Reinforced Temporal Attention and Split-Rate Transfer for Depth-Based Person Re-Identification

Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, Stefano Soatto

{nikarian, zliu, yiche}@microsoft.com, soatto@cs.ucla.edu

ECCV 2018 European Conference on Computer Vision

Highlights

- First, we enhance re-identification from video by **implementing temporal attention as a Bernoulli-Sigmoid unit** acting upon frame-level features. The introduced unit is trained end-to-end with reinforcement learning and thus it is termed as **Reinforced Temporal Attention (RTA)**.
- Second, we address **data scarcity** in depth-based person re-identification by introducing **Split-Rate Transfer** from large RGB data. Our scheme encourages parameter sharing at the bottom CNN layers between RGB and depth data while the remaining layers are rapidly fine-tuned from RGB.

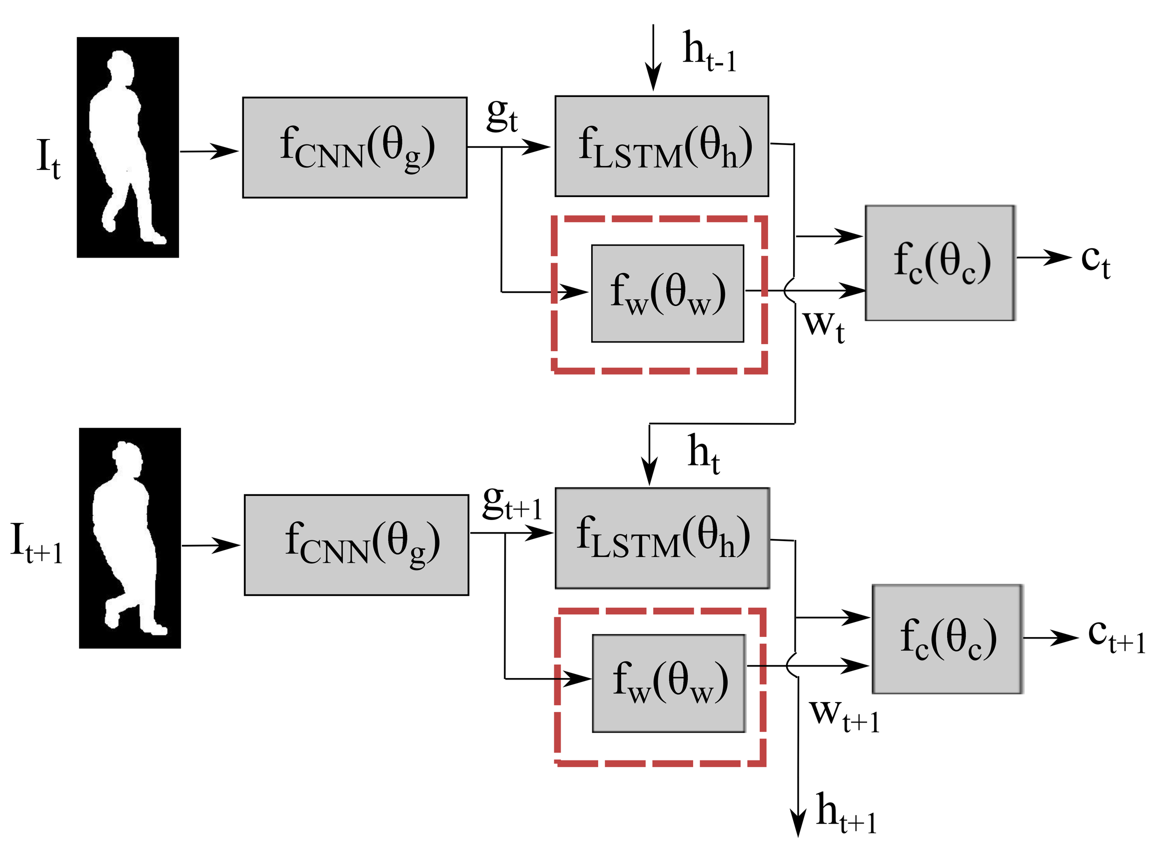
Reinforced Temporal Attention

- We extend Recurrent Attention models [1, 2] to temporal domain by modeling temporal attention as a Bernoulli-Sigmoid stochastic unit:

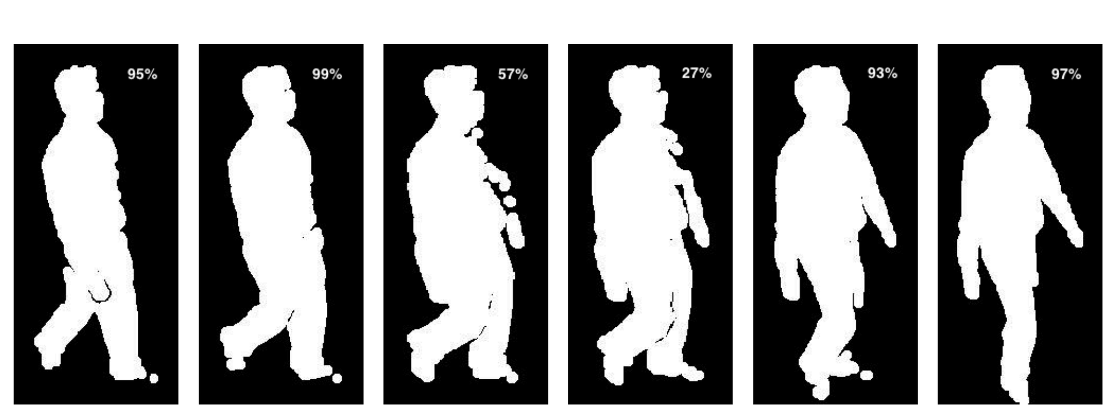
$$f(w_t; f_w(g_t;\theta_w)) = \begin{cases} f_w(g_t;\theta_w), & w_t = 1 \\ 1 - f_w(g_t;\theta_w), & w_t = 0 \end{cases}$$

- Due to non-differentiability a gradient sample approximation is used:

$$\nabla_{\theta_g,\theta_w} J = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\theta_g,\theta_w)}[\nabla_{\theta_g,\theta_w} \log \pi_1(w_t|s_{1:t};\theta_g,\theta_w)(R_t - b_t)]$$
$$\approx \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T} \frac{w_t^i - p_t^i}{p_t^i(1-p_t^i)}(R_t^i - b_t)$$

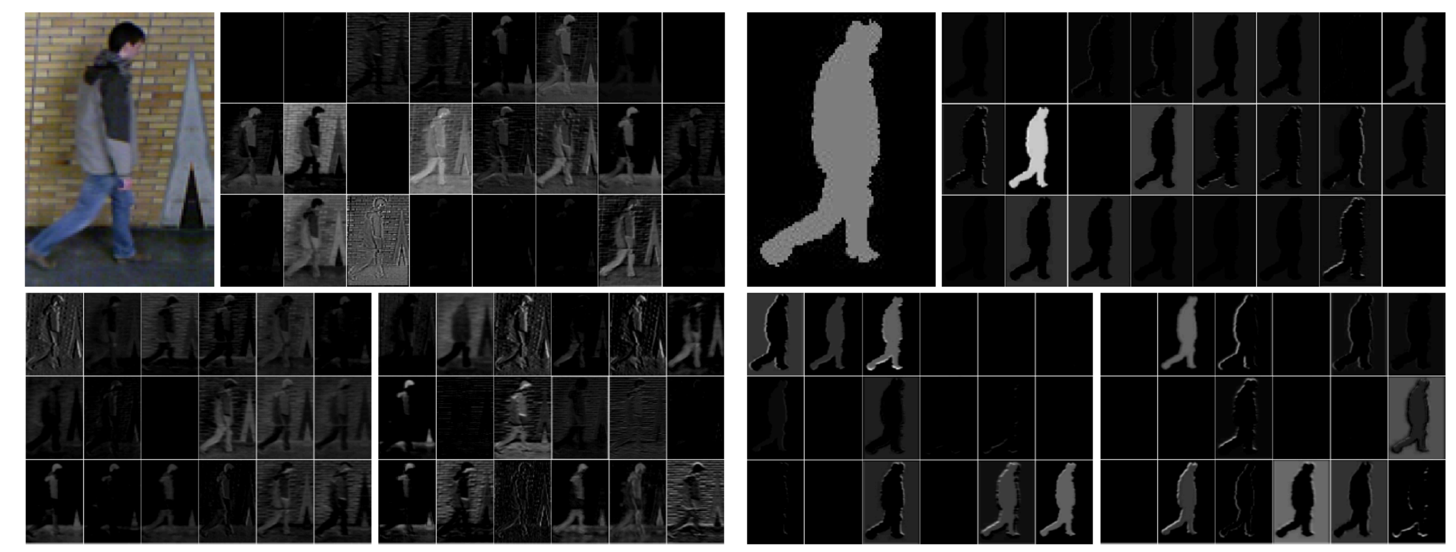


Schematic diagram of the end-to-end model with RTA drawn in red.



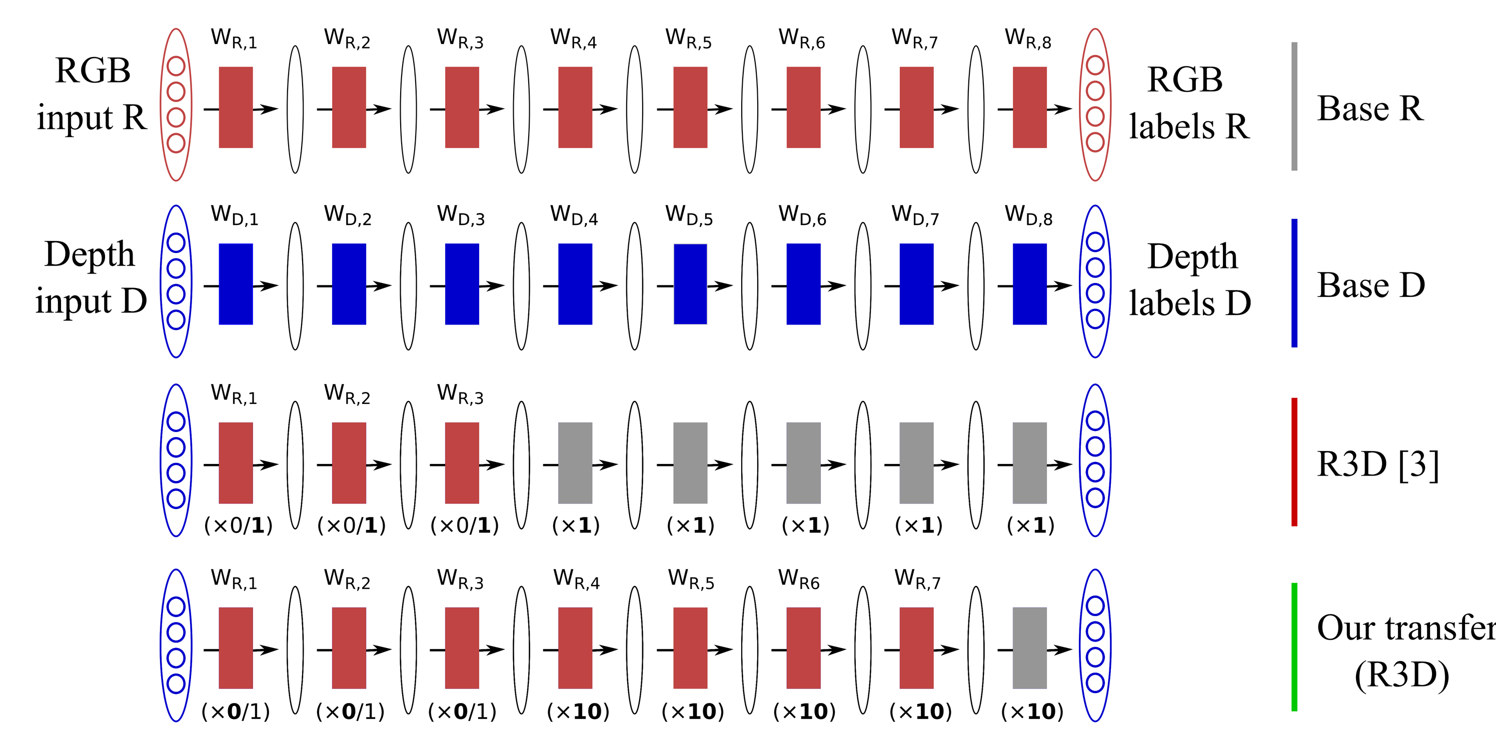
Example sequence with the predicted Bernoulli parameter.

RGB to Depth transfer learning



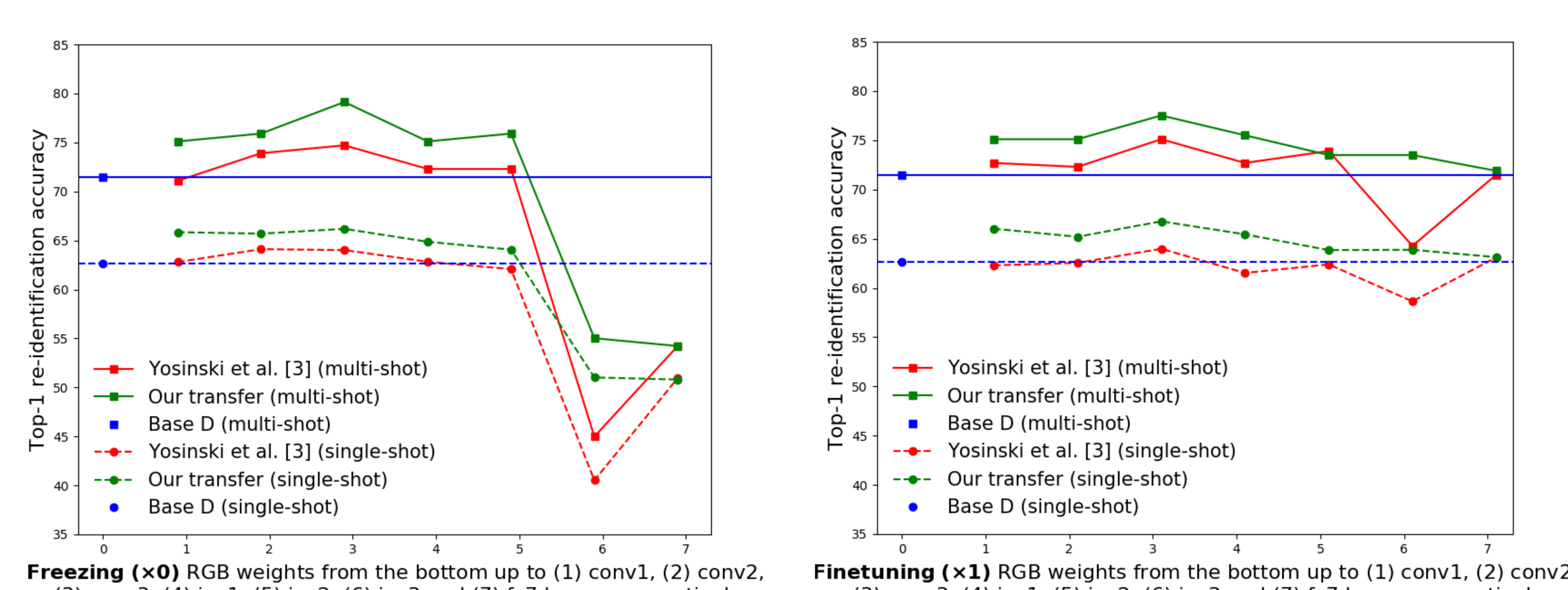
Person ReID from RGB   Person ReID from Depth

Split-Rate Transfer



Our transfer scheme has 3 key differences compared to fine-tuning [3]:

1. Despite apparent differences between RGB and depth data, their **bottom layers can be directly shared**.
2. **Fine-tuning** from RGB works **better than re-training** for top layers.
3. Using **lower** (or 0) learning rate for the bottom layers and **higher** for the top layers is **better than using uniform rate** across the hierarchy.

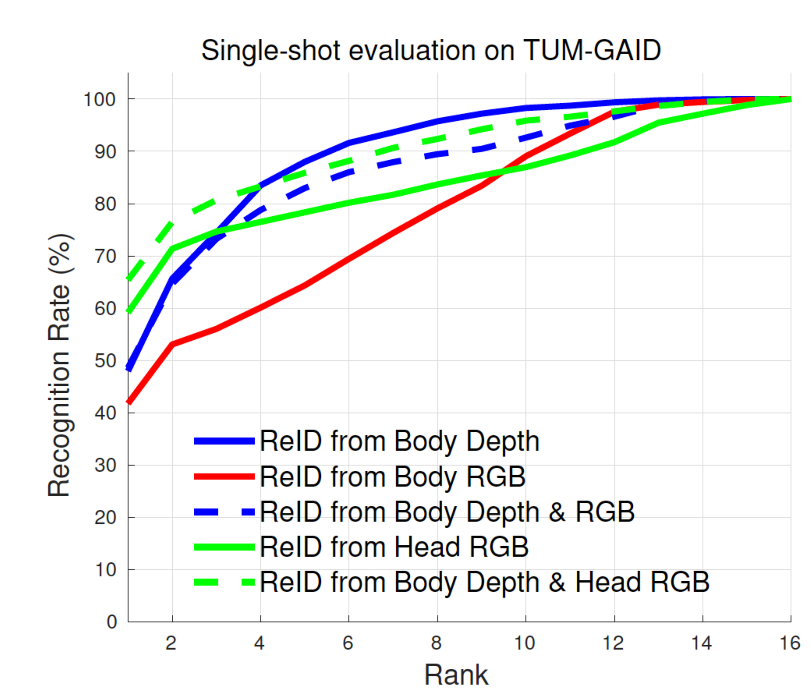


Comparison of our RGB-to-Depth transfer with Yosinski et al [3] in terms of top-1 accuracy on DPI-T.

Comparisons in Person ReID from Depth

Mode	Method	Top-1 Accuracy (%) DPI-T	BIWI	IIT PAVIS
	Random	8.3	2.0	1.3
Single-shot	Skeleton (NN)	–	21.1	28.6
	Skeleton (SVM)	–	13.8	35.7
	3D RAM [2]	47.5	**30.1**	41.3
	Our method (CNN)	**66.8**	25.4	**43.0**
Multi-shot	Skeleton (NN)	–	39.3	–
	Skeleton (SVM)	–	17.9	–
	Energy Volume	14.2	25.7	18.9
	3D CNN+Avg Pooling	28.4	27.8	27.5
	4D RAM [2]	55.6	45.3	43.0
	Our method (CNN-LSTM+Avg Pooling)	75.5	45.7	50.1
	Our method with RTA attention	**76.3**	**50.0**	**52.4**

- Methods that learn **end-to-end features perform much better** than the ones that rely on hand-crafted biometrics on all datasets.
- Our algorithm is the **top performer in multi-shot mode**, as our RTA unit effectively learns to re-weight the most effective frames based on a task-specific reward.
- We note that spatial attention is also important in datasets with significant variation in human pose and partial body occlusions, as in BIWI, but less critical on DPI-T, which contains views from the top and the visible region is mostly uniform across frames.



Modality	top-1	nAUC
Body RGB (ss)	41.8	74.3
Body Depth (ss)	48.0	**85.0**
Body Depth & RGB (ss)	48.6	81.9
Head RGB (ss)	59.4	79.5
Body Depth & Head RGB (ss)	**65.4**	85.2
Body RGB (ms: LSTM & RTA)	50.0	79.9
Body Depth (ms: LSTM)	56.3	87.7
Body Depth (ms: LSTM & RTA)	59.4	**89.6**
Head RGB (ms: LSTM & RTA)	65.6	81.0
Body Depth & Head RGB (ms: LSTM & RTA)	**75.0**	88.1

- In scenario with **unseen clothes**, Depth-based ReID more robust, while combined with head information performs the best.

References

1. *Recurrent models of visual attention*. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K. (NIPS, 2014)
2. *Recurrent attention models for depth-based person identification*. Haque, A., Alahi, A., Fei-Fei, L. (CVPR, 2016)
3. *How transferable are features in deep neural networks?* Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (NIPS, 2014)