PAPER ANALYSIS

Presented by Yannis He

-

Paper: **Unsupervised Domain Adaptation in Semantic Segmentation via Orthogonal and Clustered Embeddings**

Conference: WACV 2021

Authors: Marco Toldo, Umberto Michieli, Pietro Zanuttigh | Department of Information Engineering, University of Padova

https://arxiv.org/abs/2011.12616

INTRODUCTION

- Background:
  - Semantic segmentation is powerful but convolutional networks have nature of data hungry
  - → lead to high demand for adaptation techniques, which is:
    - Transfer learned knowledge from label-abundant domains to unlabeled ones.
  - Typical methods rely on auto-encoder structure
    - Cons: high complexity, massive labeled data required
  - Current popular solution: Adversarial
    - Cons:
      - perform a semantically unaware alignment, as they neglect the underlying class-conditional data distribution. Additionally, they typically require a long training time to converge and the process may be unstable.

INTRODUCTION (CONT')

- Proposal:
  - Unsupervised Domain Adaptation (UDA) strategy, based on a **feature clustering** method that captures the different semantic modes of the **feature** distribution and **groups features of the same class into tight and well-separated clusters**
  - Two novel learning objectives to enhance discriminative clustering:
    - Orthogonality loss:
      - Forces spaced out individual representations to be orthogonal
    - A sparsity loss:
      - Reduces the number of active feature channels class-wise
- Achievement:
  - State-of-the-art performance in *synthetic-to-real*

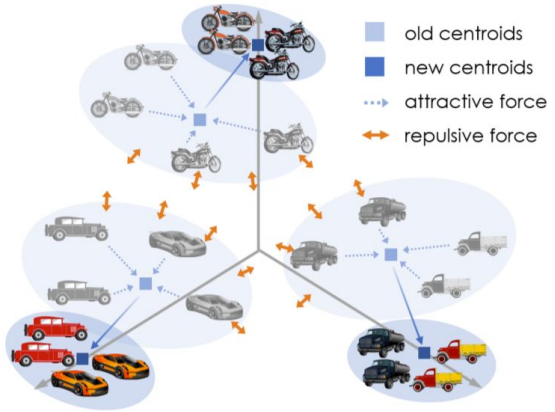


Figure 1: The proposed domain adaptation scheme is driven by 3 main components, i.e., feature clustering, orthogonality and sparsity. These push features in the previous step (in light gray) to new locations (colored) where features of the same class are clustered, while features of distinct classes are pushed away. To further improve performances, features of distinct classes are forced to be orthogonal and sparse.

METHOD - OVERVIEW

- Outline:
  - Discriminative Clustering
    - Applied over individual feature representations
    - Enforced the clustering objective on both source & target representations
  - Orthogonality
    - Two-fold objective:
    1. Forces feature vectors of kindred semantic connotations to activate the same channels, while turning off the remaining ones
    2. Constraints feature vectors of dissimilar semantic connotations to activate different channels to reduce cross interference.
       a. I.e. with no overlap
  - Sparsity
    - Encourage lower volume of active feature channel from latent representations (concentrate on few dimensions)

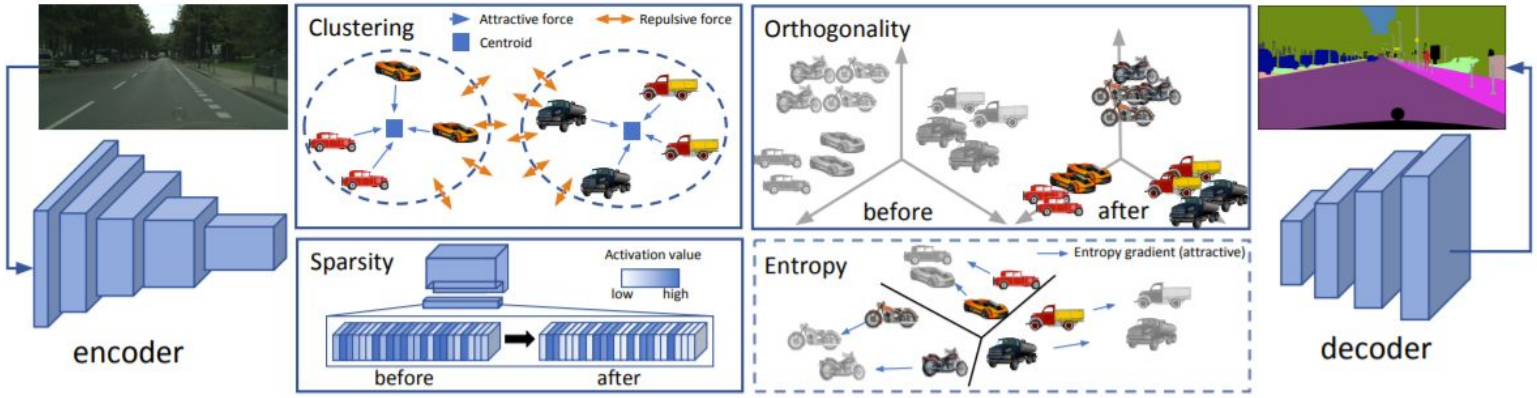


Figure 2: Overview of the proposed approach. Features after supervised training on the source domain are represented in light gray, while features of the current step are colored. A set of techniques is employed to better shape the latent feature space spanned by the encoder. Features are clustered and the clusters are forced to be disjoint. At the same time, features belonging to different classes are forced to be orthogonal with respect to each other. Additionally, features are forced to be sparse and an entropy minimization loss could also be added to guide target samples far from the decision boundaries.

$$\mathcal{L}'_{tot} = \mathcal{L}_{ce} + \lambda_{cl} \cdot \mathcal{L}_{cl} + \lambda_{or} \cdot \mathcal{L}_{or} + \lambda_{sp} \cdot \mathcal{L}_{sp} \quad (1)$$

$$\mathcal{L}_{tot} = \mathcal{L}'_{tot} + \lambda_{em} \cdot \mathcal{L}_{em} \quad (2)$$

L_{ce}: standard supervised cross entropy loss

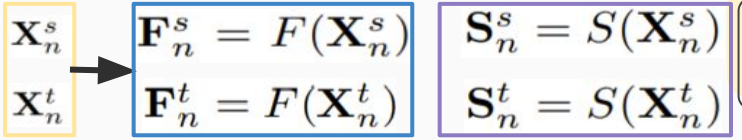
L_{cl}: main clustering objective

L_{or}: orthogonality constraint

L_{sp}: sparsity constraint

L_{em}: off-the-shelf entropy-minimization like objective

λ: parameters balance the multiple losses



METHOD - DISCRIMINATIVE CLUSTERING

Given a batch of source and target training images, they extract the feature tensors along with the computed output segmentation maps.

$$\mathcal{L}_{cl} = \frac{1}{|\mathbf{F}_n^{s,t}|} \sum_{\substack{\mathbf{f}_i \in \mathbf{F}_n^{s,t} \\ \hat{y}_i \in \mathbf{S}_n^{s,t}}} d(\mathbf{f}_i, \mathbf{c}_{\hat{y}_i}) - \frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)} \sum_{j \in \mathcal{C}} \sum_{\substack{k \in \mathcal{C} \\ k \neq j}} d(\mathbf{c}_j, \mathbf{c}_k) \tag{3}$$

$\mathbf{X}_n^s \in \mathbb{R}^{H \times W \times 3}$: source dataset sample

$\mathbf{Y}_n^s \in \mathbb{R}^{H \times W}$: source semantic maps

$\mathbf{X}_n^t \in \mathbb{R}^{H \times W \times 3}$: target dataset samples, with no semantic maps

$$\mathbf{c}_j = \frac{\sum_{\mathbf{f}_i} \sum_{\hat{y}_i} \delta_{j,\hat{y}_i} \mathbf{f}_i}{\sum_{\hat{y}_i} \delta_{j,\hat{y}_i}}, \quad j \in \mathcal{C}$$

METHOD - ORTHOGONALITY

- orthogonality constraint in the form of a training objective.
- Feature vectors from either domains, but of different semantic classes are forced to be orthogonal
    - Meaning that their scalar product should be small.
- The paper devises the orthogonality objective as an entropy minimization loss that forces each feature to be orthogonal with respect to all the centroids but one:

$$\mathcal{L}_{or} = - \sum_{\mathbf{f}_i \in F(\mathbf{X}_n^{s,t})} \sum_{j \in \mathcal{C}} p_j(\mathbf{f}_i) \log p_j(\mathbf{f}_i) \qquad (5)$$

where $\{p_j(\mathbf{f}_i)\}$ denotes a probability distribution derived as:

$$p_j(\mathbf{f}_i) = \frac{e^{\langle \mathbf{f}_i, \mathbf{c}_j \rangle}}{\sum_{k \in \mathcal{C}} e^{\langle \mathbf{f}_i, \mathbf{c}_k \rangle}}, \quad j \in \mathcal{C} \qquad (6)$$

METHOD - SPARSITY

- Introduce Sparsity constraint to shape class-wise feature structures inside the latent space
    - A sparsity loss, with the intent of decreasing the number of active feature channels of latent vectors.

$$\mathcal{L}_{sp} = -\sum_{i \in \mathcal{C}} ||\tilde{\mathbf{c}}_i - \boldsymbol{\rho}||_2^2 \quad (7)$$

where $\tilde{\mathbf{c}}_i$ stands for the normalized centroid $\mathbf{c}_i$ in $[0,1]^D$ and D denotes the number of feature maps in the encoder output. empirically set $\boldsymbol{\rho} = [0.5]^D$

RESULT

- Metrics:
  - GTA5 → Cityscapes
    - State-of-the-art performance in feature-level UDA for semantic segmentation: 45.9% (from 37.0%)
  - SYNTHIA → Cityscapes
    - State-of-the-art performance for feature-level UDA: 48.2% (from 40.5%)

	B	Method	Road	Sidewalk	Building	Wall*	Fence*	Pole*	T. Light	T. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mIoU (all)	mIoU* (13-cl)
GTA5 → Cityscapes	VGG16	Source Only	26.5	13.3	45.1	6.0	15.2	16.5	21.3	8.5	78.0	8.3	59.7	45.0	10.5	69.1	22.8	17.9	0.0	16.4	2.7	25.4	-
		FCNs ITW [22]	70.4	**32.4**	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1	-
		CyCADA (feat) [21]	85.6	30.7	74.7	14.4	13.0	17.6	13.7	5.8	74.6	15.8	69.9	38.2	3.5	72.3	16.0	5.0	0.1	3.6	0.0	29.2	-
		CBST [75]	66.7	26.8	73.7	14.8	9.5	**28.3**	25.9	10.1	75.5	15.7	51.6	**47.2**	6.2	71.9	3.7	2.2	**5.4**	**18.9**	**32.4**	30.9	-
		MinEnt [61]	85.1	18.9	76.3	**32.4**	**19.7**	19.9	21.0	8.9	76.3	**26.2**	63.1	42.8	5.9	**80.8**	**20.2**	9.8	0.0	14.8	0.6	32.8	-
		MaxSquare IW (r) [7]	81.4	20.0	75.4	19.4	19.1	16.1	24.4	7.9	78.8	22.9	65.9	45.0	12.3	74.6	16.1	10.3	0.2	11.3	1.0	31.7	-
		Ours ($\mathcal{L}'_{tot}$)	83.6	16.6	79.0	19.8	18.7	21.5	27.3	**15.9**	80.2	14.3	**72.6**	47.0	17.5	76.8	16.6	**13.9**	0.1	16.0	3.4	33.7	-
		Ours ($\mathcal{L}_{tot}$)	**86.0**	13.5	**79.4**	20.4	18.5	21.5	**27.6**	15.2	**80.8**	21.9	**72.6**	46.3	**18.1**	80.0	16.9	13.1	1.0	14.6	2.0	**34.2**	-
	ResNet101	Source Only	81.8	16.3	74.4	18.6	12.7	23.5	29.3	18.1	73.5	21.4	77.6	55.6	25.6	74.1	28.6	10.2	3.0	25.8	32.7	37.0	-
		AdaptSegNet (feat) [57]	83.7	27.6	75.5	20.3	19.9	27.4	28.3	27.4	79.0	28.4	70.1	55.1	20.2	72.9	22.5	35.7	**8.3**	20.6	23.0	39.3	-
		MinEnt [61]	84.4	18.7	80.6	23.8	23.2	28.4	36.9	23.4	83.2	25.2	**79.4**	59.0	**29.9**	78.5	33.7	29.6	1.7	29.9	33.6	42.3	-
		SAPNet [25]	88.4	38.7	79.5	**29.4**	**24.7**	27.3	32.6	20.4	82.2	32.9	73.3	55.5	26.9	82.4	31.8	41.8	2.4	26.5	24.1	43.2	-
		MaxSquare IW [7]	89.3	**40.5**	81.2	29.0	20.4	25.6	34.4	19.0	83.6	34.4	76.5	59.2	27.4	83.8	**38.4**	43.6	7.1	**32.2**	32.5	45.2	-
		Ours ($\mathcal{L}'_{tot}$)	88.7	32.2	81.8	24.1	22.1	**30.8**	**37.6**	**32.8**	83.4	36.3	76.0	60.0	27.0	81.0	34.2	43.0	8.0	23.4	38.1	45.3	-
		Ours ($\mathcal{L}_{tot}$)	**89.4**	30.7	**82.1**	23.0	22.0	29.2	**37.6**	31.7	**83.9**	**37.9**	78.3	**60.7**	27.4	**84.6**	37.6	**44.7**	7.3	26.0	**38.9**	**45.9**	-
SYNTHIA → Cityscapes	VGG16	Source Only	7.8	13.7	66.6	2.2	0.0	23.9	4.8	13.3	71.2	-	76.5	49.2	12.1	67.1	-	24.5	-	9.8	9.2	28.3	32.8
		FCNs ITW [22]	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	-	68.7	51.2	3.8	54.0	-	3.2	-	0.2	0.6	20.2	22.9
		Cross-City [9]	62.7	25.6	**78.3**	-	-	-	1.2	5.4	81.3	-	81.0	37.4	6.4	63.5	-	16.1	-	1.2	4.6	-	35.7
		CBST [75]	69.6	28.7	69.5	**12.1**	**0.1**	**25.4**	11.9	13.6	**82.0**	-	**81.9**	49.1	14.5	66.0	-	6.6	-	3.7	**32.4**	35.4	36.1
		MinEnt [61]	37.8	18.2	65.8	2.0	0.0	15.5	0.0	0.0	76.0	-	73.9	45.7	11.3	66.6	-	13.3	-	1.5	13.1	27.5	32.5
		MaxSquare IW (r) [7]	9.1	12.7	72.5	1.0	0.0	22.3	7.0	8.4	80.0	-	77.9	49.4	10.0	71.8	-	23.8	-	6.0	13.5	29.1	34.0
		Ours ($\mathcal{L}'_{tot}$)	**78.5**	29.9	77.7	1.2	**0.1**	24.1	11.9	**15.0**	78.7	-	78.5	51.0	15.4	73.7	-	**24.7**	-	**10.1**	23.5	**37.1**	**43.7**
		Ours ($\mathcal{L}_{tot}$)	78.3	**30.1**	78.0	1.7	**0.1**	24.1	**12.0**	14.6	79.7	-	79.1	**51.4**	**15.5**	**74.4**	-	23.7	-	9.1	22.7	**37.1**	**43.7**
	ResNet101	Source Only	39.5	18.1	75.5	10.5	0.1	26.3	9.0	11.7	78.6	-	81.6	57.7	21.0	59.9	-	30.1	-	15.7	28.2	35.2	40.5
		AdaptSegNet (feat) [57]	62.4	21.9	76.3	-	-	-	11.7	11.4	75.3	-	80.9	53.7	18.5	59.7	-	13.7	-	**20.6**	24.0	-	40.8
		MinEnt [61]	73.5	29.2	77.1	7.7	0.2	27.0	7.1	11.4	76.7	-	82.1	57.2	**21.3**	69.4	-	29.2	-	12.9	27.9	38.1	44.2
		SAPNet [25]	81.7	33.5	75.9	-	-	-	7.0	6.3	74.8	-	78.9	52.1	**21.3**	75.7	-	**30.6**	-	10.8	28.0	-	44.3
		MaxSquare IW [7]	78.5	34.7	76.3	6.5	0.1	**30.4**	12.4	12.2	**82.2**	-	**84.3**	**59.9**	17.9	80.6	-	24.1	-	15.2	31.2	40.4	46.9
		Ours ($\mathcal{L}'_{tot}$)	64.4	25.5	77.3	**14.3**	**0.9**	29.6	**21.2**	**24.2**	76.6	-	79.7	53.7	15.5	79.7	-	11.0	-	11.0	**35.2**	38.7	44.2
		Ours ($\mathcal{L}_{tot}$)	**88.3**	**42.2**	**79.1**	7.1	0.2	24.4	16.8	16.5	80.0	-	**84.3**	56.2	15.0	**83.5**	-	27.2	-	6.3	30.7	**41.1**	**48.2**

Table 1: Numerical evaluation of the GTA5 and SYNTHIA to Cityscapes adaptation scenarios in terms of per-class and mean IoU. Evaluations are performed on the validation set of the Cityscapes dataset. In all the experiments the DeepLab-V2 segmentation network is employed, with VGG-16 (top) or ResNet-101 (bottom) backbones. The mIoU* results in the last column refer to the 13-classes configuration, i.e., classes marked with * are ignored. MaxSquares IW (r) denotes our re-implementation, as original results are provided only for the ResNet-101 backbone.

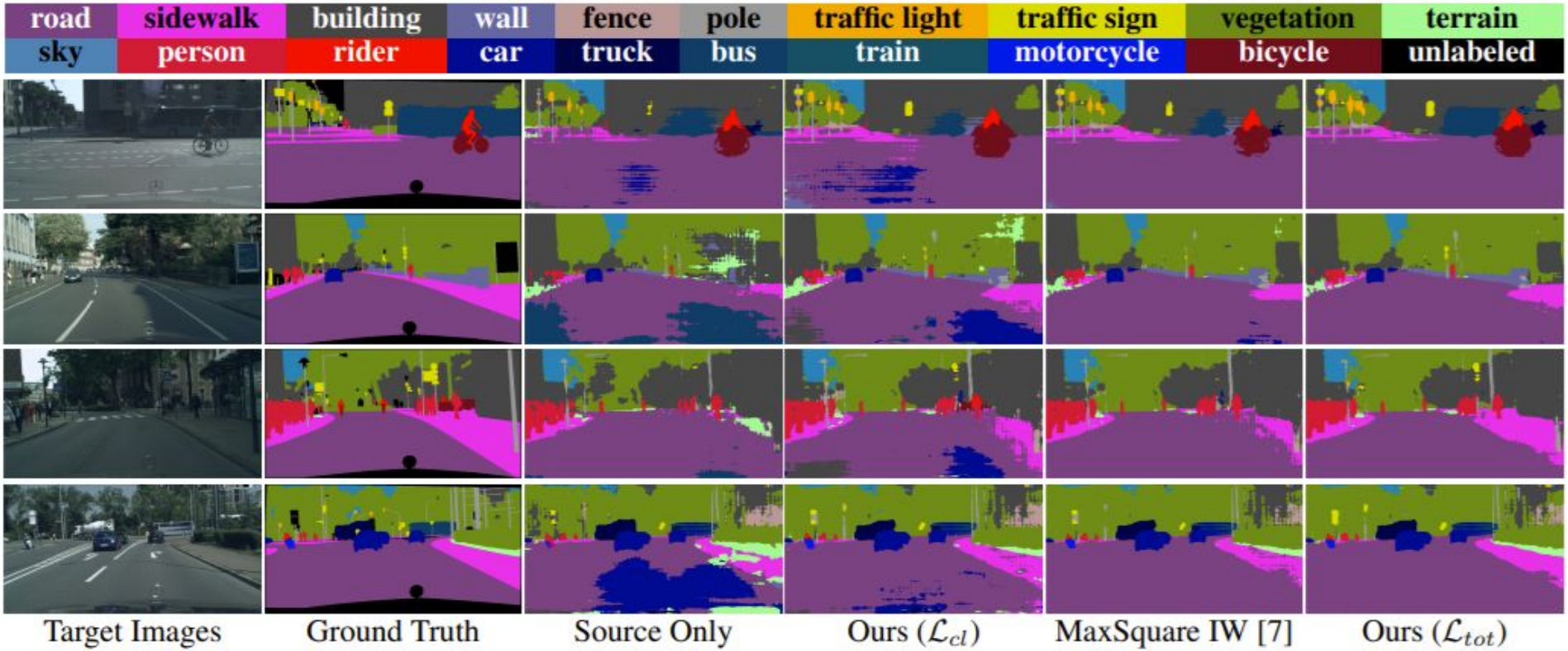


Figure 3: Semantic segmentation of some sample scenes from the Cityscapes validation dataset when adaptation is performed from the GTA5 source dataset and the DeepLab-V2 with ResNet-101 backbone is employed (*best viewed in colors*).

ABLATION STUDY

$\mathcal{L}_{cl}$	$\mathcal{L}_{or}$	$\mathcal{L}_{sp}$	$\mathcal{L}_{em}$	mIoU
				37.0
✓				42.3
	✓			43.2
		✓		43.7
			✓	44.8
✓	✓	✓		45.3
✓	✓	✓	✓	45.9

Table 2: Ablation results on the contribution of each adaptation module in the GTA5 to Cityscapes scenario and with ResNet-101 as backbone.