Video Computing Group
Center for Robotics & Intelligent Systems
UC RIVERSIDE Marlan and Rosemary Bourns College of Engineering

ECCV TEL AVIV 2022

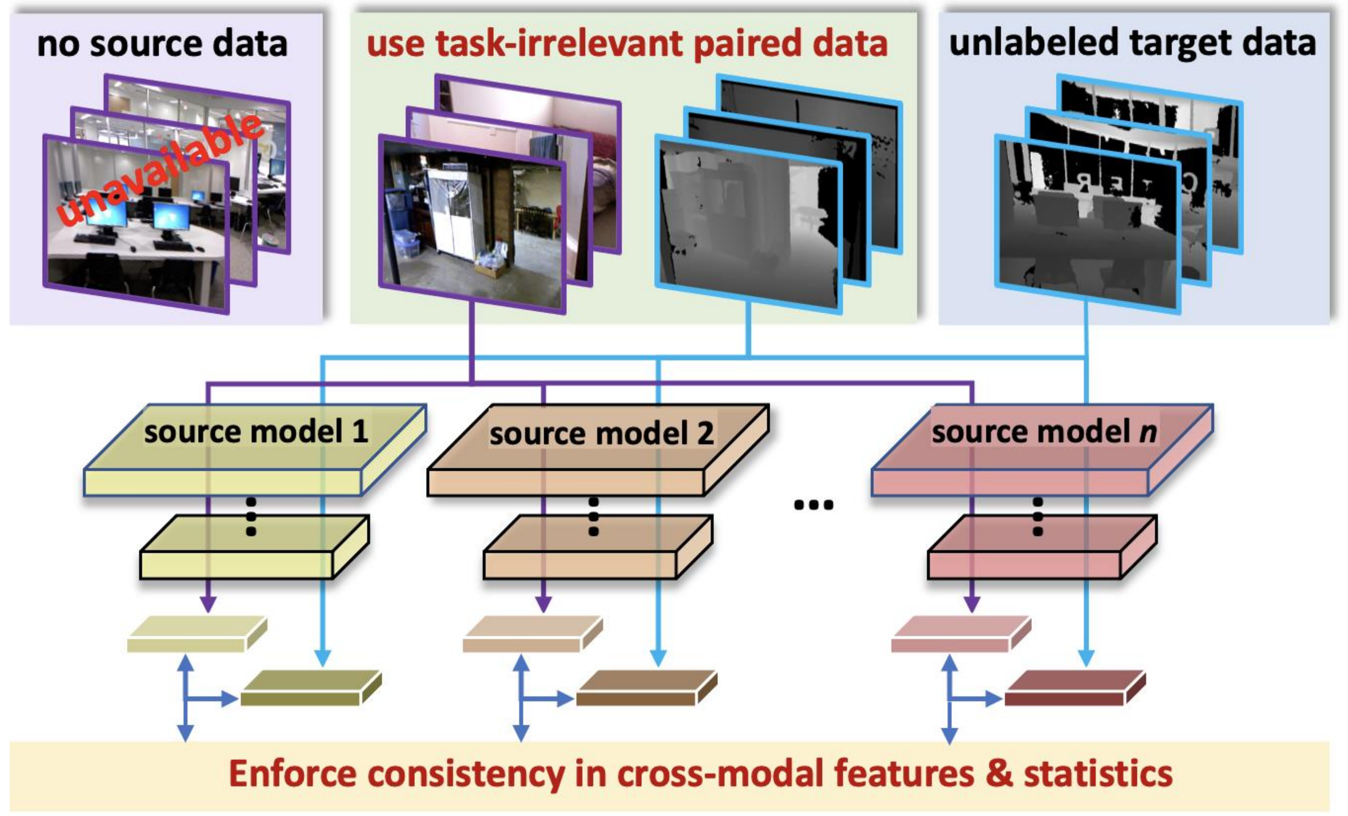
Cross-Modal Knowledge Transfer Without Task-Relevant Source Data

Sk Miraj Ahmed[1], Suhas Lohit[2], Kuan-Chuan Peng[2], Michael J. Jones[2], Amit K. Roy-Chowdhury[1]
University of California, Riverside (UCR)[1], Mitsubishi Electric Research Laboratories (MERL)[2]

MITSUBISHI ELECTRIC Changes for the Better

Problem Definition

- Conventional source free Unsupervised Domain Adaptation (UDA) approaches assume source and target data to be of same modality.
- Contrary to that we tackle a novel problem where the unlabeled target is of different modality than the source, assuming only trained source model is available with no Task-Relevant source data.
- We generalize our method for both single and multiple sources.

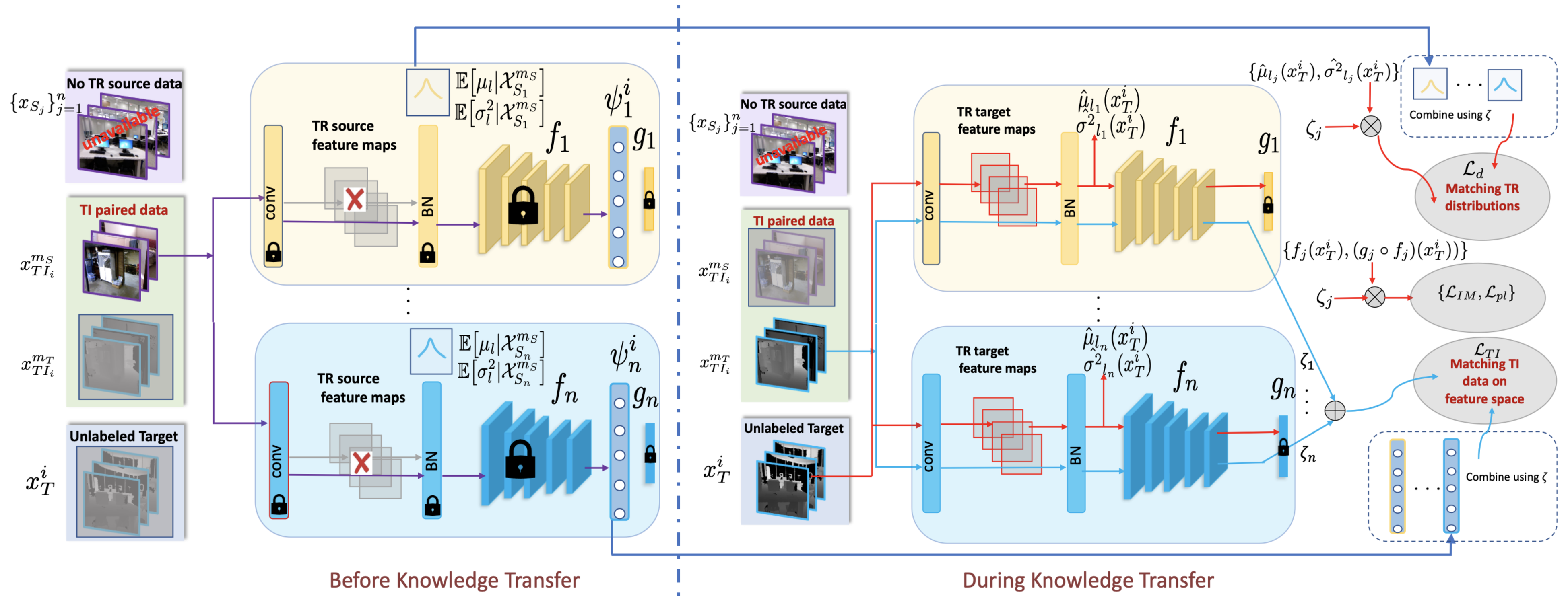


Problem setup. Difference between standard source-free and cross-modal source free UDA.

Our Contribution

- We formulate a novel problem for knowledge transfer from a model trained for a source modality to a different target modality without any access to task-relevant source data and when the target data is unlabeled.
- In order to bridge the gap between modalities, we propose a novel framework, SOCKET, for cross-modal knowledge transfer without access to source data (a) using an external task-irrelevant paired dataset, and (b) by matching the moments obtained from the normalization layers in the source models with the moments computed on the unlabeled target data.

Framework Overview



Overall framework of our approach. Our framework can be split into two parts: (i) Before Knowledge Transfer (left): We freeze the source models and pass the task-irrelevant (TI) source data through the source feature encoders to extract the TI source features. As task-relevant (TR) source feature maps are not available, we extract the stored moments of its distribution from the BN layers. (ii) During Knowledge Transfer (right): We freeze only the classification layers and feed the TI and unlabeled TR target data through the models to get batch-wise TI target features and the TR target moments, respectively, which we match with pre-extracted source features and moments to jointly train all the feature encoders along with the mixing weights. The final target model is the optimal linear combination of the updated source models

Results

Source RGB \ Target depth	Kinect v1			Kinect v2			Realsense			Xtion		
	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET
Kinect v1	14.8	16.7	**25.3**	14.6	20.3	**23.6**	9.0	11.9	**13.4**	7.1	15.3	**18.1**
Kinect v2	4.0	12.8	**13.6**	17.0	29.4	**35.2**	10.8	19.3	**22.8**	**10.6**	7.0	8.3
Realsense	2.0	7.9	**20.3**	7.1	18.4	**23.5**	14.7	27.4	**30.0**	5.1	9.5	**11.8**
Xtion	0.7	9.5	**14.2**	6.0	20.2	**24.2**	9.0	21.8	**23.5**	8.1	13.2	**22.2**
Average	5.4	11.7	**18.4**	11.2	22.1	**26.6**	10.9	20.1	**22.4**	7.7	11.3	**15.1**

Source RGB \ Target depth	Kinect v1		Kinect v2		Realsense		Xtion	
	DECISION	SOCKET	DECISION	SOCKET	DECISION	SOCKET	DECISION	SOCKET
Kinect v1 + Kinect v2	17.9	**19.5**	34.2	**36.6**	18.8	**19.8**	14.6	**18.0**
Kinect v1 + Realsense	12.6	**18.0**	23.3	**26.8**	24.3	**24.7**	10.9	**12.2**
Kinect v1 + Xtion	11.7	**23.9**	29.6	**35.7**	20.3	**21.1**	16.7	**20.0**
Kinect v2 + Realsense	7.4	**11.7**	22.7	**33.1**	28.4	**29.4**	6.9	**9.1**
Kinect v2 + Xtion	14.8	**16.2**	27.0	**31.0**	**25.4**	25.0	11.6	**18.3**
Realsense + Xtion	8.3	**10.7**	23.1	**25.2**	30.1	**31.5**	9.5	**10.8**
Average	12.1	**16.6**	26.7	**31.4**	24.6	**25.3**	11.7	**14.7**

Results On SUN RGB-D. for both single and multi source knowledge transfer. On average SOCKET outperforms the single source baseline SHOT [1] and multi-source baseline DECISION [2] for all four target domains by good margins.

Learning Losses

1) Task-irrelevant feature matching

$$\mathcal{L}_{TI} = \sum_{i=1}^{n_{TI}} \sum_{j=1}^{n} \|\zeta_j(\psi_j^i - f_j(x_{TI_i}^{m_T}))\|^2 \qquad \text{where} \qquad \psi_j^i = f_j(x_{TI_i}^{m_S})$$

This loss helps reducing the modality gap by using external Task-irrelevant source data.

2) Task-relevant distribution matching

$$\mathcal{L}_d = \sum_{l=1}^{b} \left(\| \sum_{j=1}^{n} \zeta_j \mathbb{E}[\mu_l | \mathcal{X}_{S_j}^{m_S}] - \sum_{j=1}^{n} \zeta_j \hat{\mu}_{l_j} \| + \| \sum_{j=1}^{n} \zeta_j \mathbb{E}[\sigma_l^2 | \mathcal{X}_{S_j}^{m_S}] - \sum_{j=1}^{n} \zeta_j \hat{\sigma^2}_{l_j} \| \right)$$

This loss matches the Task-relevant feature statistics from the BN layers across the source and target, to reduce the modality gap further .

3) Modality agnostic unsupervised losses

$$\mathcal{L}_{ent} = -\frac{1}{n_T} \left[\sum_{i=1}^{n_T} (\mathcal{F}_T^{m_T}(x_T^i)) \log(\mathcal{F}_T^{m_T}(x_T^i)) \right], \mathcal{L}_{div} = -\sum_{j=1}^{N} \bar{p}_j \log \bar{p}_j$$

$$\mathcal{L}_{pl} = -\frac{1}{n_T} \sum_{i=1}^{n_T} \sum_{k=1}^{K} \mathbf{1}\{\hat{y}_T^i = k\} \log \left[\mathcal{F}_T^{m_T}(x_T^i) \right]_k \qquad \text{where}$$

$$\mathcal{F}_T^{m_T}(x_T^i) = \sum_{k=1}^{n} \zeta_k \mathcal{F}_{S_k}^{m_S}(x_T^i) \text{ and } \bar{p} = \frac{1}{n_T} \sum_{i=1}^{n_T} \left[\mathcal{F}_T^{m_T}(x_T^i) \right]$$

Modality agnostic losses: entropy, diversity and pseudo-label loss respectively widely used in standard source-free UDA settings .

Overall Optimization

$$\min_{\{f_j\}_{j=1}^n, \zeta} \quad \mathcal{L}_{ent} - \mathcal{L}_{div} + \lambda_{pl}\mathcal{L}_{pl} + \lambda_{TI}\mathcal{L}_{TI} + \lambda_d \mathcal{L}_d$$

$$\text{s.t.} \quad \sum_{k=1}^{n} \zeta_k = 1, \zeta_k \geq 0$$

Acknowledgements

SMA, SL, KCP and MJ were supported by Mitsubishi Electric Research Laboratories. SMA and ARC were partially supported by ONR grant N00014-19-1-2264 and the NSF grants CCF-2008020 and IIS-1724341.

References

[1] Liang, Jian, Dapeng Hu, and Jiashi Feng. "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation." International Conference on Machine Learning. PMLR, 2020.

[2] Ahmed, S.M., Raychaudhuri, D.S., Paul, S., Oymak, S., Roy-Chowdhury, A.K.: Unsupervised multi-source domain adaptation without access to source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 10103–10112