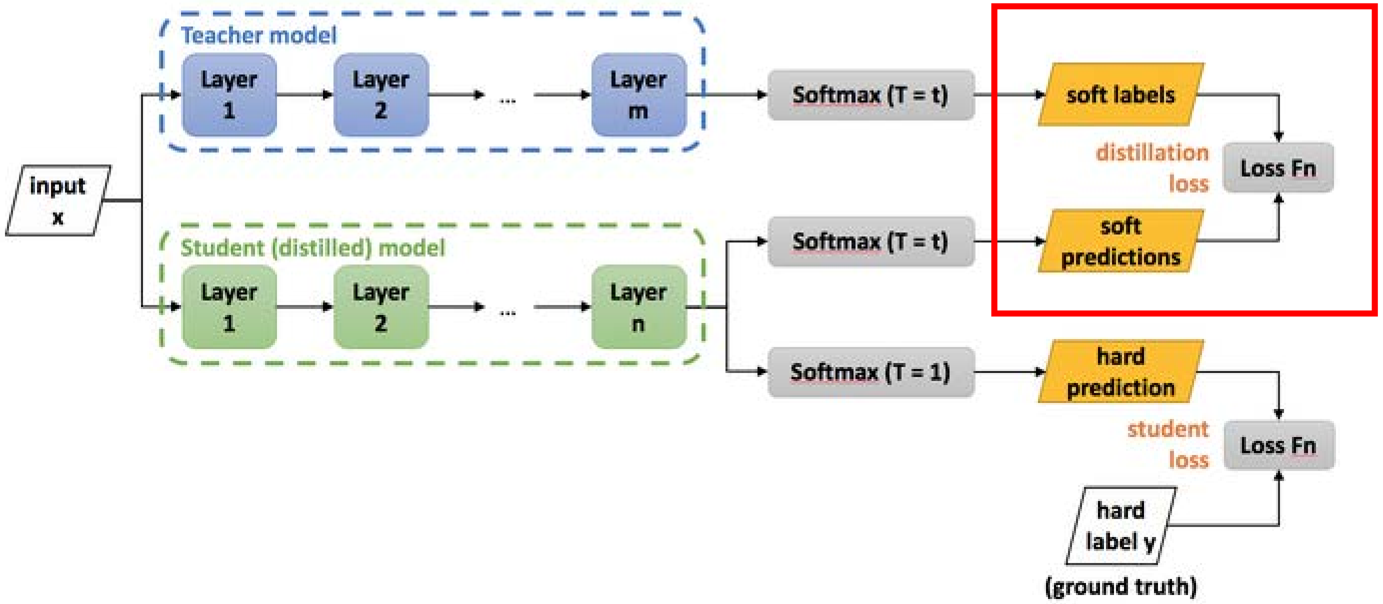
Lightweight Deep CNN for Natural Image Matting via Similarity-Preserving Knowledge Distillation (SPKD)

- Introduction
- Methods
- Experiment
- Reference

Introduction

Knowledge Distillation

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network.



0.08	0.72	0.04	0.02	0.06	0.08

Soft labels

0	1	0	0	0	0

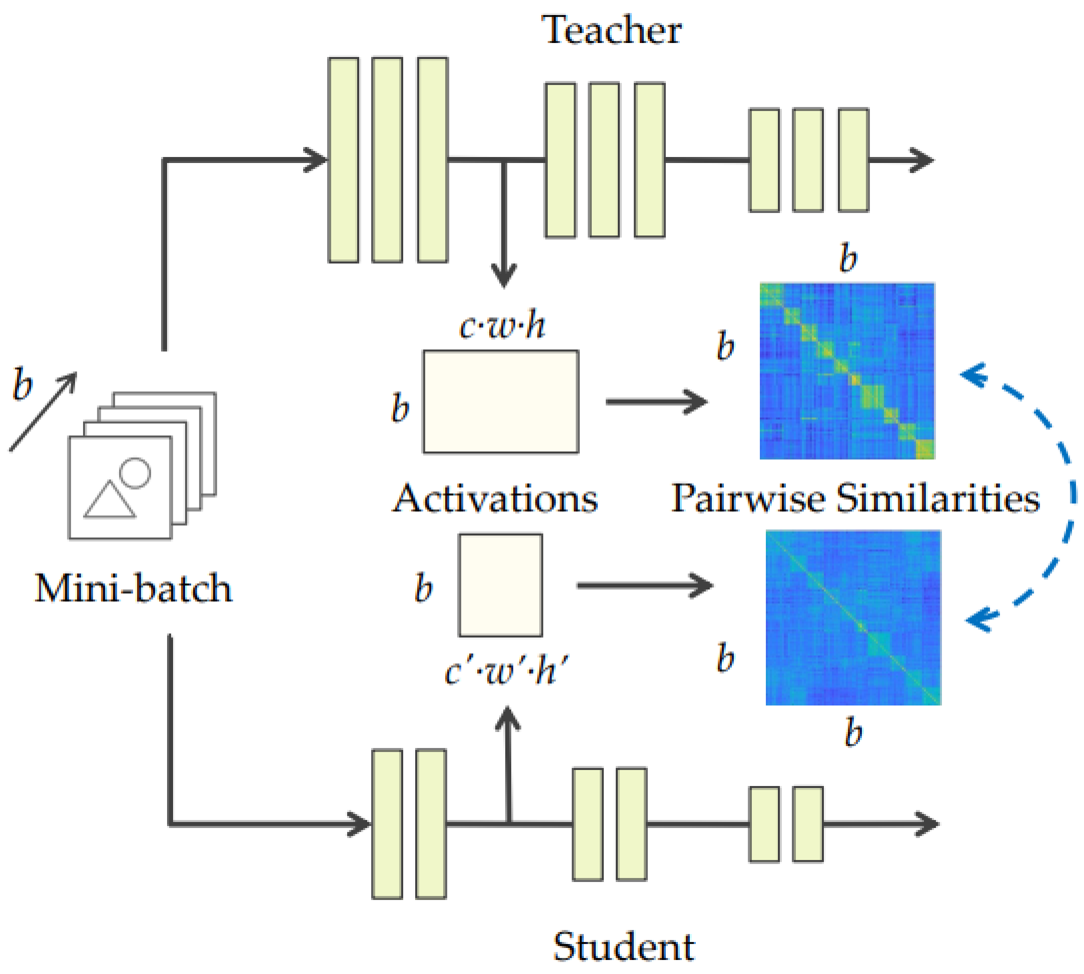
hard labels

Total loss = Cross entropy loss + KD loss

Introduction

Knowledge Distillation

F. Tung and G. Mori, (2019) Similarity-preserving knowledge distillation



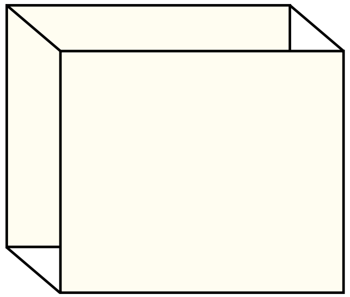
Total loss = Cross entropy loss + KD loss $\boxed{\text{+ Similarity loss}}$

Introduction

Knowledge Distillation

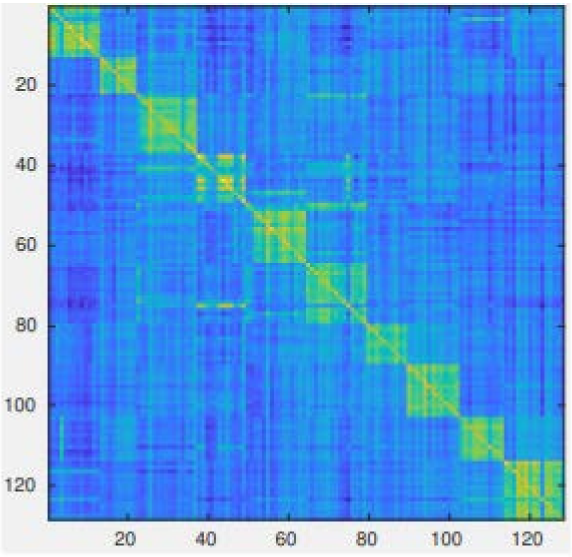
F. Tung and G. Mori, (2019) Similarity-preserving knowledge distillation

feature map



$b \times c \times w \times h$

$b \times cwh \quad \times \quad cwh \times b \quad =$

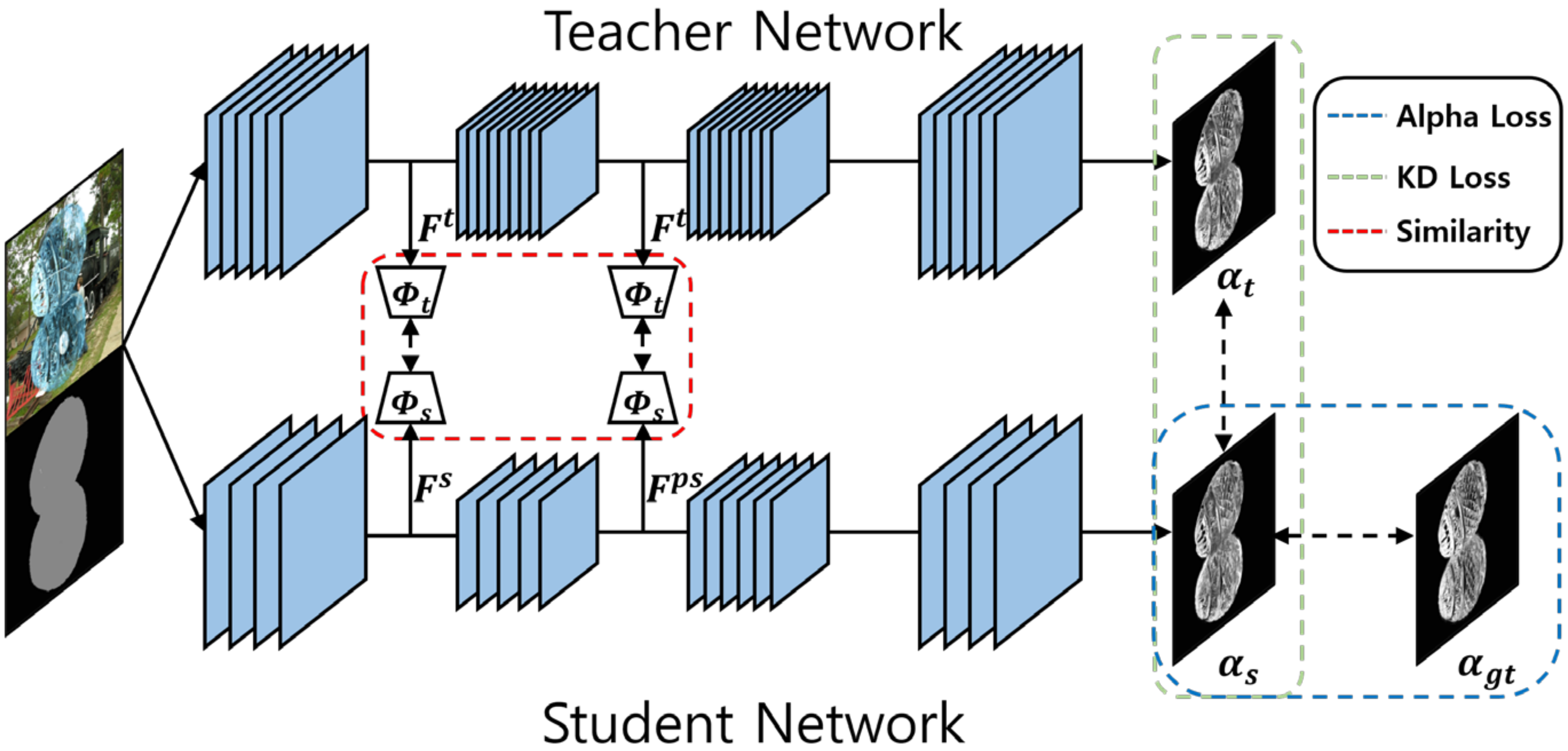


$b \times b$

Batch similarity

Methods

Overview



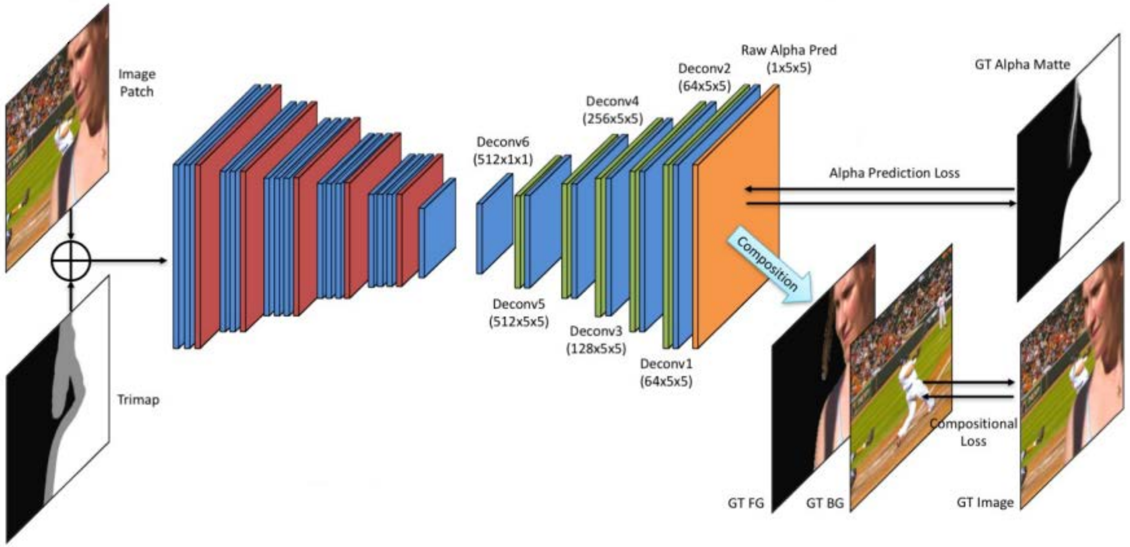
$$L_{overall} = w_1 L_a(\alpha_t, \alpha_s) + (1 - w_1) L_\alpha(\alpha_s, \alpha_{gt}) + w_2 \sum_{i=2}^{5} L_F(A_{ti}, A_{si})$$

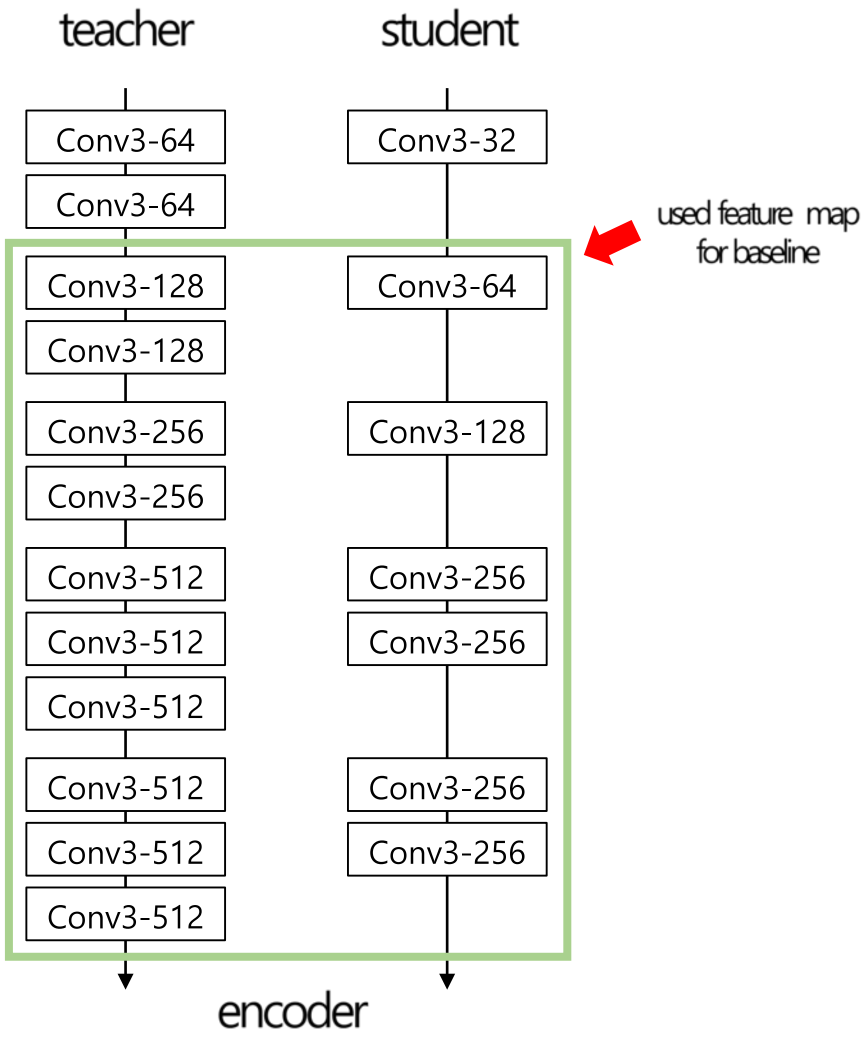
Methods

Baseline Network Architecture

Baseline Network Architecture : Deep image matting

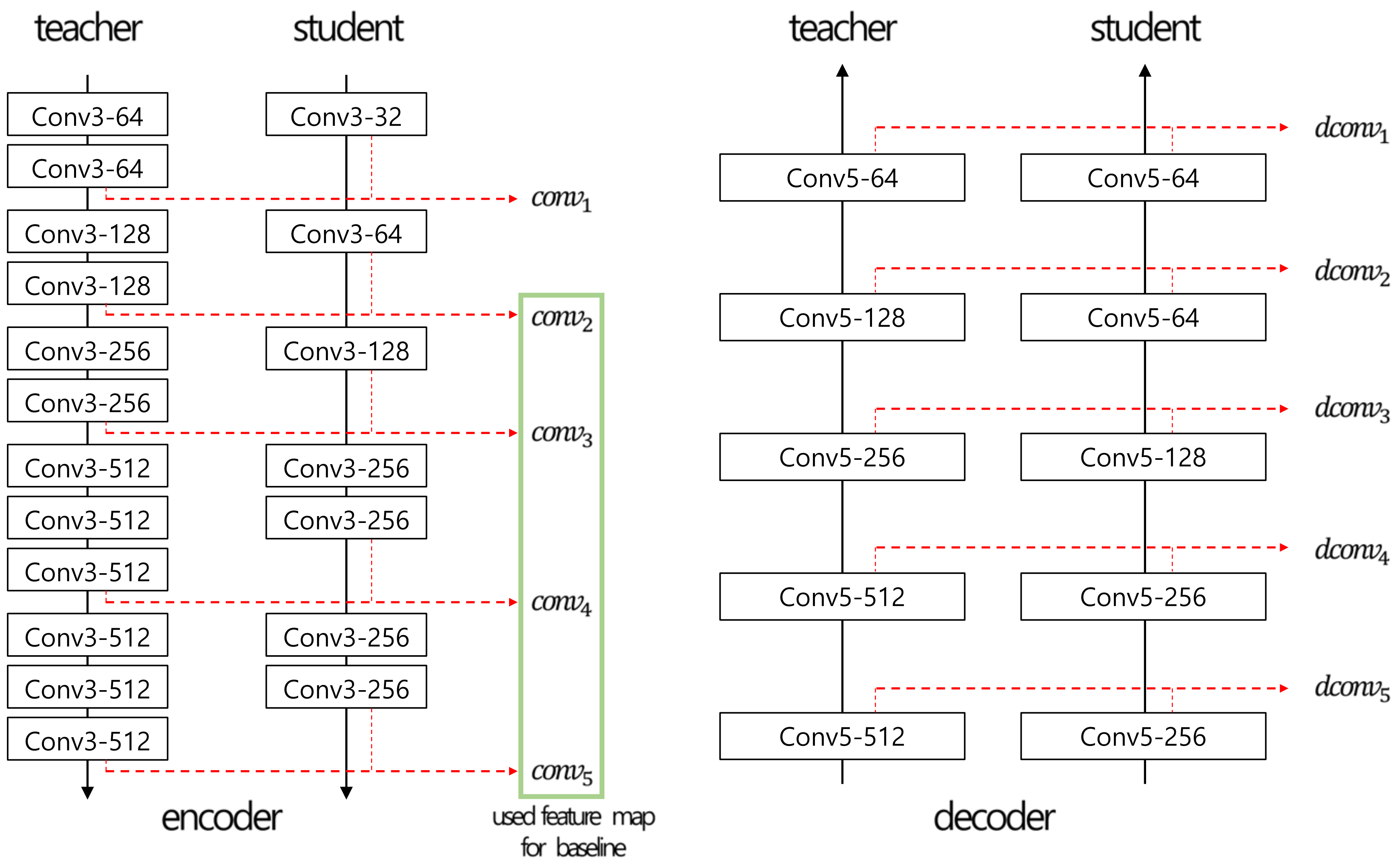
Student Network :
- The number of channels is half the number of teacher channels
- Reduction of layers in the encoder





Methods

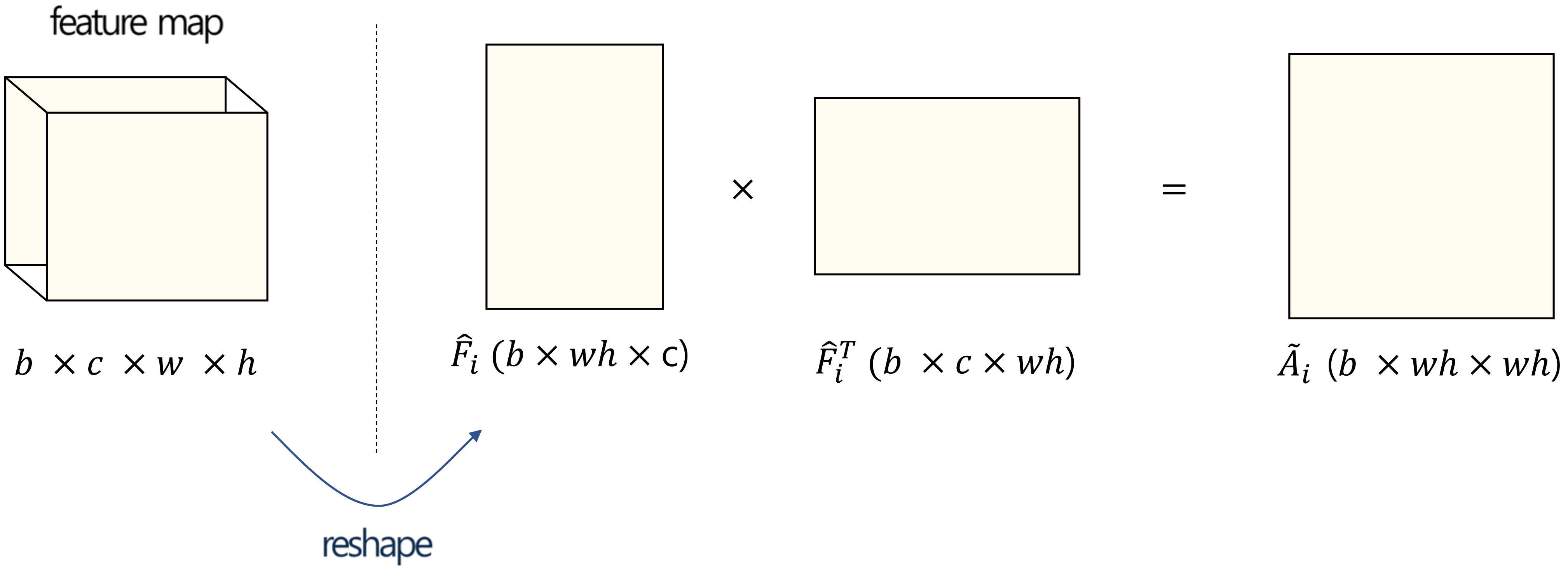
Baseline Network Architecture



Methods

Similarity-Preserving Knowledge Distillation

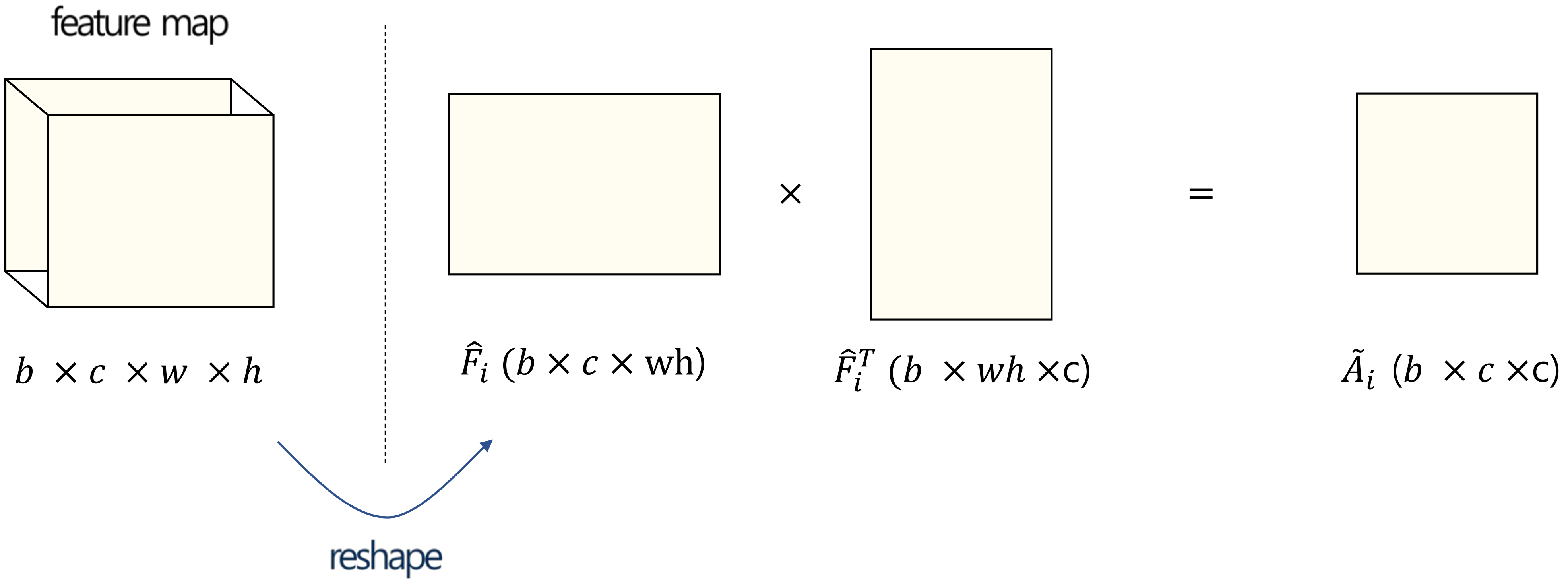
Spatial Similarity Map



Methods

Similarity-Preserving Knowledge Distillation

Channel Similarity Map



Methods

Loss

$$L_{overall} = w_1 L_a(\alpha_t, \alpha_s) + (1 - w_1) L_\alpha(\alpha_s, \alpha_{gt}) + w_2 \sum_{i=2}^{5} L_F(A_{ti}, A_{si})$$

w_1 , w_2: balancing parameters

L_a: alpha prediction loss

L_F: distillation loss ($MSE\ loss$)

A_{ti}: i-th teacher similarity map

α_t : teacher prediction

α_s : student prediction

α_{gt}: Ground Truth

A_{si}: i-th student similarity map

Experiment

Quantitative results

Quantitative evaluation on the Adobe-1k

Method	MSE	SAD	Gradient Error	Connectivity
DIM-teacher	0.021	65.37	33.20	67.58
DIM-student	0.058	121.77	75.36	129.55
DIM-Hinton [31]	0.055	120.32	69.56	128.19
DIM-AT [32]	0.052	117.00	74.90	125.60
DIM-FSP [33]	0.065	136.12	95.47	147.83
DIM-OFD [34]	0.042	94.47	54.56	100.66
DIM-batch	0.055	124.43	74.36	132.25
DIM-spatial	0.039	95.40	54.71	100.92
DIM-channel	0.038	94.76	56.36	100.36
DIM-spatial+channel	**0.034**	**84.37**	**47.63**	**89.35**
DIM-batch+spatial+channel	0.037	91.30	56.20	97.20

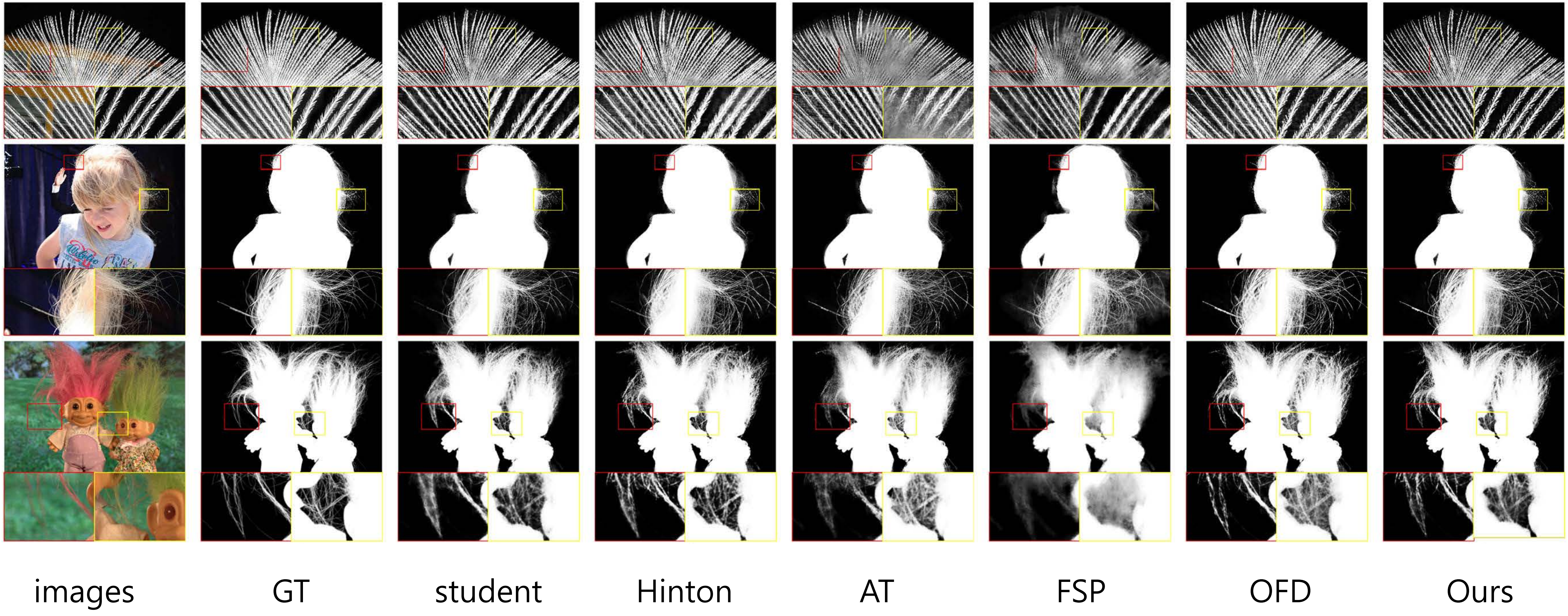
Quantitative evaluation on the alphamatting.com

Method	MSE	SAD	Gradient Error	Connectivity
DIM-teacher	0.009	3.85	3.81	3.25
DIM-student	0.020	5.91	7.90	5.47
DIM-Hinton [31]	0.021	6.02	8.72	5.69
DIM-AT [32]	0.022	6.23	9.03	5.89
DIM-FSP [33]	0.037	9.29	15.65	9.06
DIM-OFD [34]	**0.016**	5.39	**6.51**	4.99
DIM-batch	0.020	5.92	8.01	5.49
DIM-spatial	0.017	5.50	7.22	5.04
DIM-channel	0.018	5.47	7.74	5.13
DIM-spatial+channel	**0.016**	**5.23**	7.19	**4.74**
DIM-batch+spatial+channel	0.018	5.60	7.64	5.21

Experiment

Qualitative results

Qualitative evaluation on the Adobe-1k



Experiment

Ablation Studies

Impact on Layers for Distillation

Method	MSE	SAD	Gradient Error	Connectivity
$\mathrm{conv}_{2\text{-}5}$	0.034	84.37	47.63	89.35
$\mathrm{conv}_{3\text{-}5}$ + dconv_{5}	0.041	98.25	58.90	104.80
$\mathrm{conv}_{4\text{-}5}$ + $\mathrm{dconv}_{5\text{-}4}$	0.042	95.98	62.44	102.08
conv_{5} + $\mathrm{dconv}_{5\text{-}3}$	0.050	115.33	72.49	124.66

Best : baseline

Experiment

Ablation Studies

Impact on Baseline Backbone Network

Method	MSE	SAD	Gradient Error	Connectivity
IndexNet-teacher [25]	0.013	45.63	27.28	43.80
GCAM-teacher [27]	0.009	34.94	17.54	30.92
IndexNet-student	0.027	65.21	42.56	66.83
IndexNet-spatial+channel	0.019	55.74	33.78	55.85
GCAM-student	0.017	52.61	34.27	46.24
GCAM-spatial+channel	0.015	48.06	30.28	42.81

The number of Parameters in the network

	Teacher	Student
DIM	25,582,595	5,181,795
IndexNet	5,953,515	3,509,879
GCAM	25,269,144	6,394,224

- DIM (50%)
- IndexNet (75%)
- GCAM (50%)

(-%) is the remaining channel ratio

Reference

- N. Xu, B. L. Price, S. Cohen, and T. S. Huang, “Deep image matting,” in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 311–320.

- H. Lu, Y. Dai, C. Shen, and S. Xu, “Indices matter: Learning to index for deep image matting,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3266–3275.

- Y. Li and H. Lu, “Natural image matting via guided contextual attention,” in *Proc. Assoc. Adv. Artif. Intell.*, 2020, pp. 13673–13682.

- G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. Neural Inf. Process. Syst. Workshops (NeurIPSW)*, 2015, pp. 1–9.

- F. Tung and G. Mori, “Similarity-preserving knowledge distillation,” in *Proc. Int. Conf. Comput. Vis*, 2019, pp. 1365–1374.

- S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in Proc. Int. Conf. Learn. Representation, 2017, pp. 1–13.

- J. Yim, D. Joo, J. Bae, and J. Kim, “A gift from knowledge distillation: Fast optimization, network minimization and transfer learning,” in Proc. Comput. Vis. Pattern Recognit., 2017, pp. 7130–7138.

- B. Heo, J. Kim, S.Yun, H. Park,N.Kwak, and J.Y.Choi, “Acomprehensive overhaul of feature distillation,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 1921–1930.