Effective Use of Synthetic Data for Urban Scene Semantic Segmentation

European Conference on Computer Vision (ECCV), 2018

F. Saleh[1,2], M. Aliakbarian[1,2,3], M. Salzmann[4], L. Petersson[2], J. Alvarez[5]

{[1]ANU, [2]Data61-CSIRO, [3]ACRV}, Canberra, Australia

[4]CVLab, EPFL, Lausanne, Switzerland

[5]NVIDIA, Santa Clara, USA

Presented by: Ricardo B. Sousa (up201503004) FEUP, PDEEC, Computer Vision, 2020/2021

Outline

- Introduction
- Related Work
- Method
- VEIS: Virtual Environment for Instance Segmentation
- Experiments
- Conclusions

INTRODUCTION: CONTEXT

- Semantic segmentation
  - Images → Regions (limited number of classes)
  - The main idea is to recognise and understand what's in the image in pixel level
  - Applications: robot vision, understanding, autonomous driving, medical purposes
- Deep networks have proven highly effective to perform semantic segmentation

- **Problem:** deep networks require vast amounts of labelled data
  - Pixel labelling 1 image of the Cityscapes [1] (real data) dataset takes 90min on average

Introduction: Motivation

- Advances on computer graphics allow the generation of synthetic datasets
  - Datasets such as GTA5 [2], VYPER [3] or SYNTHIA [4]

- Domain adaptation methods require the access to large sets of real data (even though unsupervised)
  - Account the domain shift between real and synthetic data
  - Cannot deploy a model trained off-line on synthetic data in a new real-world environment

Introduction: Goals

- **Assumption:** texture of background classes in synthetic images look more realistic than foreground classes

- Treat the two different kinds of classes differently [0]
  - Background classes: semantic segmentation
  - Foreground classes: detector-based approach

- Creation of the Virtual Environment for Instance Segmentation (VEIS) [0]
  - Existent synthetic datasets does not provide instance-level annotations for all the foreground classes of standard real datasets (e.g., Cityshapes [1] or CamVid [5])
  - Automatic instance-level annotation
  - Created using the game engine Unity

Related Work

- **Semantic segmentation:** understanding an image at pixel-level
  - The most recent techniques rely on deep networks

- **Problem:** fully-supervised data for semantic segmentation with pixel-level annotations is very expensive and time-consuming

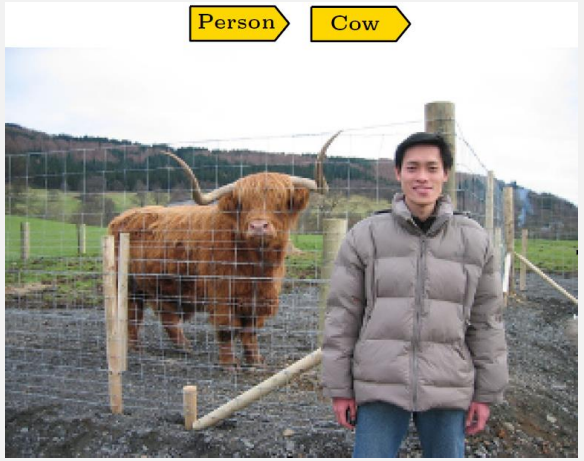
2 trends

Weakly-supervised methods

Use of synthetic data

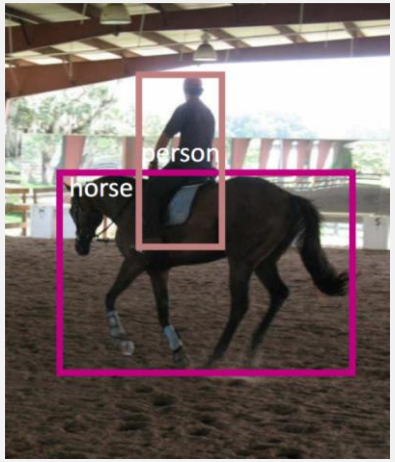
Related Work: Weakly-Supervised Semantic Segmentation

- Image tags



Source: Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation (2016) [6]

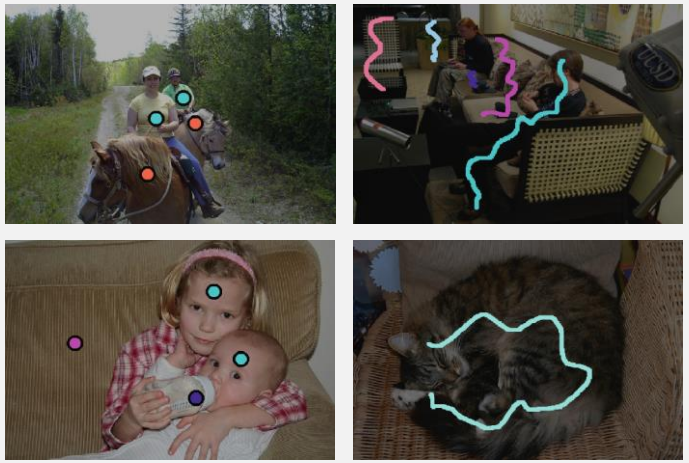
- Bounding boxes



Source: Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation (2015) [7]

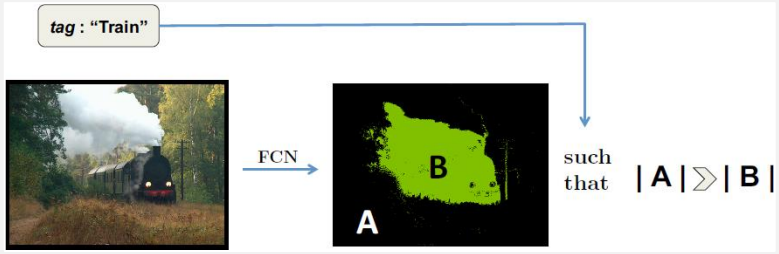
Related Work: Weakly-Supervised Semantic Segmentation

- Points and scribbles



Source: Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision (2016) [8]

- Object size statistics



Source: Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation (2015) [9]

Related Work: Weakly-Supervised Semantic Segmentation

- **Problem:** most of existent methods focus on foreground classes
  - Treat the background as a single entity

- Background classes are important for scenarios such as automated driving
  - Example: differentiating road from grass

- Only few approaches, such as Saleh et al. [10], consider multiple background classes with weakly-supervisor semantic segmentation
- Still a huge gap compared to fully-supervised methods (especially in the foreground classes)

Related Work: Synthetic Data

- Generate fully-supervised synthetic data
  - Examples of datasets: GTA5 [2], VIPER [3], and SYNTHIA [4]

- **Problem:** domain shift between real and synthetic data

- Domain adaptation methods are applied to reduce the gap between the feature distributions of the two domains
  - Curriculum style learning to align label distribution over both entire images and superpixels [11]
  - Feature regularizer based on the notion of distillation to align the distributions of source and target [12]
  - Generative approach with cycle consistency to adapt pixel and feature-level representations [13]

- **Requirement:** have access to real images during training (even though without supervision)

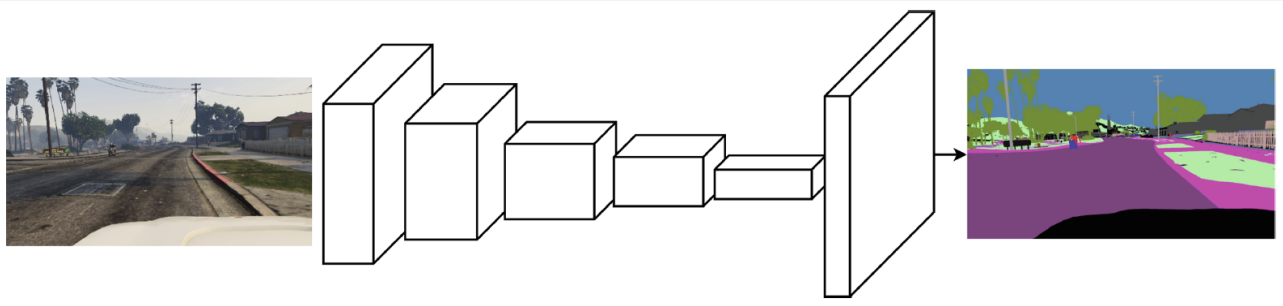
Method

- **Background Classes**
  - VGG16-based DeepLab model (w/ large field of view and dilated convolution layers) [14]
  - Model trained on the GTA5 [2] dataset (background classes look photo-realistic)

- **Foreground Classes**
  - Mask R-CNN (detection-based instance-level semantic segmentation) [15]
    - 1. Initial object detection stage (bounding box)
    - 2. Binary mask extraction + object classification
  - Model trained with the author's own dataset (retrieved from VEIS)

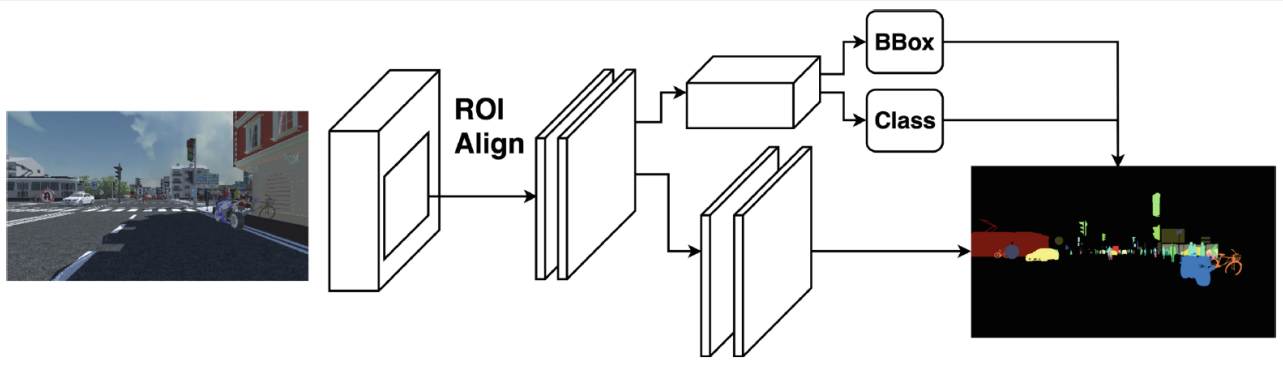
Method

- **Background Classes**



Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

- **Foreground Classes**



Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

METHOD

- **Prediction on Real Images**
  - The networks for the 2 different types of classes are trained using **only** synthetic data
  - **Goal:** pixel-level segmentation
  - Non-maximum suppression-based approach (starting with the Mask R-CNN [15] predictions):
    - 1. Sort the predicted segments according to their confidence scores
    - 2. Iterate over the sorted list from the higher to the lowest confidence score
      - IF (*current segment overlaps with a previous one*) THEN *remove the pixels in the overlapping region*
  - The prediction given by the DeepLab [14] network fill the remaining holes with a similar approach
    - Every pixel that is not already assigned to a foreground class takes the label with highest probability at that pixel location in the DeepLab [14] result

METHOD

- **Leveraging Unsupervised Real Images**
  - The predictions obtained with the proposed method are used as pseudo ground-truth labels for the real images
  - **Modification:** pixels predicted as foreground classes by the DeepLab [14] model are set to an **ignore** label
    - Standard segmentation network are not reliable for foreground classes when trained only on synthetic data
  - Train DeepLab [14] segmentation network with the pseudo ground-truth labels

VEIS: Virtual Environment for Instance Segmentation

- **Current synthetic datasets have drawbacks** (especially in terms of foreground classes)
  - GTA5 [2] does not have instance-level annotations for foreground classes
  - VIPER [3] and SYNTHIA [4] have instance-level annotations, but not for all foreground classes of standard real datasets (e.g., Cityshapes [1] or CamVid [5])

- **Virtual Environment for Instance Segmentation (VEIS)**
  - Uses the Unity3D game engine
    - Easy to generate annotations automatically (instance-level pixel-wise)
    - Easy to set up
  - Less photo-realistic than GTA5 [2] and VIPER [3] → However, VEIS is used for the foreground classes and it has a larger diversity of foreground objects
  - 61305 frames w/ instance-level semantic segmentation (obtained with no human intervention)

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

		road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
DeepLab	GTA5	**80.5**	26.0	74.7	23.0	9.8	9.1	13.4	7.3	79.4	28.6	72.1	40.4	5.1	77.8	23.0	18.6	1.2	5.3	0.0	31.3
	SYNTHIA	36.7	22.7	51.0	0.3	0.1	**16.6**	0.1	9.5	72.5	0.0	78.4	47.5	5.6	61.4	0.0	13.0	0.0	3.2	3.1	22.1
	VIPER	36.9	19.0	74.7	0.0	5.3	7.1	10.0	10.1	78.7	13.6	69.6	43.0	0.0	41.2	20.8	13.9	0.0	9.1	0.0	23.9
	VEIS	70.8	9.5	50.9	0.0	0.0	0.3	15.6	26.8	66.8	12.7	52.3	44.0	14.2	60.6	10.2	8.2	3.2	5.5	11.8	24.4
	GTA5 + VEIS	66.2	21.6	72.3	15.7	18.3	12.3	22.3	23.8	78.4	11.3	74.6	48.7	13.3	75.1	14.3	21.2	2.1	**24.2**	7.3	32.8
	GTA5 + VEIS & ps-gt	77.6	26.8	75.5	19.4	19.5	4.8	18.7	19.8	79.5	21.7	**78.9**	47.3	8.7	77.6	23.1	16.1	2.2	15.6	0.0	33.3
[0]	[0]	71.9	23.8	75.5	23.4	14.9	9.3	**26.7**	42.5	80.1	34.0	76.3	52.2	**28.5**	76.2	19.6	31.6	6.9	18.1	9.8	38.0
	[0] & ps-gt	79.8	**29.3**	**77.8**	**24.2**	**21.6**	6.9	23.5	**44.2**	**80.5**	**38.0**	76.2	**52.7**	22.2	**83.0**	**32.3**	**41.3**	**27.0**	19.3	**27.7**	**42.5**

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

		road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
DeepLab	GTA5	**80.5**	26.0	74.7	23.0	9.8	9.1	13.4	7.3	79.4	28.6	72.1	40.4	5.1	77.8	23.0	18.6	1.2	5.3	0.0	31.3
	SYNTHIA	36.7	22.7	51.0	0.3	0.1	**16.6**	0.1	9.5	72.5	0.0	78.4	47.5	5.6	61.4	0.0	13.0	0.0	3.2	3.1	22.1
	VIPER	36.9	19.0	74.7	0.0	5.3	7.1	10.0	10.1	78.7	13.6	69.6	43.0	0.0	41.2	20.8	13.9	0.0	9.1	0.0	23.9
	VEIS	70.8	9.5	50.9	0.0	0.0	0.3	15.6	26.8	66.8	12.7	52.3	44.0	14.2	60.6	10.2	8.2	3.2	5.5	11.8	24.4
	GTA5 + VEIS	66.2	21.6	72.3	15.7	18.3	12.3	22.3	23.8	78.4	11.3	74.6	48.7	13.3	75.1	14.3	21.2	2.1	**24.2**	7.3	32.8
	GTA5 + VEIS & ps-gt	77.6	26.8	75.5	19.4	19.5	4.8	18.7	19.8	79.5	21.7	**78.9**	47.3	8.7	77.6	23.1	16.1	2.2	15.6	0.0	33.3
[0]	[0]	71.9	23.8	75.5	23.4	14.9	9.3	**26.7**	42.5	80.1	34.0	76.3	52.2	**28.5**	76.2	19.6	31.6	6.9	18.1	9.8	38.0
	[0] & ps-gt	79.8	**29.3**	**77.8**	**24.2**	**21.6**	6.9	23.5	**44.2**	**80.5**	**38.0**	76.2	**52.7**	22.2	**83.0**	**32.3**	**41.3**	**27.0**	19.3	**27.7**	**42.5**

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

		road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
DeepLab	GTA5	**80.5**	26.0	74.7	23.0	9.8	9.1	13.4	7.3	79.4	28.6	72.1	40.4	5.1	77.8	23.0	18.6	1.2	5.3	0.0	31.3
	SYNTHIA	36.7	22.7	51.0	0.3	0.1	**16.6**	0.1	9.5	72.5	0.0	78.4	47.5	5.6	61.4	0.0	13.0	0.0	3.2	3.1	22.1
	VIPER	36.9	19.0	74.7	0.0	5.3	7.1	10.0	10.1	78.7	13.6	69.6	43.0	0.0	41.2	20.8	13.9	0.0	9.1	0.0	23.9
	VEIS	70.8	9.5	50.9	0.0	0.0	0.3	15.6	26.8	66.8	12.7	52.3	44.0	14.2	60.6	10.2	8.2	3.2	5.5	11.8	24.4
	GTA5 + VEIS	66.2	21.6	72.3	15.7	18.3	12.3	22.3	23.8	78.4	11.3	74.6	48.7	13.3	75.1	14.3	21.2	2.1	**24.2**	7.3	32.8
	GTA5 + VEIS & ps-gt	77.6	26.8	75.5	19.4	19.5	4.8	18.7	19.8	79.5	21.7	**78.9**	47.3	8.7	77.6	23.1	16.1	2.2	15.6	0.0	33.3
[0]	[0]	71.9	23.8	75.5	23.4	14.9	9.3	**26.7**	42.5	80.1	34.0	76.3	52.2	**28.5**	76.2	19.6	31.6	6.9	18.1	9.8	38.0
	[0] & ps-gt	79.8	**29.3**	**77.8**	**24.2**	**21.6**	6.9	23.5	**44.2**	**80.5**	**38.0**	76.2	**52.7**	22.2	**83.0**	**32.3**	**41.3**	**27.0**	19.3	**27.7**	**42.5**

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

		road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
DeepLab	GTA5	**80.5**	26.0	74.7	23.0	9.8	9.1	13.4	7.3	79.4	28.6	72.1	40.4	5.1	77.8	23.0	18.6	1.2	5.3	0.0	31.3
	SYNTHIA	36.7	22.7	51.0	0.3	0.1	**16.6**	0.1	9.5	72.5	0.0	78.4	47.5	5.6	61.4	0.0	13.0	0.0	3.2	3.1	22.1
	VIPER	36.9	19.0	74.7	0.0	5.3	7.1	10.0	10.1	78.7	13.6	69.6	43.0	0.0	41.2	20.8	13.9	0.0	9.1	0.0	23.9
	VEIS	70.8	9.5	50.9	0.0	0.0	0.3	15.6	26.8	66.8	12.7	52.3	44.0	14.2	60.6	10.2	8.2	3.2	5.5	11.8	24.4
	GTA5 + VEIS	66.2	21.6	72.3	15.7	18.3	12.3	22.3	23.8	78.4	11.3	74.6	48.7	13.3	75.1	14.3	21.2	2.1	**24.2**	7.3	32.8
	GTA5 + VEIS & ps-gt	77.6	26.8	75.5	19.4	19.5	4.8	18.7	19.8	79.5	21.7	**78.9**	47.3	8.7	77.6	23.1	16.1	2.2	15.6	0.0	33.3
[0]	[0]	71.9	23.8	75.5	23.4	14.9	9.3	**26.7**	42.5	80.1	34.0	76.3	52.2	**28.5**	76.2	19.6	31.6	6.9	18.1	9.8	38.0
	[0] & ps-gt	79.8	**29.3**	**77.8**	**24.2**	**21.6**	6.9	23.5	**44.2**	**80.5**	**38.0**	76.2	**52.7**	22.2	**83.0**	**32.3**	**41.3**	**27.0**	19.3	**27.7**	**42.5**

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Fully Sup. DeepLab*	**95.8**	**70.4**	85.4	42.7	41.0	21.2	**33.7**	**44.8**	**86.2**	51.4	88.4	58.1	**30.1**	86.4	43.8	56.7	42.8	33.9	54.8	56.2
Fully Sup. [0]*	95.6	70.1	**86.1**	**43.8**	**41.4**	16.6	31.3	43.3	85.9	**52.0**	**89.6**	**67.0**	29.9	**87.7**	**61.8**	**72.7**	**53.1**	**50.8**	**60.5**	**60.0**
Weakly Sup. [10]*	75.9	1.5	41.7	14.1	15.3	6.3	4.4	7.7	58.4	12.6	56.2	16.2	6.1	41.2	22.7	16.6	20.4	15.7	14.9	23.6
[11]	74.8	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
[12]	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
[13]	85.2	37.2	76.5	21.8	15.0	**23.8**	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
[0]	71.9	23.8	75.5	23.4	14.9	9.3	26.7	42.5	80.1	34.0	76.3	52.2	28.5	76.2	19.6	31.6	6.9	18.1	9.8	38.0
[0] & ps-gt	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

(*) Methods were trained on the dataset Cityshapes and evaluated on the Cityshapes's validation set

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Fully Sup. DeepLab*	**95.8**	**70.4**	85.4	42.7	41.0	21.2	**33.7**	**44.8**	**86.2**	51.4	88.4	58.1	**30.1**	86.4	43.8	56.7	42.8	33.9	54.8	56.2
Fully Sup. [0]*	95.6	70.1	**86.1**	**43.8**	**41.4**	16.6	31.3	43.3	85.9	**52.0**	**89.6**	**67.0**	29.9	**87.7**	**61.8**	**72.7**	**53.1**	**50.8**	**60.5**	**60.0**
Weakly Sup. [10]*	75.9	1.5	41.7	14.1	15.3	6.3	4.4	7.7	58.4	12.6	56.2	16.2	6.1	41.2	22.7	16.6	20.4	15.7	14.9	23.6
[11]	74.8	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
[12]	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
[13]	85.2	37.2	76.5	21.8	15.0	**23.8**	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
[0]	71.9	23.8	75.5	23.4	14.9	9.3	26.7	42.5	80.1	34.0	76.3	52.2	28.5	76.2	19.6	31.6	6.9	18.1	9.8	38.0
[0] & ps-gt	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

(*) Methods were trained on the dataset Cityshapes and evaluated on the Cityshapes's validation set

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Fully Sup. DeepLab*	**95.8**	**70.4**	85.4	42.7	41.0	21.2	**33.7**	**44.8**	**86.2**	51.4	88.4	58.1	**30.1**	86.4	43.8	56.7	42.8	33.9	54.8	56.2
Fully Sup. [0]*	95.6	70.1	**86.1**	**43.8**	**41.4**	16.6	31.3	43.3	85.9	**52.0**	**89.6**	**67.0**	29.9	**87.7**	**61.8**	**72.7**	**53.1**	**50.8**	**60.5**	**60.0**
Weakly Sup. [10]*	75.9	1.5	41.7	14.1	15.3	6.3	4.4	7.7	58.4	12.6	56.2	16.2	6.1	41.2	22.7	16.6	20.4	15.7	14.9	23.6
[11]	74.8	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
[12]	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
[13]	85.2	37.2	76.5	21.8	15.0	**23.8**	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
[0]	71.9	23.8	75.5	23.4	14.9	9.3	26.7	42.5	80.1	34.0	76.3	52.2	28.5	76.2	19.6	31.6	6.9	18.1	9.8	38.0
[0] & ps-gt	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

(*) Methods were trained on the dataset Cityshapes and evaluated on the Cityshapes's validation set

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Fully Sup. DeepLab*	**95.8**	**70.4**	85.4	42.7	41.0	21.2	**33.7**	**44.8**	**86.2**	51.4	88.4	58.1	**30.1**	86.4	43.8	56.7	42.8	33.9	54.8	56.2
Fully Sup. [0]*	95.6	70.1	**86.1**	**43.8**	**41.4**	16.6	31.3	43.3	85.9	**52.0**	**89.6**	**67.0**	29.9	**87.7**	**61.8**	**72.7**	**53.1**	**50.8**	**60.5**	**60.0**
Weakly Sup. [10]*	75.9	1.5	41.7	14.1	15.3	6.3	4.4	7.7	58.4	12.6	56.2	16.2	6.1	41.2	22.7	16.6	20.4	15.7	14.9	23.6
[11]	74.8	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
[12]	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
[13]	85.2	37.2	76.5	21.8	15.0	**23.8**	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
[0]	71.9	23.8	75.5	23.4	14.9	9.3	26.7	42.5	80.1	34.0	76.3	52.2	28.5	76.2	19.6	31.6	6.9	18.1	9.8	38.0
[0] & ps-gt	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

(*) Methods were trained on the dataset Cityshapes and evaluated on the Cityshapes's validation set

Background classes

Foreground classes

EXPERIMENTS: TRAIN ON SYNTHETIC DATA – TEST ON CITYSHAPES

Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Fully Sup. DeepLab*	**95.8**	**70.4**	85.4	42.7	41.0	21.2	**33.7**	**44.8**	**86.2**	51.4	88.4	58.1	**30.1**	86.4	43.8	56.7	42.8	33.9	54.8	56.2
Fully Sup. [0]*	95.6	70.1	**86.1**	**43.8**	**41.4**	16.6	31.3	43.3	85.9	**52.0**	**89.6**	**67.0**	29.9	**87.7**	**61.8**	**72.7**	**53.1**	**50.8**	**60.5**	**60.0**
Weakly Sup. [10]*	75.9	1.5	41.7	14.1	15.3	6.3	4.4	7.7	58.4	12.6	56.2	16.2	6.1	41.2	22.7	16.6	20.4	15.7	14.9	23.6
[11]	74.8	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
[12]	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
[13]	85.2	37.2	76.5	21.8	15.0	**23.8**	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
[0]	71.9	23.8	75.5	23.4	14.9	9.3	26.7	42.5	80.1	34.0	76.3	52.2	28.5	76.2	19.6	31.6	6.9	18.1	9.8	38.0
[0] & ps-gt	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

(*) Methods were trained on the dataset Cityshapes and evaluated on the Cityshapes's validation set

Background classes

Foreground classes

EXPERIMENTS: TEST ON CAMVID

		building	vegetation	sky	car	sign	road	pedestrian	fence	pole	sidewalk	cyclist	mIoU
Fully Sup.	DeepLab [14]	81.5	74.6	89.0	82.2	42.3	**92.2**	48.4	27.2	14.3	**75.4**	50.1	61.6
	Dilation8 [16]	**82.6**	**76.2**	**89.9**	**84.0**	**46.9**	**92.2**	**56.3**	**35.8**	**23.4**	75.3	**55.5**	**65.3**
[0]	[0]	66.3	55.0	61.9	73.4	37.4	82.7	41.4	23.9	9.2	57.7	14.9	47.6
	[0] & ps-gt	72.3	55.2	72.6	73.1	37.4	83.9	39.9	33.2	1.2	55.5	12.8	48.8

Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation (2018) [0]

Background classes

Foreground classes

CONCLUSIONS

- Proposed approach outperforms the domain adaption methods
- Proposed approach outperformed a standard semantic segmentation network (DeepLab [14]) from synthetic data
- The results validated the assumption that the shape remains realistic for foreground classes between synthetic and real data

Open Questions

- Could domain adaption improve the results of the proposed method?
- If VEIS would have been developed within the CARLA [17] open source framework (based on the Unreal Engine 4), could it be used for both foreground and background classes and improve performance?
  - GTA5 was released in 2013
  - Unreal Engine 4 is considered as one of the most photorealistic graphic motor engines
  - Also, Unreal Engine 5 was released in May 2020 (after the article was proposed back in 2018) increasing the photorealism

REFERENCES

[0] Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation. In: European Conference on Computer Vision (ECCV), pp. 86–103 (2018). doi: 10.1007/978-3-030-01216-8_6

[1] Cordts, M., et al.: The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016). doi: 10.1109/CVPR.2016.350

[2] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 102–118. Springer, Cham (2016). doi: 10.1007/978-3-319-46475-6_7

[3] Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: International Conference on Computer Vision (ICCV) (2017). doi: 10.1109/ICCV.2017.243

[4] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016). doi: 10.1109/CVPR.2016.352

[5] Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: a high-definition ground truth database. Pattern Recognition Letters 30(2), 88–97 (2009). doi: 10.1016/j.patrec.2008.04.005

[6] Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham. doi: 10.1007/978-3-319-46493-0_42

[7] Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1635–1643. IEEE, Santiago, Chile (2015). doi: 10.1109/ICCV.2015.191

[8] Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9911. Springer, Cham. doi: 10.1007/978-3-319-46478-7_34

[9] Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1796–1804. IEEE, Santiago, Chile (2015). doi: 10.1109/ICCV.2015.209

[10] Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Bringing background into the foreground: making all classes equal in weakly-supervised video semantic segmentation. In: IEEE International Conference on Computer Vision (ICCV), pp. 2125–2135. IEEE, Venice, Italy (2017). doi: https://doi.org/10.1109/ICCV.2017.232

References

[11] Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: IEEE International Conference on Computer Vision (ICCV), pp. 2039–2049. IEEE, Venice, Italy (2017). doi: 10.1109/ICCV.2017.223

[12] Chen, Y., Li, W., Gool, L.V.: ROAD: reality oriented adaptation for semantic segmentation of urban scenes. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7892–7901 (2018). IEEE, Salt Lake City, UT, USA (2018). doi: 10.1109/CVPR.2018.00823

[13] Hoffman, J., Tzeng, E., Park, T., Jun-Yan, Z., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: cycle-consistent adversarial domain adaptation. In: Proceedings of the 35th International Conference on Machine Learning, PMLR 80: 1989–1998 (2018). url: arXiv:1711.03213

[14] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848 (2018). doi: 10.1109/TPAMI.2017.2699184

[15] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE, Venice, Italy (2017). doi: 10.1109/ICCV.2017.322

[16] Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: 4th International Conference on Learning Representations, ICLR 2016 – Conference Track Proceedings (2016). url: arXiv:1511.07122

[17] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Proceedings of the 1st Annual Conference on Robot Learning, PMLR 78:1–16 (2017). doi: arXiv:1711.03938

Effective Use of Synthetic Data for Urban Scene Semantic Segmentation

European Conference on Computer Vision (ECCV), 2018

F. Saleh[1,2], M. Aliakbarian[1,2,3], M. Salzmann[4], L. Petersson[2], J. Alvarez[5]

{[1]ANU, [2]Data61-CSIRO, [3]ACRV}, Canberra, Australia

[4]CVLab, EPFL, Lausanne, Switzerland

[5]NVIDIA, Santa Clara, USA

Presented by: Ricardo B. Sousa (up201503004)

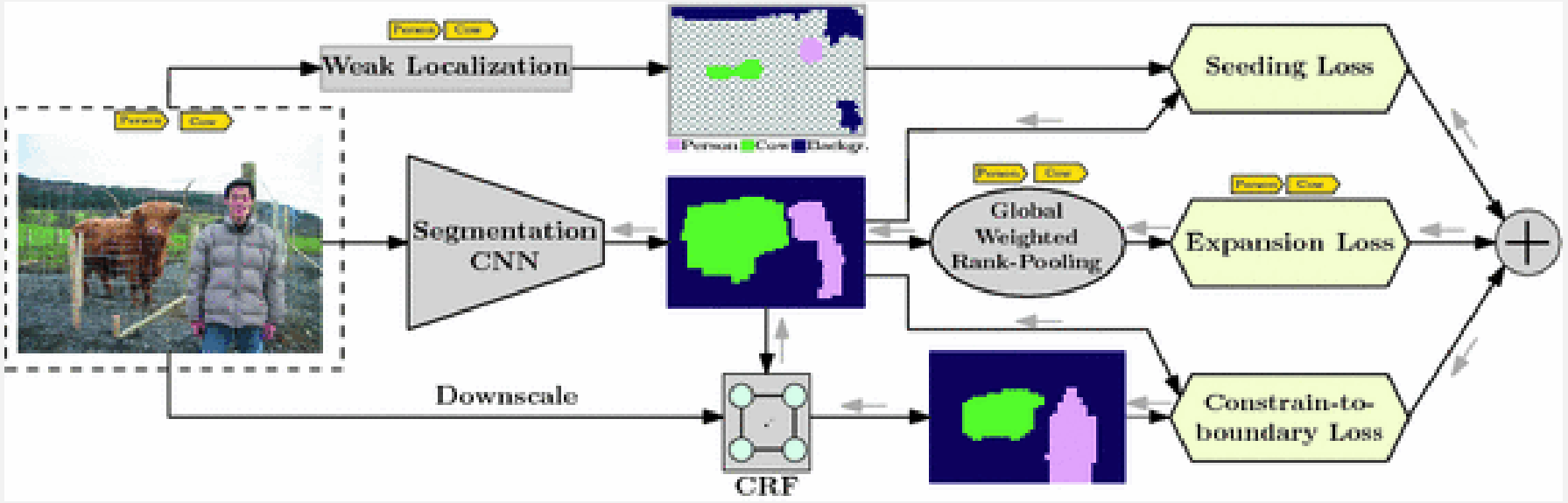
PDEEC, Computer Vision, 2020/2021

Related Work: Weakly-Supervised SS [6]

- **Seed, Expand and Constrain (SEC)**
  - Deep Convolution Neural Network (DCNN) with proposed loss functions for training
  - Use of a per-image annotation

- Seeding loss function to match localization of the objects
  - Generation of reliable object localization seeds (e.g., AlexNet or VGG)
  - Agnostic about the rest of the image (at this stage, the extension of the object does not matter)
- Global weighted rank pooling that is leverage by expansion loss to expand the object seeds to regions of a reasonable size
- Constrain-to-boundary loss to alleviate the problem of imprecise boundaries at training time

Related Work: Weakly-Supervised SS [6]



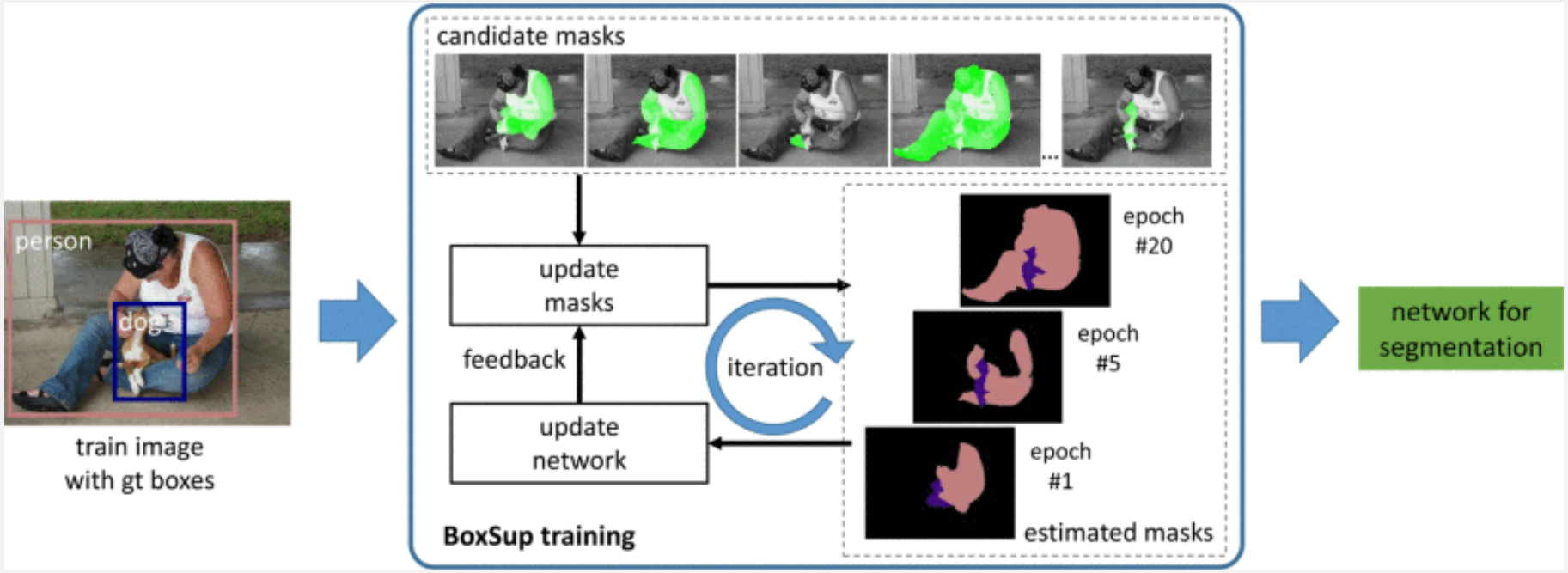
Source: Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation (2016) [6]

Related Work: Weakly-Supervised SS [7]

- **BoxSup**
  - Requires only bounding box annotations
  - Iterates between automatically generating region proposals + training convolution networks

- Generating region proposals
  - Usually, these methods have high recall rates
  - Generate candidates of greater variance → provides a kind of data augmentation for network training
- Deep Convolutional Neural Networks (DCNN)
  - Candidate segments are used to pick better candidates
  - Semantic features learned by the network are used to pick better candidates

Related Work: Weakly-Supervised SS [7]



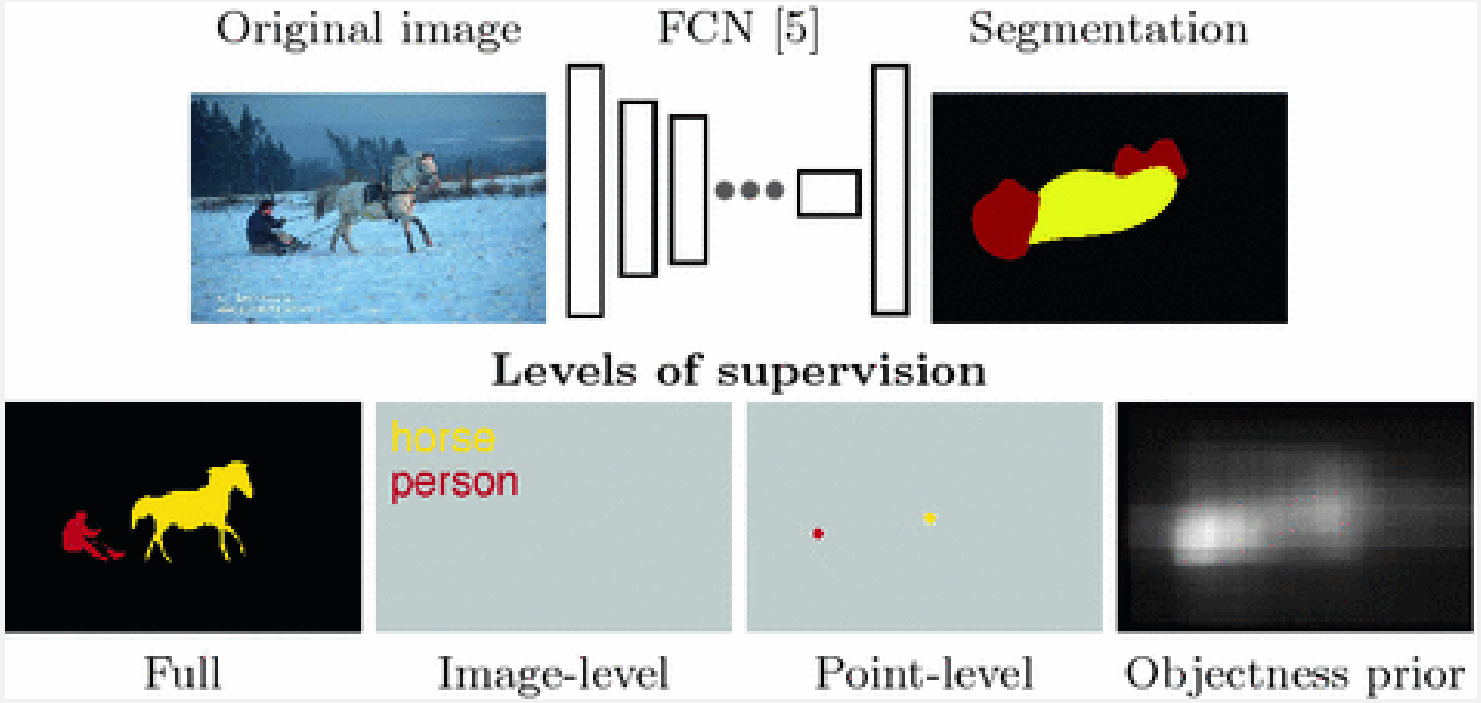
Source: Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation (2015) [7]

Related Work: Weakly-Supervised SS [8]

- **Semantic Segmentation with Point Supervision**
  - Require point annotation for the objects in the foreground (1 point / object class)

- Objectness Prior
  - Inferring the spatial extent of the objects
  - Provides a probability for whether each pixel belong to any foreground class as opposed to background
- Point-level supervision
  - 1. Ask the user to either determine that the class is not present or to point to one object instance
  - 2. Ask multiple annotators to do the same task as 1. and set a confidence coefficient of the annotator's accuracy that provided the point
  - 3. Ask the annotator(s) to point every instance of the classes in the image and compute the order of the points
  - Goal: the first points is more likely to correspond to the largest object instant

Related Work: Weakly-Supervised SS [8]



Source: Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision (2016) [8]

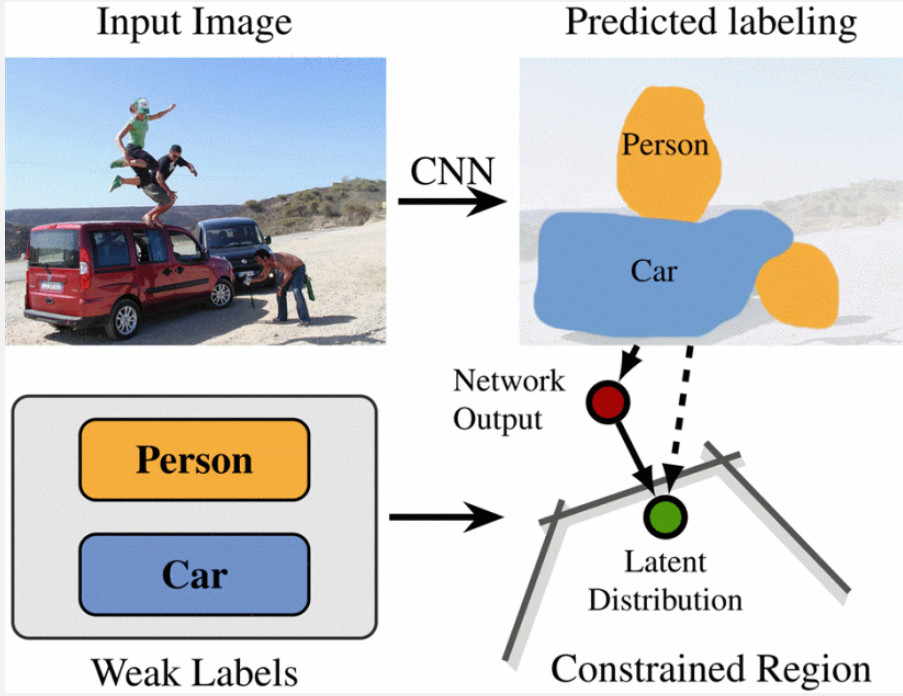
F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

Related Work: Weakly-Supervised SS [9]

- **Constrained Convolutional Neural Network (CCNN)**
  - Optimises any set of linear constraints on the output space (i.e., predicted label distribution) of a CNN

- Linear constraints
  - Can describe the existence and expected distribution of labels from image-level tags
  - Example: if a car is present in an image, a certain number of pixels should be labelled as car

Related Work: Weakly-Supervised SS [9]



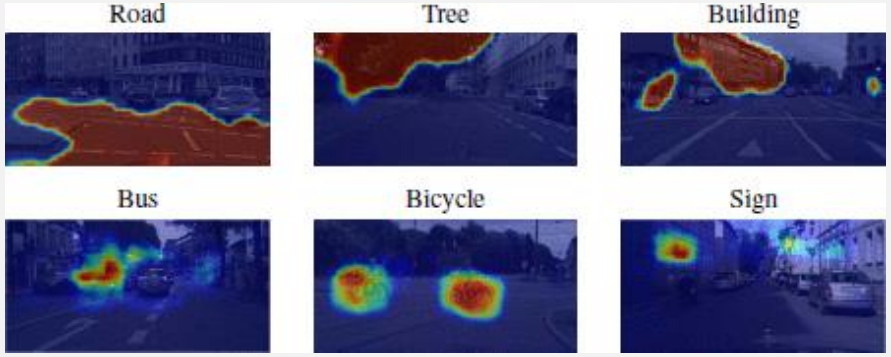
Source: Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation (2015) [9]

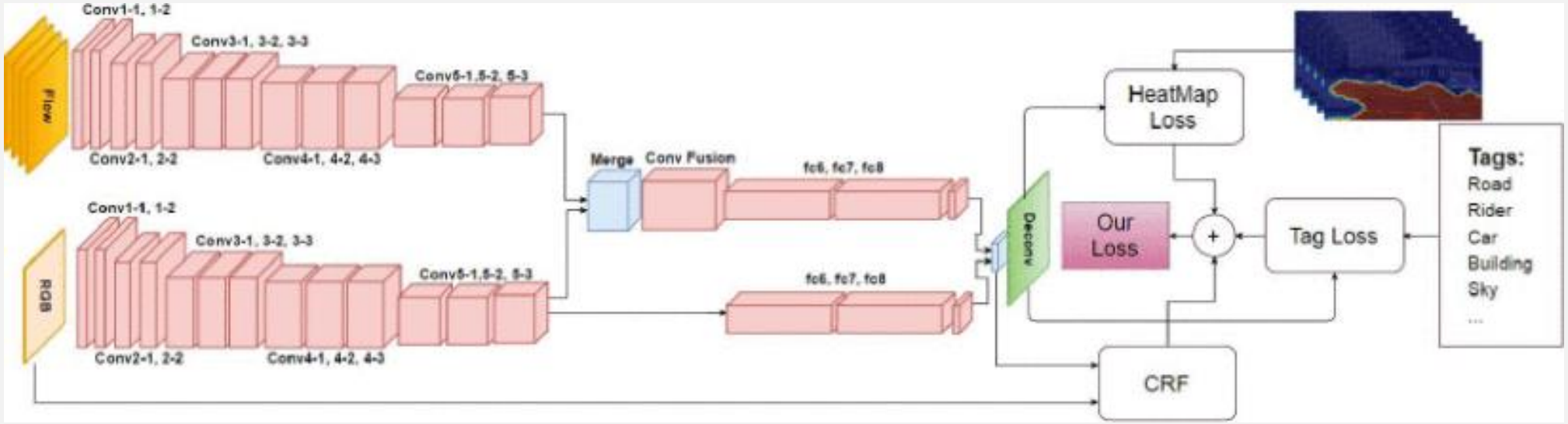
Related Work: Weakly-Supervised SS – Fore+Back [10]

- **Weakly-supervised video semantic segmentation**
  - Does not require pixel-level annotations
  - Relies on class-dependent heatmaps obtained from classifiers trained for image-level recognition
  - Treats all classes, foreground and background ones, equally

- Two-stream deep network
  - Jointly leverages appearance and motion
  - Early fusion: learns to combinate the spatial and temporal information into a spatio-temporal stream
  - Late fusion: leveraging the valuable semantic information of the spatial stream to merge it with the spatio-temporal one for final prediction

Related Work: Weakly-Supervised SS – Fore+Back [10]





Source: Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Bringing background into the foreground: making all classes equal in weakly-supervised video semantic segmentation (2017) [10]

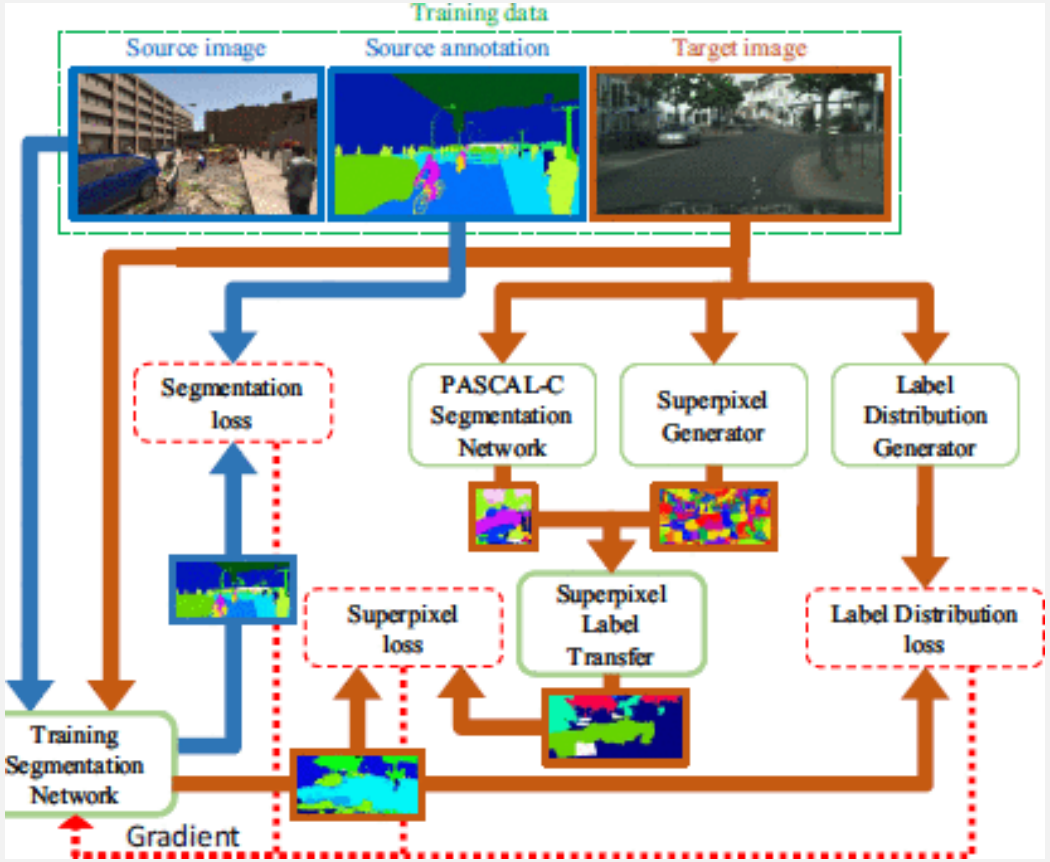
Related Work: Synthetic Data – Domain Adaptation [11]

- **Curriculum-style domain adaptation**
  - Begins with the easy tasks (suffer less due to domain discrepancy) to gain some high-level proprieties
  - Then, had task: predictions over the target image are forced to follow those necessary properties as possible

- Urban traffic scene images have strong idiosyncrasies
  - E.g., the size and spatial relations of buildings, streets, cars, etc.
- Structured output in semantic segmentation enables convenient posterior regularization
  - Express structural constraints in latent variables (variables that are not directly observed, only inferred) arising from prior knowledge and indirect supervision

F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

Related Work: Synthetic Data – Domain Adaptation [11]



Source: Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes (2017) [11]

F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

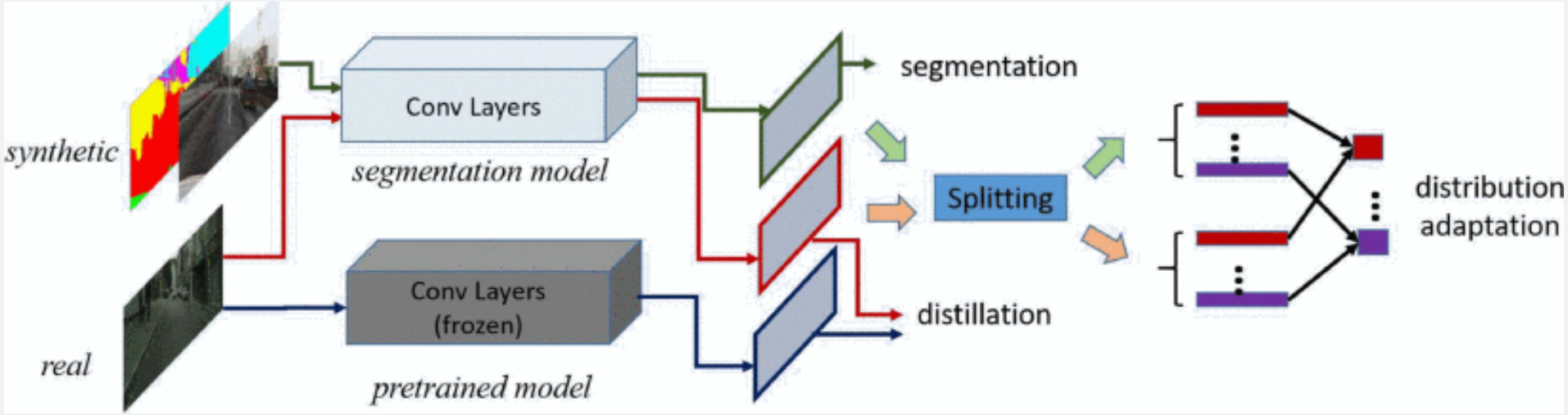
Related Work: Synthetic Data – Domain Adaptation [12]

- **ROAD: Reality Oriented Adaptation**
  - Target guided distillation approach to learn the real image style
  - Spatial-aware adaptation scheme to effectively align the distribution of two domains

- Real style orientation
  - Use target real images to imitate a pretrained real style model
  - Model distillation to enforce the output from segmentation model similar with the output of a pretrained model
- Real distribution orientation
  - Exploit the intrinsic geometry information presented in urban scene

F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

Related Work: Synthetic Data – Domain Adaptation [12]



Source: Chen, Y., Li, W., Gool, L.V.: ROAD: reality oriented adaptation for semantic segmentation of urban scenes (2018) [12]

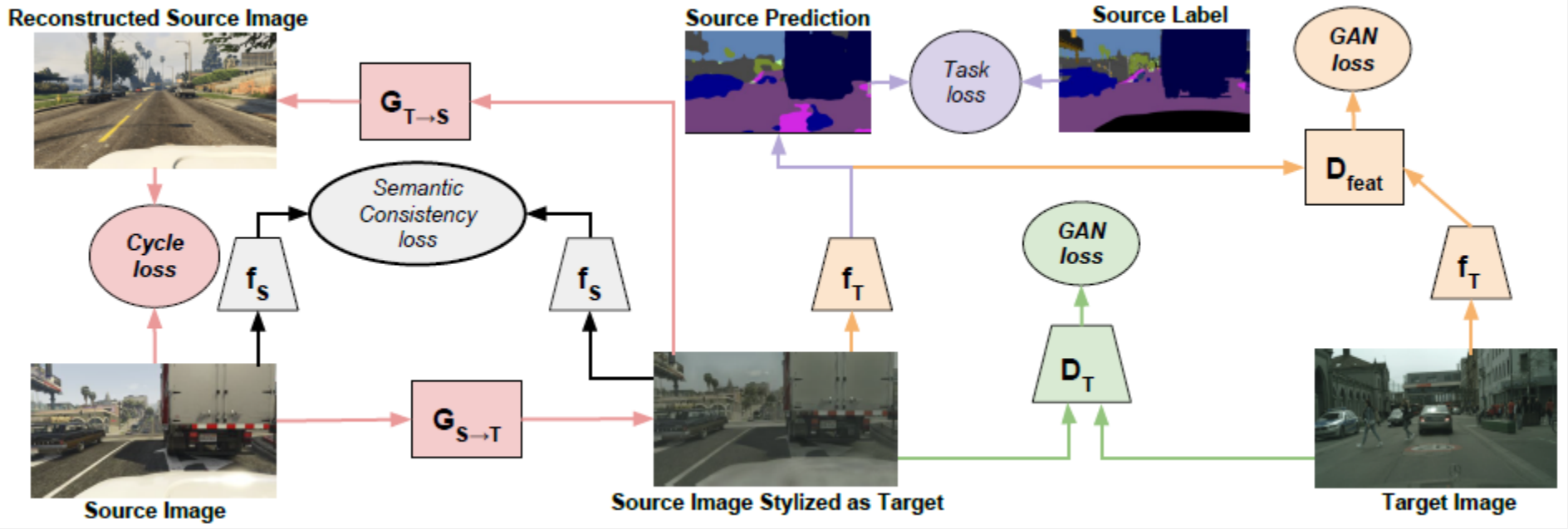
Related Work: Synthetic Data – Domain Adaptation [13]

- **CyCADA: Cycle-Consistent Adversarial Domain Adaptation**
  - Adapts representations at both pixel-level and feature-level
  - Enforces local and global structural consistency through pixel cycle-consistency and semantic losses

- Reconstruction (cycle-consistency) loss
  - Encourages the cross-domain transformation to preserve local structural information
- Semantic loss
  - Enforces semantic consistency

F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

Related Work: Synthetic Data – Domain Adaptation [13]



Source: Hoffman, J., Tzeng, E., Park, T., Jun-Yan, Z., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: cycle-consistent adversarial domain adaptation (2018) [13]

F. Saleh et al.: Effective Use of Synthetic Data for Urban Scene Semantic Segmentation (2018) [0]

METHOD: DEEPLAB [14]

- **Semantic image segmentation with deep learning**
  - Convolution with upsampled filters
  - Atrous convolution allow to explicitly control the resolution at which feature responses are computed within DCNN
  - Atrous Spatial Pyramid Pooling (ASPP) to robustly segment objects at multiple scales
  - Improve localization of object boundaries by combining methods from DCNN and probabilistic graphical models

- DeepLab-CRF-LargeFOV (used in [0])
  - Matches the performance of VGG-16
  - 3.36 times faster
  - Significantly fewer parameters (20.5M vs 134.3M)

METHOD: MASK R-CNN [15]

- **Mask R-CNN**
  - Extends the Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition
  - Predicts a segmentation mask in a pixel-to-pixel manner

- RoIAlign
  - Faster R-CNN was not designed for pixel-to-pixel alignment between network inputs and outputs
  - RoIAlign tries to fix the misalignment
- Decoupling mask and class prediction
  - Prediction of a mask for each class **independently**, without competition among classes
  - Relies on the RoI classification branch to predict the category

EXPERIMENTS: DATASETS

- **Cityscapes** [1]
  - Real data (street scenes from 50 different cities)
  - 5000 images w/ pixel-level annotation
- **CamVid** [5]
  - Real data (camera mounted inside a car)
  - 18000 (~10min at 30Hz) images (960 x 720) w/ pixel-level segmentation

Experiments: Datasets

- **GTA5** [2]
  - Synthetic data (Grand Theft Auto V)
  - 24966 photo-realistic images (1920 x 1080) w/ pixel-level annotations
  - Class definitions compatible with the Cityshapes dataset
- **VIPER** [3]
  - Synthetic data (Grand Theft Auto V)
  - Wider range of weather conditions relative to GTA5
  - 250k video frames all annotated w/ ground-truth labels for, e.g., optical flow, semantic instance segmentation, object detection and tracking, or visual odometry
  - Class definitions are not compatible with the Cityshapes dataset
  - Does not have a large diversity of foreground classes
    - Missing: rider, traffic, sign, train, bicycle

EXPERIMENTS: DATASETS

- **SYNTHIA** [4]
  - Synthetic data (Unity)
  - 9400 images w/ pixel-level annotation
  - Class definitions compatible with the Cityshapes dataset (subset SYNTHIA-RAND-CITYSHAPES)
  - Does not have a large diversity of foreground classes
    - Missing: train, truck, traffic light, traffic sign

Effective Use of Synthetic Data for Urban Scene Semantic Segmentation

European Conference on Computer Vision (ECCV), 2018

F. Saleh[1,2], M. Aliakbarian[1,2,3], M. Salzmann[4], L. Petersson[2], J. Alvarez[5]

{[1]ANU, [2]Data61-CSIRO, [3]ACRV}, Canberra, Australia

[4]CVLab, EPFL, Lausanne, Switzerland

[5]NVIDIA, Santa Clara, USA

Presented by: Ricardo B. Sousa (up201503004)

PDEEC, Computer Vision, 2020/2021