	5.3	12.37	23.33
Basic + S + D + G	27.87	4.7	11.57	20.83

Table 5. Ablation studies. The basic model is trained with reconstruction losses. “S”, “D”, “G” denotes semantic trimap , multi-class discriminator and gradient-related losses respectively.

$$\nabla I = (F - B)\nabla\alpha + \alpha\nabla F + (1 - \alpha)\nabla B$$

$$\nabla I = \lambda_1\nabla\alpha + (1 - \lambda_2)\nabla F + \lambda_2\nabla B$$

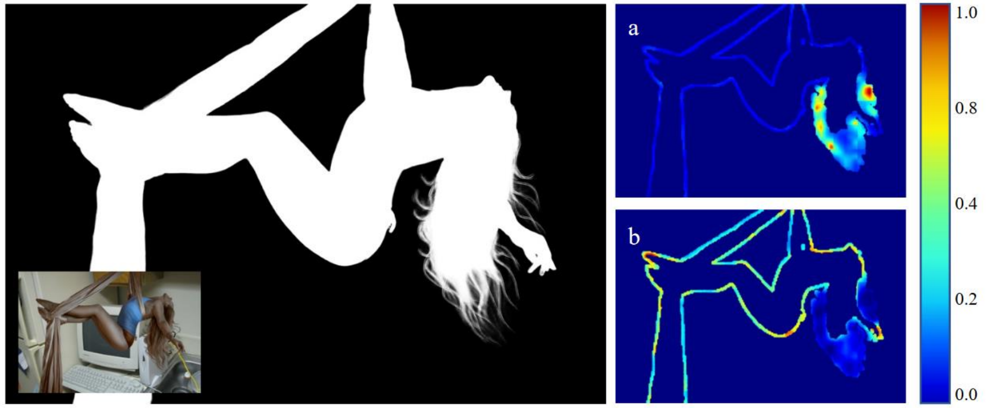


Figure 10. An example semantic trimap. We visualize two channels of the score maps: a. $hair\_hard$; b. $sharp$.

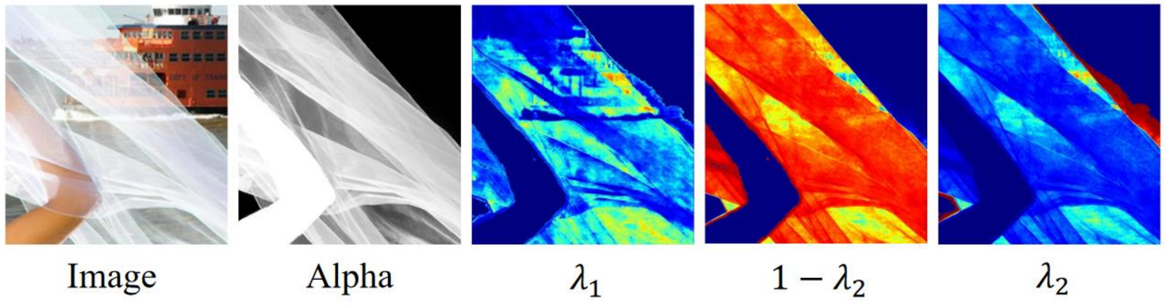


Figure 11. Visualization of learnable weights.