Webly Supervised Semantic Segmentation

Bin Jin[1], Maria V. Ortiz-Segovia[2], Sabine Süsstrunk[1]

[1]IC, EPFL [2]Océ Print Logic Technologies

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

océ A CANON COMPANY

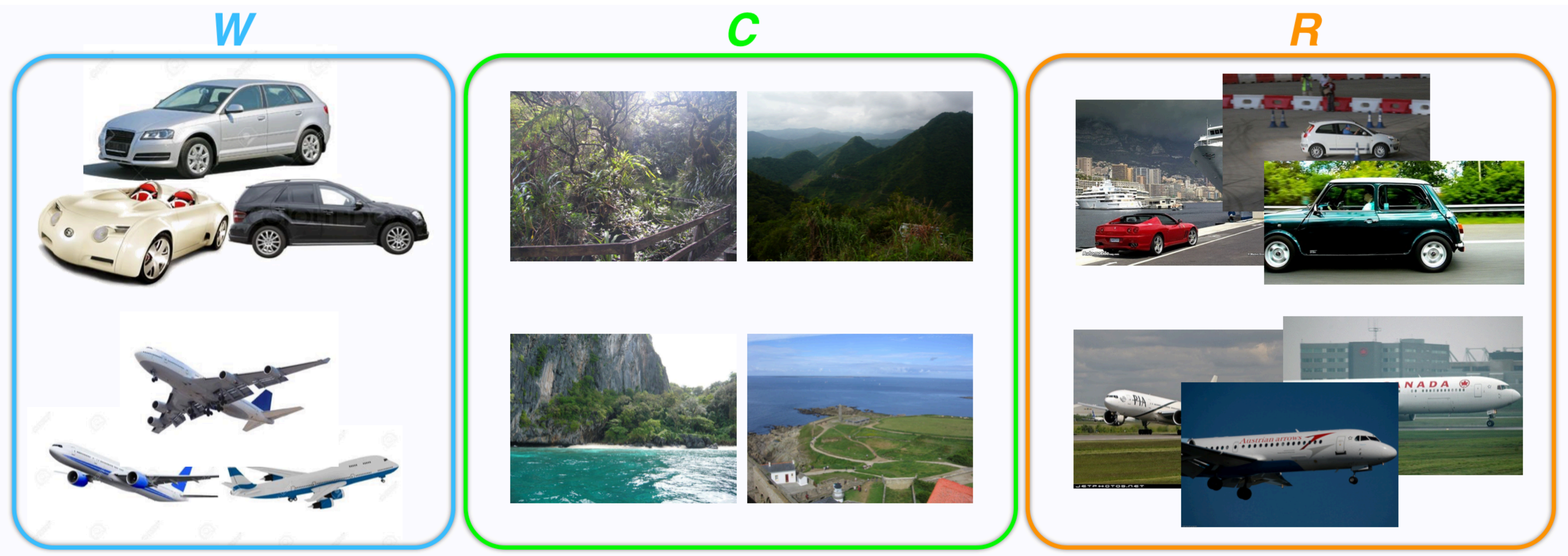
Paper and code available at:

Motivation:

The pixel-wise annotation of images to obtain accurate semantic segmentation ground-truth is both expensive and time-consuming. By exploiting the vast collection of labeled web images with rich context, we can bypass this tedious task. Our webly supervised semantic segmentation outperforms the state-of-the-art weakly supervised segmentation methods by a significant margin.

Web images:

From these websites, we collect three sets of web images as training data:



—{*W*}: a white background set, built by querying the text-based image search engine, e.g., Google or Microsoft Bing, with the query "<class> on white background".

—{*C*}: a common background set, built by retrieving images from image sharing websites, e.g., Flickr or Imgur1, with common background keywords.

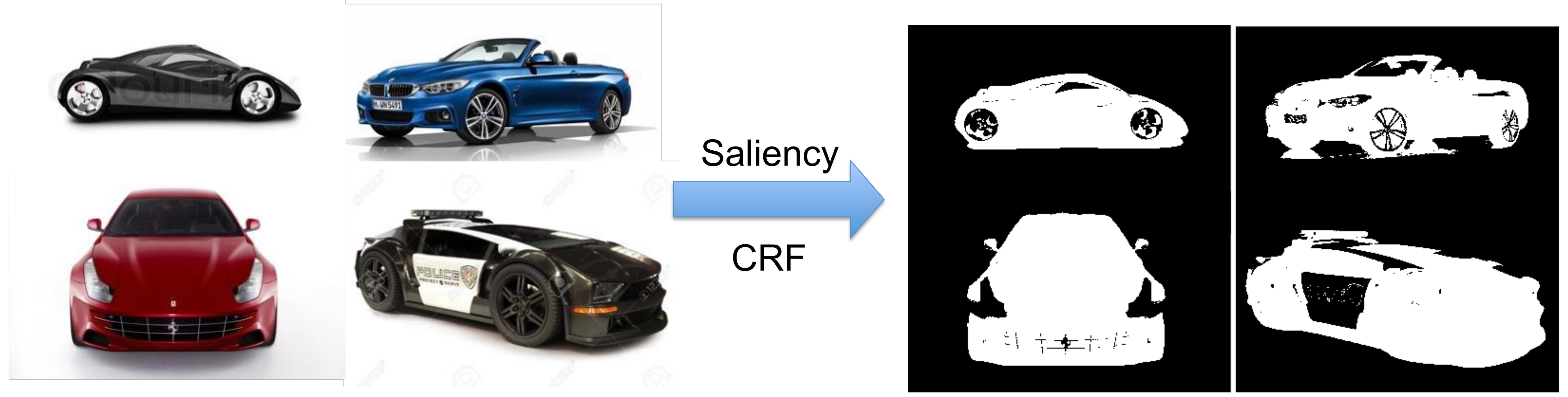
—{*R*}: a realistic images set, constructed by crawling image sharing websites with the given class name or using existing datasets.

References:

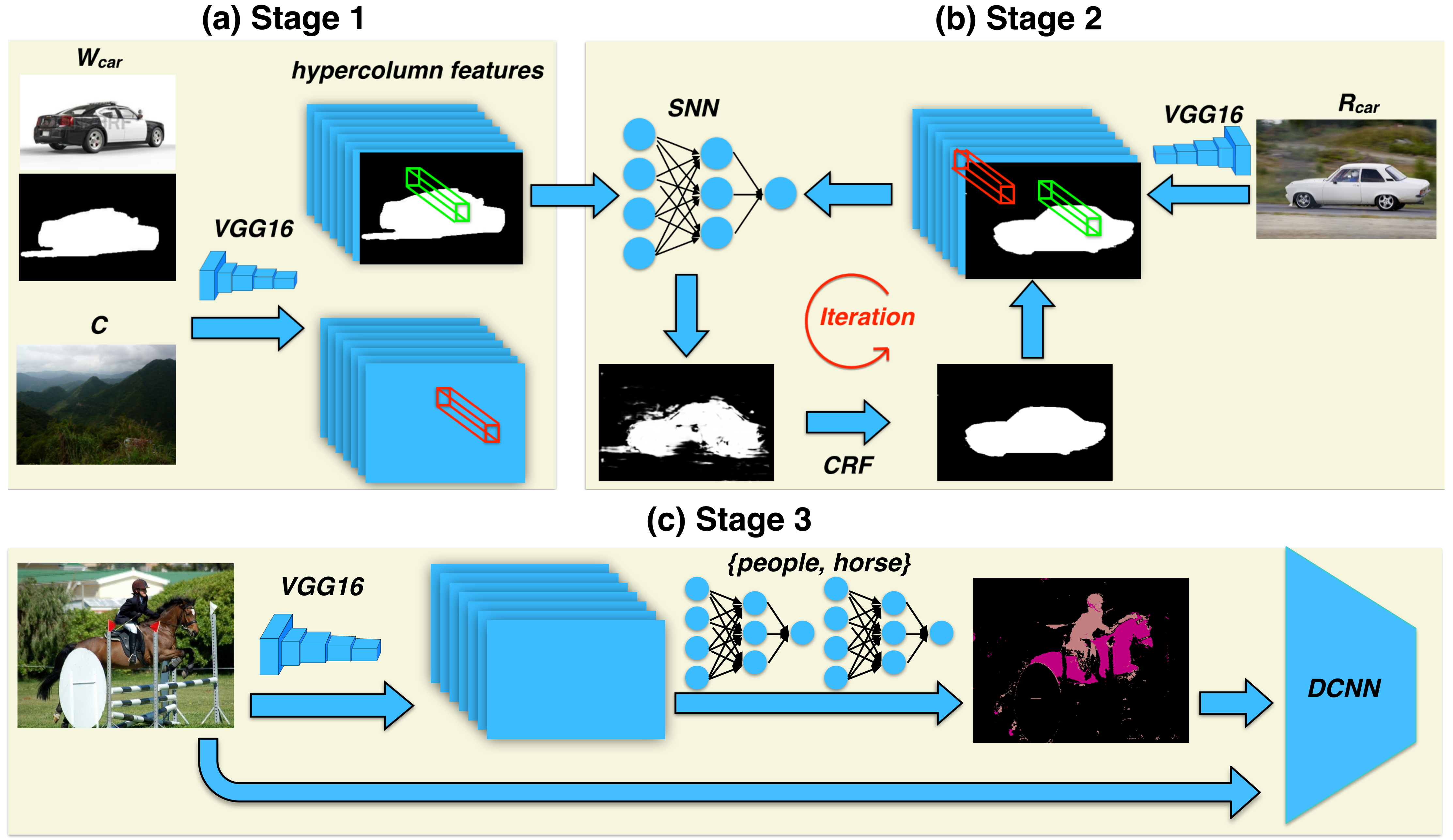
[1] C. Yang et al. "Saliency detection via graph-based manifold ranking," in CVPR, 2013.
[2] P. Krähenbühl et al. "Efficient inference in fully connected CRFs with Gaussian edge potentials," in NIPS, 2011.
[3] D. Pathak, et al. "Constrained convolutional neural networks for weakly supervised segmentation," in ICCV, 2015.
[4] W. Shimoda et al. "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in ECCV, 2016.
[5] F. Saleh, et al. "Built-in foreground/background prior for weakly-supervised semantic segmentation," in ECCV, 2016.
[6] Y. Wei et al. "STC: A simple to complex framework for weakly-supervised semantic segmentation," TPAMI, 2016.
[7] A. Kolesnikov et al. "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in ECCV, 2016.

Methods:

Images in {*W*} are first segmented with a saliency algorithm[1] combined with dense CRF[2].



We then train a semantic segmentation network in a three-stage pipeline:



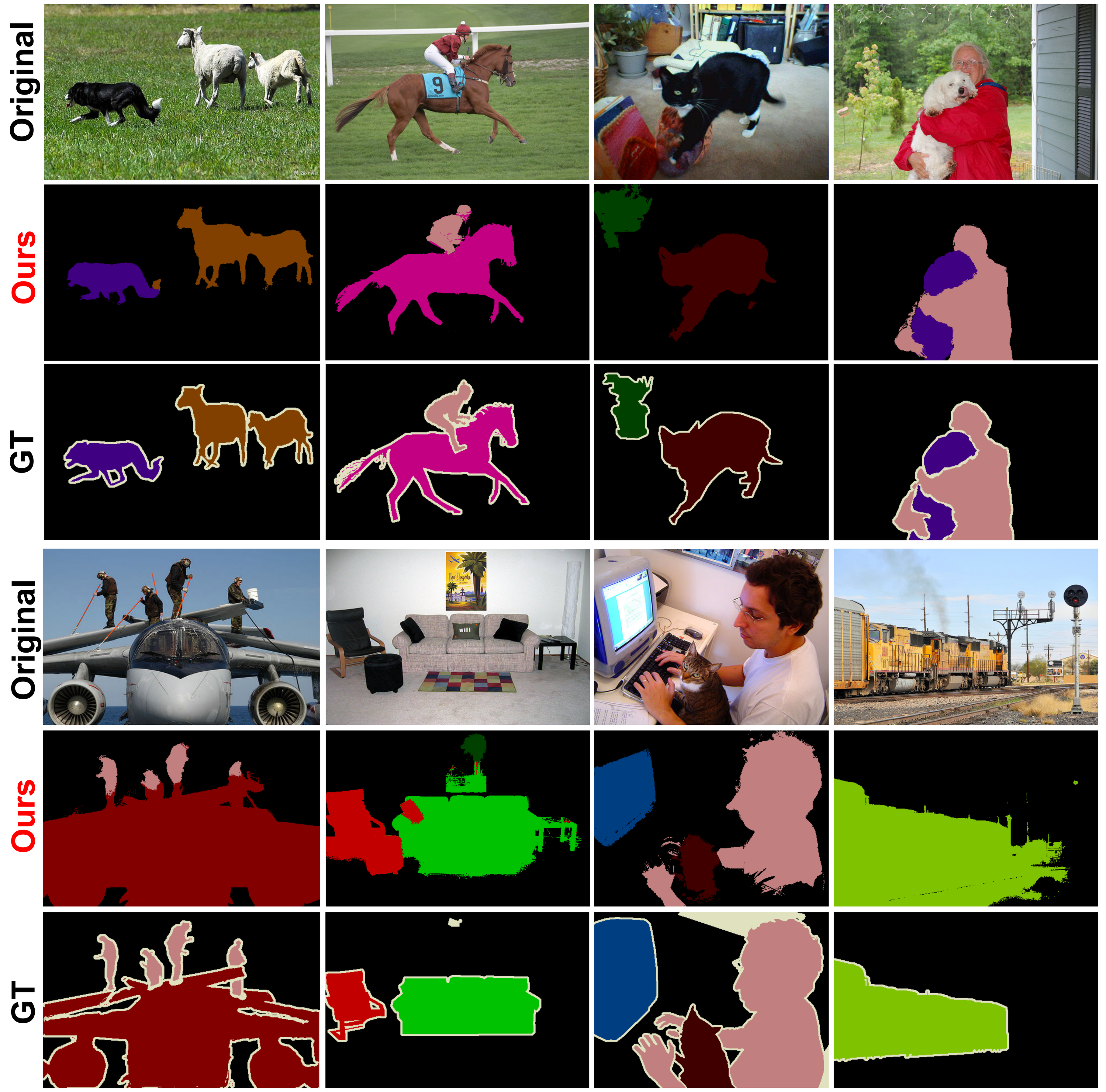
—**Stage 1**: we train a Shallow Neural Network (SNN) for each class to output class-specific segmentation masks, using the hypercolumn features from a pre-trained network.

—**Stage 2**: we iteratively refine the SNNs based on the realistic images in {*R*}. A Conditional Random Field (CRF) is applied during each iteration.

—**Stage 3**: we assemble all SNNs into one deep convolution neural network (DCNN) by training the DCNN end-to-end with the segmentation masks generated by the SNNs.

Results:

Dataset: For training, we collect 6807 white background images for {*W*}, 1491 images for {*C*} and 10,582 images for {*R*}. We evaluate the performance on the PASCAL VOC 2012 segmentation benchmark with the *Intersection over Union* (IoU) metric.



IoU: Our method produces excellent results on both the validation and test set, outperforming all state-of-the-art weakly-supervised semantic segmentation methods.

Method	CCCN[3]	DCSM[4]	BFBP[5]	STC[6]	SEC[7]	OURS
Validation set	35.3	44.1	46.6	49.8	50.7	53.4
Test set	35.6	45.1	48.0	51.2	51.7	55.3