Museum Exhibit Identification Challenge for the Supervised Domain Adaptation and Beyond.

Authors: Piotr.Koniusz[1,2] Yusuf.Tas[1,2] Hongguang.Zhang[2,1]
Mehrtash.Harandi[3] Fatih.Porikli[2] Rui Zhang[4]
[1]@data61.csiro.au [2]@anu.edu.au [3]@monash.edu.au [4]renata_zhang@sina.com

## *** The contents ***

- Motivation (saturated performance).
- Open MIC dataset (details, challenges).
- Our supervised domain adaptation pipeline+Results
- Our few-shot learning pipeline+Results
- Conclusions.

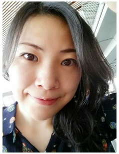
[Koniusz et al., ECCV'18]
[Zhang & Koniusz, WACV'19]


Yusuf

Hongguang


Mehrtash


Fatih


Rui

# Motivation

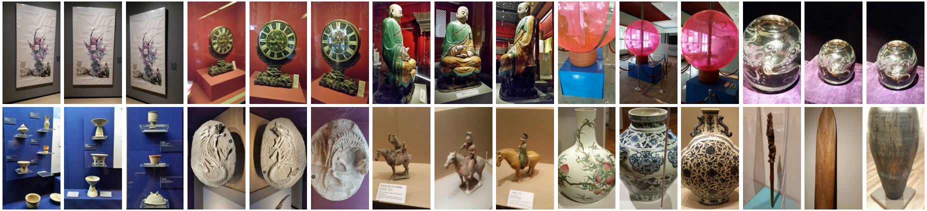
- Results on Office 31 dataset [K. Saenko et al., ECCV'10] reached $\sim$90% accuracy (still a good dataset for the sanity check!).
- New dataset Open Museum Identification Challenge (Open MIC) to stimulate research in domain adaptation, egocentric recognition and few-shot learning.
- 866 unique exhibit labels, 8560 source and 7596 target images.
- Open MIC: photos of exhibits captured in 10 distinct exhibition spaces of several museums which showcase paintings, timepieces, sculptures, glassware, relics, science exhibits, natural history pieces, ceramics, pottery, tools and indigenous crafts.

# Motivation

- Results on Office 31 dataset [K. Saenko et al., ECCV'10] reached ∼90% accuracy (still a good dataset for the sanity check!).
- New dataset Open Museum Identification Challenge (Open MIC) to stimulate research in domain adaptation, egocentric recognition and few-shot learning.
- 866 unique exhibit labels, 8560 source and 7596 target images.
- Open MIC: photos of exhibits captured in 10 distinct exhibition spaces of several museums which showcase paintings, timepieces, sculptures, glassware, relics, science exhibits, natural history pieces, ceramics, pottery, tools and indigenous crafts.
- Museums contain some of the most visually diverse objects. Cannot find a lot of wearable data of them on Flickr or YouTube.
- We study artwork identification in the context of:
    - supervised/unsupervised domain adaptation
    - one- and/or few-shot learning (follow up paper)

# Open MIC

- Source domain: we captured photos in a controlled fashion by Android phones *e.g.*, each exhibit is centered and non-occluded.
- We captured 2–30 photos per art piece from different viewpoints and distances:



Source subsets of Open MIC.

(Top) Paintings (*Shn*), Clocks (*Clk*), Sculptures (*Scl*), Science Exhibits (*Sci*) and Glasswork (*Gls*).

(Bottom) Cultural Relics (*Rel*), Natural History Exhibits (*Nat*), Historical/Cultural Exhibits (*Shx*), Porcelain (*Clv*) and Indigenous Arts (*Hon*).

# Open MIC

- Target domain: in-the-wild capture, wearable cameras took a photo every 10s.
- We captured varied materials *e.g.*, rigid, non-rigid, emitting light, in motion, extremely small or composite installations:

Examples of the target subsets of Open MIC. From left to right, each column illustrates one exhibition.

Paintings (*Shn*), Clocks (*Clk*), Sculptures (*Scl*), Science Exhibits (*Sci*) and Glasswork (*Gls*), Cultural Relics (*Rel*), Natural History Exhibits (*Nat*), Historical/Cultural Exhibits (*Shx*), Porcelain (*Clv*) and Indigenous Arts (*Hon*).

# Open MIC

- Our target exhibits various photometric and geometric challenges *e.g.*, sensor noises, motion blur, occlusions, background clutter, varying viewpoints, scale changes, rotations, glares, transparency, non-planar surfaces, clipping, multiple exhibits, active light, color inconsistency, zoomed in/out photos, intra-exhibit variations:

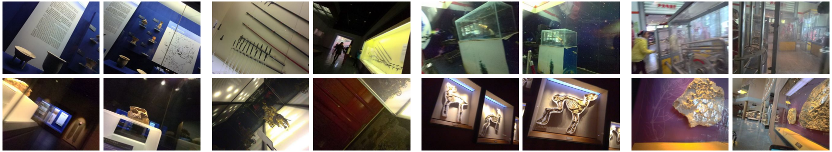


Illustration of the significant domain shift from the source to target.

# Open MIC

- Our target exhibits various photometric and geometric challenges *e.g.*, sensor noises, motion blur, occlusions, background clutter, varying viewpoints, scale changes, rotations, glares, transparency, non-planar surfaces, clipping, multiple exhibits, active light, color inconsistency, zoomed in/out photos, intra-exhibit variations:

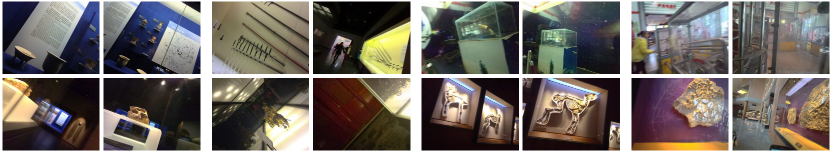


Illustration of the significant domain shift from the source to target.

- Some of the hardest to identify instances in Open MIC:

# Open MIC

- Supervised Domain Adaptation:
  - Use small or large source data (lebelled).
  - Transfer to improve recognition on scarce target data (lebelled).
  - Ultimately: beat combined source+target training and/or fine-tuning.
  - Not all is big data! Quote: learning quickly from only a few examples is definitely the desired characteristic to emulate in any brain-like system [Rajapakse & Wang, Research & Development, 2004].

# Open MIC

- Supervised Domain Adaptation:
  - Use small or large source data (lebelled).
  - Transfer to improve recognition on scarce target data (lebelled).
  - Ultimately: beat combined source+target training and/or fine-tuning.
  - Not all is big data! Quote: learning quickly from only a few examples is definitely the desired characteristic to emulate in any brain-like system [Rajapakse & Wang, Research & Development, 2004].
- Evaluation protocols include:
  - training/evaluation per exhibition subset (10 exhibitions)
  - training/testing on the combined set of all 866 identities
  - testing w.r.t. various scene factors: quality of lighting, motion blur, occlusions, clutter, viewpoint and scale variations, rotations, glares, transparency, non-planarity, clipping
  - unsupervised domain adaptation ($\pm$videoclips)
- Accuracy measure we use:
  - top-$k$-$n$ tells if any of top $n$ ground-truth labels per image are contained in top $k$ predictions.

# Open MIC

One-shot protocols include:

- training on combined target sets (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) which give subproblems *p1*, ..., *p4*.
  We form 12 possible pairs: subproblem *x* is used for training and *y* for testing (x→y).
  (generalization from one task to another task)

# Open MIC

One-shot protocols include:

- training on combined target sets (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) which give subproblems *p1*, ..., *p4*.
  We form 12 possible pairs: subproblem $x$ is used for training and $y$ for testing (x→y).
  (generalization from one task to another task)
- training on each source exhibition and testing on the corresponding target exhibition (generalization from one domain to another domain: does few-shot learning cope with the domain shift?).

# Open MIC

One-shot protocols include:

- training on combined target sets (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) which give subproblems *p1*, ..., *p4*.
  We form 12 possible pairs: subproblem $x$ is used for training and $y$ for testing (x→y).
  (generalization from one task to another task)
- training on each source exhibition and testing on the corresponding target exhibition (generalization from one domain to another domain: does few-shot learning cope with the domain shift?).
- training on combined source sets and testing on non-corresponding target sets (gen. from one task to another task and from one domain to another domain).

# Open MIC

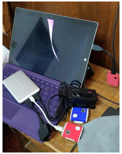
One-shot protocols include:

- training on combined target sets (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) which give subproblems *p1*, ..., *p4*.
  We form 12 possible pairs: subproblem $x$ is used for training and $y$ for testing (x$\rightarrow$y).
  (generalization from one task to another task)
- training on each source exhibition and testing on the corresponding target exhibition (generalization from one domain to another domain: does few-shot learning cope with the domain shift?).
- training on combined source sets and testing on non-corresponding target sets (gen. from one task to another task and from one domain to another domain).
- Evaluation is performed for so called K-shot L-way problems (L-way means choosing L random classes for each episode: generalization from task to task)
- Episode=query training image + $K \times L$ support images

# Open MIC

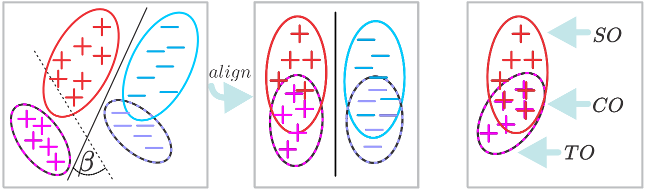
One-shot protocols include:

- training on combined target sets (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) which give subproblems *p1*, ..., *p4*.
  We form 12 possible pairs: subproblem *x* is used for training and *y* for testing (x$\rightarrow$y).
  (generalization from one task to another task)
- training on each source exhibition and testing on the corresponding target exhibition (generalization from one domain to another domain: does few-shot learning cope with the domain shift?).
- training on combined source sets and testing on non-corresponding target sets (gen. from one task to another task and from one domain to another domain).
- Evaluation is performed for so called K-shot L-way problems (L-way means choosing L random classes for each episode: generalization from task to task)
- Episode=query training image + $K \times L$ support images
- Charging all wearable cameras is the hardest part but ...
- We plan to release next iteration of the dataset
  (20 exhibition spaces: some challenging subsets such as fossils)



# DA pipeline

- We build on the *So-HoT* model [Koniusz et al., CVPR'17] posed as a trade-off between the classifier $\ell$ and source-target alignment loss $\hbar$.
- Essentially, a trade-off between within- and between-class statistics (LDA)
- Idea: establish so-called commonality between class-wise stats. of source and target.
- The commonality: partial alignment of statistics (full alignment is bad assumption).



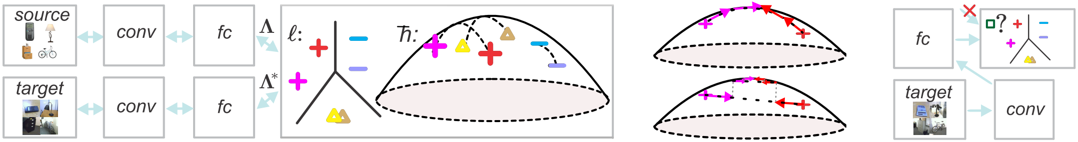

(a) (b) (c)

## Alignment problem:

- How to separate two classes + and - for two domains given $\beta$.
- Partially aligned distributions have the commonality (*CO*).
- Source and target specific parts (*SO*) and (*TO*) – dissimilarity between source/target.

# DA pipeline

- We combine the source and target CNN streams:




(a) (b) (c)

DA pipeline:
*(a)* Source/target streams $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$ merge at the classifier level.
*(b)* Loss $\hbar$ aligns covariances on the manifold of $\mathcal{S}_{++}$ matrices.
*(c)* At the test time, we use the target stream and the trained classifier.

- For alignment of covariances, the Euclidean distance is suboptimal in the light of Riemannian geometry.

# DA pipeline

- The loss $\hbar$ depends on two sets of variables $(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_C)$ and $(\mathbf{\Phi}^*_1, ..., \mathbf{\Phi}^*_C)$ – one set per network stream.
- $\mathbf{\Phi}(\mathbf{\Theta})$ and $\mathbf{\Phi}^*(\mathbf{\Theta}^*)$ depend on parameters of the source/target streams $\mathbf{\Theta}$ and $\mathbf{\Theta}^*$ that we optimize over.
- $\mathbf{\Sigma}_c \equiv \mathbf{\Sigma}(\mathbf{\Phi}_c)$, $\mathbf{\Sigma}^*_c \equiv \mathbf{\Sigma}(\mathbf{\Phi}^*_c)$, $\boldsymbol{\mu}_c(\mathbf{\Phi})$ and $\boldsymbol{\mu}^*_c(\mathbf{\Phi}^*)$ denote the covariances and means, respectively. We solve:

$$\begin{aligned}
&\underset{\substack{\boldsymbol{W},\boldsymbol{W}^*\!,\mathbf{\Theta},\mathbf{\Theta}^*\\ \text{s. t. } ||\boldsymbol{\phi}_n||_2^2\leq\tau,\\ ||\boldsymbol{\phi}^*_{n'}||_2^2\leq\tau,\\ \forall n\in\mathcal{I}_N, n'\in\mathcal{I}^*_N}}{\arg\min} \ \ell(\boldsymbol{W},\mathbf{\Lambda})+\ell(\boldsymbol{W}^*\!,\mathbf{\Lambda}^*)+\eta||\boldsymbol{W}-\boldsymbol{W}^*||_F^2 + \\
&\underbrace{\frac{\alpha_1}{C}\sum_{c\in\mathcal{I}_C} d^2\left(\mathbf{\Sigma}_c,\mathbf{\Sigma}^*_c\right)+\frac{\alpha_2}{C}\sum_{c\in\mathcal{I}_C}||\boldsymbol{\mu}_c-\boldsymbol{\mu}^*_c||_2^2}_{\hbar(\mathbf{\Phi},\mathbf{\Phi}^*)}.
\end{aligned} \tag{1}$$

# DA pipeline

- The loss $\hbar$ depends on two sets of variables $(\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_C)$ and $(\boldsymbol{\Phi}^*_1, ..., \boldsymbol{\Phi}^*_C)$ – one set per network stream.
- $\boldsymbol{\Phi}(\boldsymbol{\Theta})$ and $\boldsymbol{\Phi}^*(\boldsymbol{\Theta}^*)$ depend on parameters of the source/target streams $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$ that we optimize over.
- $\boldsymbol{\Sigma}_c \equiv \boldsymbol{\Sigma}(\boldsymbol{\Phi}_c)$, $\boldsymbol{\Sigma}^*_c \equiv \boldsymbol{\Sigma}(\boldsymbol{\Phi}^*_c)$, $\boldsymbol{\mu}_c(\boldsymbol{\Phi})$ and $\boldsymbol{\mu}^*_c(\boldsymbol{\Phi}^*)$ denote the covariances and means, respectively. We solve:

$$\begin{aligned}&\underset{\substack{\boldsymbol{W},\boldsymbol{W}^*\!,\boldsymbol{\Theta},\boldsymbol{\Theta}^*\\ \text{s. t. } ||\phi_n||_2^2\leq\tau,\\ ||\phi^*_{n'}||_2^2\leq\tau,\\ \forall n\in\mathcal{I}_N, n'\in\mathcal{I}^*_N}}{\arg\min}\ \ell(\boldsymbol{W},\boldsymbol{\Lambda})+\ell(\boldsymbol{W}^*,\boldsymbol{\Lambda}^*)+\eta||\boldsymbol{W}-\boldsymbol{W}^*||_F^2 + \\ &\underbrace{\frac{\alpha_1}{C}\sum_{c\in\mathcal{I}_C} d^2\left(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}^*_c\right)+\frac{\alpha_2}{C}\sum_{c\in\mathcal{I}_C}||\boldsymbol{\mu}_c-\boldsymbol{\mu}^*_c||_2^2}_{\hbar(\boldsymbol{\Phi},\boldsymbol{\Phi}^*)}.\end{aligned} \tag{1}$$

- For alignment of covariances/SPD matrices, the Euclidean distance is suboptimal in the light of Riemannian geometry.

| Dist./Ref. | $d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$ | Invar. | Tr. Ineq. | Geo. | $d$ if $\mathcal{S}_+$ | $\nabla_{\boldsymbol{\Sigma}}$ if $\mathcal{S}_+$ | $\frac{\partial d^2(\boldsymbol{\Sigma},\boldsymbol{\Sigma}^*)}{\partial\boldsymbol{\Sigma}}$ |
|---|---|---|---|---|---|---|---|
| Frobenius | $\|\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^*\|\|_F^2$ | rot. | yes | no | fin. | fin. | $2(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^*)$ |
| AIRM | $\|\|\log(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-\frac{1}{2}})\|\|_F^2$ | aff./inv. | yes | yes | $\infty$ | $\infty$ | $-2\boldsymbol{\Sigma}^{-\frac{1}{2}}\log(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-\frac{1}{2}})\boldsymbol{\Sigma}^{-\frac{1}{2}}$ |
| JBLD | $\log\left\|\frac{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*}{2}\right\|-\frac{1}{2}\log\|\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\|$ | aff./inv. | no | no | $\infty$ | $\infty$ | $(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*)^{-1}-\frac{1}{2}\boldsymbol{\Sigma}^{-1}$ |

We use Affine Inv. Riemannian Metric (AIRM) and Jensen-Bregman LogDet Divergence (JBLD).

# DA pipeline

- For GPU/CPU, SVD of large matrices ($d \geq 2048$) in CUDA BLAS is extremely slow.
- Idea: we exploit the low-rank nature of our covariance matrices + low number of datapoints (RKHS-friendly setting).
- For typical $N \approx 30$, $N^* \approx 3$, we get $33 \times 33$ dim. covariances rather than $4096 \times 4096$.

# DA pipeline

- For GPU/CPU, SVD of large matrices ($d \geq 2048$) in CUDA BLAS is extremely slow.
- Idea: we exploit the low-rank nature of our covariance matrices + low number of datapoints (RKHS-friendly setting).
- For typical $N \approx 30$, $N^* \approx 3$, we get $33 \times 33$ dim. covariances rather than $4096 \times 4096$.
- For each class $c \in \mathcal{I}_C$, we choose $\boldsymbol{X} = \boldsymbol{Z} = [\boldsymbol{\Phi}_c, \boldsymbol{\Phi}^*_c]$.
- From the Nyström projection, we obtain:
  $\boldsymbol{\Pi}(\boldsymbol{X}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-0.5}\boldsymbol{Z}^T\boldsymbol{X} = \cancel{\boldsymbol{Z}}\boldsymbol{X} = (\boldsymbol{Z}^T\boldsymbol{Z})^{0.5} = (\boldsymbol{X}^T\boldsymbol{X})^{0.5}$.
- Then $\boldsymbol{\Pi}(\boldsymbol{\Phi}) = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$ and $\boldsymbol{\Pi}(\boldsymbol{\Phi}^*) = [\boldsymbol{y}_{N+1}, ..., \boldsymbol{y}_{N+N*}]$.

# DA pipeline

- For GPU/CPU, SVD of large matrices ($d \geq 2048$) in CUDA BLAS is extremely slow.
- Idea: we exploit the low-rank nature of our covariance matrices + low number of datapoints (RKHS-friendly setting).
- For typical $N \approx 30$, $N^* \approx 3$, we get $33 \times 33$ dim. covariances rather than $4096 \times 4096$.
- For each class $c \in \mathcal{I}_C$, we choose $\boldsymbol{X} = \boldsymbol{Z} = [\boldsymbol{\Phi}_c, \boldsymbol{\Phi}^*_c]$.
- From the Nyström projection, we obtain:
  $\boldsymbol{\Pi}(\boldsymbol{X}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-0.5}\boldsymbol{Z}^T\boldsymbol{X} = \cancel{\boldsymbol{Z}}\boldsymbol{X} = (\boldsymbol{Z}^T\boldsymbol{Z})^{0.5} = (\boldsymbol{X}^T\boldsymbol{X})^{0.5}$.
- Then $\boldsymbol{\Pi}(\boldsymbol{\Phi}) = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$ and $\boldsymbol{\Pi}(\boldsymbol{\Phi}^*) = [\boldsymbol{y}_{N+1}, ..., \boldsymbol{y}_{N+N*}]$.
- $\boldsymbol{\Pi}(\boldsymbol{X})$ is isometric w.r.t. AIRM/JBLD, that is
  $d^2_g(\boldsymbol{\Sigma}(\boldsymbol{\Phi}), \boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)) = d^2_g(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})), \boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))$ (!!!)

# DA pipeline

- For GPU/CPU, SVD of large matrices ($d \geq 2048$) in CUDA BLAS is extremely slow.
- Idea: we exploit the low-rank nature of our covariance matrices + low number of datapoints (RKHS-friendly setting).
- For typical $N \approx 30$, $N^* \approx 3$, we get $33 \times 33$ dim. covariances rather than $4096 \times 4096$.
- For each class $c \in \mathcal{I}_C$, we choose $\boldsymbol{X} = \boldsymbol{Z} = [\boldsymbol{\Phi}_c, \boldsymbol{\Phi}^*_c]$.
- From the Nyström projection, we obtain: $\boldsymbol{\Pi}(\boldsymbol{X}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-0.5}\boldsymbol{Z}^T\boldsymbol{X} = \cancel{\boldsymbol{Z}}\boldsymbol{X} = (\boldsymbol{Z}^T\boldsymbol{Z})^{0.5} = (\boldsymbol{X}^T\boldsymbol{X})^{0.5}$.
- Then $\boldsymbol{\Pi}(\boldsymbol{\Phi}) = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$ and $\boldsymbol{\Pi}(\boldsymbol{\Phi}^*) = [\boldsymbol{y}_{N+1}, ..., \boldsymbol{y}_{N+N*}]$.
- $\boldsymbol{\Pi}(\boldsymbol{X})$ is isometric w.r.t. AIRM/JBLD, that is $d^2_g(\boldsymbol{\Sigma}(\boldsymbol{\Phi}), \boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)) = d^2_g(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})), \boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))$ (!!!)
- $\cancel{\boldsymbol{Z}}(\boldsymbol{X})$ can be treated as a constant in differentiation $\frac{\partial \boldsymbol{\Pi}(\boldsymbol{X})}{\partial X_{mn}} = \frac{\partial \cancel{\boldsymbol{Z}}(\boldsymbol{X})\boldsymbol{X}}{\partial X_{mn}} = \cancel{\boldsymbol{Z}}(\boldsymbol{X})\frac{\partial \boldsymbol{X}}{\partial X_{mn}} = \cancel{\boldsymbol{Z}}(\boldsymbol{X})\boldsymbol{J}_{mn}$ (!!!)
- Our proof shows that $\cancel{\boldsymbol{Z}}$ is a composite rotation (!!!) and the Euclidean, JBLD and AIRM distances are rotation-invariant (!!!), hence isometry (!!!)

We provide baselines such as:

- Fine-tuning CNNs on the source subsets (*S*) and testing on the randomly chosen target splits.
- Fine tuning on target only (*T*) and evaluating on remaining disjoint target splits.
- Fine-tuning on the source+target (*S+T*) and evaluating on remaining disjoint target splits.
- Training state-of-the-art domain adaptation So-HoT algorithm equipped by us with non-Euclidean distances (*So*).

| | *Shn* | *Clk* | *Scl* | *Sci* | *Gls* | *Rel* | *Nat* | *Shx* | *Clv* | *Hon* | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Inst.* | 79 | 113 | 41 | 37 | 98 | 100 | 111 | 166 | 81 | 40 | 866 |
| *Src.* | 417 | 650 | 160 | 391 | 575 | 587 | 695 | 2697 | 503 | 970 | 7645 |
| *Tgt.* | 404 | 305 | 112 | 1342 | 863 | 863 | 668 | 546 +307K fr | 625 | 364 +73K fr | 6092 +380K fr |

Unique exhibit instances (*Inst.*) and numbers of images in the source (*Src.*) and target (*Tgt.*) splits of Open MIC.

# Experiments

## Evaluation protocols include:

- Training/evaluation per exhibition subset (10 exhibitions).
- Training/testing on the combined set of all 866 identities.
- Testing w.r.t. various scene factors: quality of lighting, motion blur, occlusions, clutter, viewpoint and scale variations, rotations, glares, transparency, non-planarity, clipping.
- Unsupervised domain adaptation ($\pm$videoclips).

| | | *S* | *T* | S+T | *JBLD* | | *S* | *T* | S+T | *JBLD* | | *S* | *T* | S+T | *JBLD* | | *S* | *T* | S+T | *JBLD* | | *S* | *T* | S+T | *JBLD* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| top-1 | *Shn* | 47.7 | 51.6 | 58.3 | **64.3** | *Clk* | 56.9 | 49.1 | 56.0 | **61.2** | *Scl* | 53.5 | 52.2 | 54.3 | **54.4** | *Sci* | 58.5 | 58.1 | 64.9 | **66.8** | *Gls* | 15.8 | 70.2 | 72.6 | **74.4** |
| top-1-5 | | 48.2 | 54.2 | 60.2 | **66.4** | | 58.9 | 56.3 | 60.3 | **68.9** | | 54.7 | 55.4 | 57.3 | **58.4** | | 60.2 | 61.7 | 67.8 | **70.2** | | 19.4 | 85.1 | 86.0 | **89.0** |
| top-1 | *Rel* | 18.1 | 66.1 | 63.2 | **67.0** | *Nat* | 41.6 | 57.3 | 57.9 | **62.7** | *Shx* | 29.9 | 41.1 | 29.0 | **48.5** | *Clv* | 47.0 | 65.2 | 62.2 | **69.1** | *Hon* | 66.7 | 67.6 | 73.4 | **77.3** |
| top-1-5 | | 24.0 | 76.8 | 73.2 | **79.5** | | 43.5 | 62.8 | 61.9 | **67.7** | | 31.5 | 47.7 | 31.9 | **56.3** | | 50.8 | 69.5 | 66.6 | **73.9** | | 70.2 | 70.3 | 76.3 | **79.7** |

Challenge I. Open MIC accuracies on 10 subsets. Baselines (*S*), (*T*), (*S+T*), and JBLD are given.

| | *So* | *JBLD* | *AIRM* |
|---|---|---|---|
| sp1 | 55.8 | **57.7** | 57.2 |
| sp2 | 58.9 | **58.9** | 58.9 |
| sp3 | 69.6 | **71.4** | 71.4 |
| sp4 | 53.8 | **57.7** | 57.7 |
| sp5 | 58.3 | **60.4** | 60.4 |
| acc. | 59.3 | **61.2** | 61.1 |

AIRM vs. JBLD.

| | sp1 | sp2 | sp3 | sp4 | sp5 | top-1 | top-1-5 |
|---|---|---|---|---|---|---|---|
| *S* | 33.9 | 34.2 | 34.8 | 34.2 | 33.8 | 34.2 | 36.0 |
| *T* | 56.9 | 55.9 | 58.7 | 56.0 | 55.2 | 56.5 | 64.1 |
| *S+T* | 56.4 | 55.2 | 57.1 | 56.3 | 54.4 | 55.9 | 62.5 |
| *So* | 64.2 | 62.4 | 65.0 | 62.7 | 60.0 | 62.8 | 70.4 |
| *JBLD* | **65.7** | **63.8** | **65.7** | **63.7** | **62.0** | **64.2** | **72.0** |

Challenge II. Perf. on the whole dataset.

# Experiments

| | clp | lgt | blr | glr | bgr | ocl | rot | zom | vpc | sml | shd | rfl | ok |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *S* | 41.4 | 17.0 | 23.8 | 27.3 | 40.3 | 34.5 | 29.7 | 52.7 | 33.4 | 14.2 | 10.4 | 32.3 | 65.5 |
| *T* | 56.2 | 38.2 | 42.6 | 56.1 | 57.9 | 49.6 | 58.3 | 60.4 | 50.3 | 29.6 | 59.2 | 60.7 | 64.3 |
| *S+T* | 56.6 | 34.6 | 39.8 | 54.9 | 56.2 | 48.3 | 56.7 | 65.9 | 48.7 | 27.3 | 56.5 | 59.0 | 72.6 |
| *JBLD* | **65.3** | **48.6** | **51.6** | **64.0** | **65.9** | **56.4** | **65.0** | **70.0** | **58.6** | **34.1** | **70.4** | **67.5** | **81.0** |

Challenge III. Performance w.r.t. 12 distortion factors.

| ∩ | clp | lgt | blr | glr | bgr | ocl | rot | zom | vpc | sml | shd | rfl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | 65.3 | 48.6 | 51.6 | 64.0 | 65.9 | 56.4 | 65.0 | 70.0 | 58.6 | 34.1 | 70.4 | 67.5 |
| clp | 65.3 | 55.1 | 51.8 | 67.5 | 66.8 | 61.5 | 67.2 | 68.1 | 62.3 | 45.5 | 72.7 | 67.0 |
| lgt | 55.1 | 48.6 | 41.0 | 43.6 | 59.8 | 43.5 | 48.3 | 44.4 | 46.1 | 31.2 | 57.9 | 80.9 |
| blr | 51.8 | 41.0 | 51.6 | 48.7 | 48.6 | 37.0 | 52.3 | 64.2 | 43.3 | 21.0 | 39.1 | 59.4 |
| glr | 67.5 | 43.6 | 48.7 | 64.0 | 62.3 | 47.9 | 65.1 | 67.1 | 60.4 | 13.5 | 50.0 | 64.5 |
| bgr | 66.8 | 59.8 | 48.6 | 62.3 | 65.9 | 59.6 | 66.6 | 76.1 | 61.2 | 29.9 | 79.6 | 73.2 |
| ocl | 61.5 | 43.5 | 37.0 | 47.9 | 59.6 | 56.4 | 55.6 | 75.4 | 55.9 | 40.7 | 78.8 | 64.8 |
| rot | 67.2 | 48.3 | 52.3 | 65.1 | 66.6 | 55.6 | 65.0 | 75.5 | 57.6 | 32.6 | 73.4 | 70.4 |
| zom | 68.1 | 44.4 | 64.2 | 67.1 | 76.1 | 75.4 | 75.5 | 70.0 | 66.3 | n/a | 83.3 | 69.7 |
| vpc | 62.3 | 46.1 | 43.3 | 60.4 | 61.2 | 55.9 | 57.6 | 66.3 | 58.6 | 33.2 | 64.1 | 61.6 |
| sml | 45.5 | 31.2 | 21.0 | 13.5 | 29.9 | 40.7 | 32.6 | n/a | 33.2 | 34.1 | n/a | 46.4 |
| shd | 72.7 | 57.9 | 39.1 | 50.0 | 79.6 | 78.8 | 73.4 | 83.3 | 64.1 | n/a | 70.4 | 80.0 |
| rfl | 67.0 | 80.9 | 59.4 | 64.5 | 73.2 | 64.8 | 70.4 | 69.7 | 61.6 | 46.4 | 80.0 | 67.5 |

Accuracy w.r.t. pairs of 12 factors.

| ∩ | sml glr | sml blr | sml bgr | sml lgt | sml rot | sml vpc | blr ocl | blr shd | sml ocl | lgt blr | lgt ocl | lgt glr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | 13.5 | 21.0 | 29.9 | 31.2 | 32.6 | 33.2 | 37.0 | 39.1 | 40.7 | 40.9 | 43.5 | 43.6 |
| clp | 42.8 | 27.8 | 38.7 | 66.7 | 42.8 | 46.0 | 44.4 | 53.8 | 45.5 | 49.1 | 45.1 | 45.7 |
| lgt | 0.0 | 30.0 | 40.0 | 31.2 | 37.5 | 50.0 | 52.3 | 38.5 | 10.0 | 40.9 | 43.5 | 43.6 |
| blr | 0.0 | 21.0 | 18.2 | 30.0 | 24.6 | 17.8 | 37.0 | 39.1 | 11.1 | 40.9 | 52.2 | 21.0 |
| glr | 13.5 | 0.0 | 7.7 | 0.0 | 10.5 | 15.0 | 27.8 | 33.3 | 27.8 | 21.0 | 31.2 | 43.6 |
| bgr | 7.7 | 18.2 | 29.9 | 40.0 | 27.7 | 31.4 | 37.2 | 60.0 | 33.0 | 46.1 | 51.4 | 42.1 |
| ocl | 15.0 | 11.1 | 33.0 | 14.3 | 39.7 | 41.0 | 37.0 | 83.3 | 40.7 | 52.2 | 43.5 | 31.2 |
| rot | 10.2 | 24.6 | 27.7 | 37.5 | 32.6 | 31.8 | 38.0 | 50.0 | 39.7 | 43.0 | 60.0 | 32.2 |
| zom | n/a | n/a | n/a | n/a | n/a | n/a | 75.0 | 100 | n/a | 100 | n/a | n/a |
| vpc | 15.0 | 17.8 | 31.4 | 50.0 | 31.8 | 33.2 | 35.3 | 58.3 | 41.0 | 35.3 | 40.4 | 46.0 |
| sml | 13.5 | 21.0 | 29.9 | 31.2 | 32.6 | 33.2 | 11.1 | n/a | 40.7 | 30.0 | 14.3 | 0.0 |
| shd | n/a | n/a | n/a | n/a | n/a | n/a | 83.3 | 39.1 | n/a | 38.5 | 75.0 | 50.0 |
| rfl | 75.0 | 50.0 | 39.3 | n/a | 46.3 | 45.2 | 69.6 | 100 | 68.2 | 100 | 50.0 | 100 |

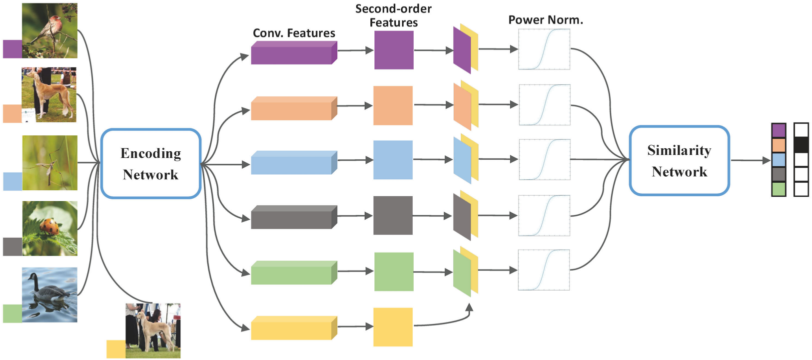
Accuracy w.r.t. selected triplets of 12 factors.

| | Shn | Clk | Scl | Sci | Gls | Rel | Nat | Shx | Clv | Hon | top-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *IHS* | 47.1 | 61.9 | 50.8 | 63.3 | 26.0 | 32.6 | 51.0 | 22.0 | 61.2 | 67.3 | 48.3 |
| *RTN* | 54.4 | 59.0 | 65.2 | 62.2 | 30.5 | 24.8 | 44.2 | 32.1 | 47.7 | 71.1 | 49.1 |
| *JAN* | 51.7 | 63.6 | 67.8 | 69.8 | 34.2 | 28.5 | 47.1 | 32.0 | 53.9 | 72.5 | **52.1** |

Challenge IV. Unsupervised Domain Adaptation.

- Invariant Hilbert Space (*IHS*) [S. Herath et al., CVPR'17].
- Unsupervised Domain Adaptation with Residual Transfer Networks (*RTN*) [M. Long et al. NIPS'16].
- Deep Transfer Learning with Joint Adaptation Networks (*JAN*) [M. Long et al. ICML'17].

# Few-shot learning pipeline




We propose Second-order Similarity Network (SoSN):

- The image encoding network.
- Second-order relation descriptors with Power Normalization.
- Similarity learning network (simple metric learning).

# Experiments

## Evaluations on the Open MIC dataset (Protocol I).

| Model | $L$ | $p1\rightarrow p2$ | $p1\rightarrow p3$ | $p1\rightarrow p4$ | $p2\rightarrow p1$ | $p2\rightarrow p3$ | $p2\rightarrow p4$ | $p3\rightarrow p1$ | $p3\rightarrow p2$ | $p3\rightarrow p4$ | $p4\rightarrow p1$ | $p4\rightarrow p2$ | $p4\rightarrow p3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relation Net | 5 | 71.1$\pm$1.0 | 53.6$\pm$1.1 | 63.5$\pm$1.0 | 47.2$\pm$1.0 | 50.6$\pm$1.1 | 68.5$\pm$1.0 | 48.5$\pm$1.1 | 49.7$\pm$1.1 | 68.4$\pm$1.0 | 45.5$\pm$1.0 | 70.3$\pm$1.0 | 50.8$\pm$1.1 |
| SoSN | | 80.8$\pm$0.9 | 64.3$\pm$1.1 | 74.9$\pm$1.1 | 58.8$\pm$1.1 | 61.2$\pm$1.1 | 76.9$\pm$0.9 | 61.3$\pm$1.1 | 80.8$\pm$0.9 | 77.2$\pm$1.0 | 58.2$\pm$1.1 | 80.1$\pm$0.9 | 61.6$\pm$1.1 |
| SoSN+SigmE | | **81.4**$\pm$0.9 | **65.2**$\pm$1.1 | **75.1**$\pm$1.0 | **60.3**$\pm$1.1 | **62.1**$\pm$1.1 | **77.7**$\pm$0.9 | **61.5**$\pm$1.1 | **82.0**$\pm$1.0 | **78.0**$\pm$1.0 | **59.0**$\pm$1.1 | **80.8**$\pm$1.0 | **62.5**$\pm$1.1 |
| SoSN+SigmE+224x224 | | **83.9**$\pm$0.9 | **68.9**$\pm$1.1 | **82.1**$\pm$0.9 | **64.7**$\pm$1.1 | **66.6**$\pm$1.1 | **82.2**$\pm$0.9 | **65.5**$\pm$1.1 | **84.5**$\pm$0.8 | **80.6**$\pm$0.8 | **64.6**$\pm$1.1 | **83.6**$\pm$0.8 | **66.0**$\pm$1.1 |
| Relation Net | 20 | 40.1$\pm$0.5 | 30.4$\pm$0.5 | 41.4$\pm$0.5 | 23.5$\pm$0.4 | 26.4$\pm$0.5 | 38.6$\pm$0.5 | 26.2$\pm$0.4 | 25.8$\pm$0.4 | 46.3$\pm$0.5 | 23.1$\pm$0.4 | 43.3$\pm$0.5 | 27.7$\pm$0.4 |
| SoSN | | 61.0$\pm$0.5 | 42.3$\pm$0.5 | 60.2$\pm$0.5 | 35.7$\pm$0.5 | 37.0$\pm$0.5 | 54.8$\pm$0.5 | 36.0$\pm$0.5 | 59.1$\pm$0.5 | 57.0$\pm$0.5 | 36.4$\pm$0.5 | 59.3$\pm$0.9 | 37.8$\pm$0.5 |
| SoSN+SigmE | | **61.5**$\pm$0.6 | **42.5**$\pm$0.5 | **61.0**$\pm$0.5 | **36.1**$\pm$0.5 | **38.3**$\pm$0.5 | **56.3**$\pm$0.5 | **38.7**$\pm$0.5 | **59.9**$\pm$0.6 | **59.4**$\pm$0.5 | **37.4**$\pm$0.5 | **59.0**$\pm$0.5 | **38.6**$\pm$0.5 |
| SoSN+SigmE+224x224 | | **63.6**$\pm$0.5 | **48.7**$\pm$0.6 | **65.6**$\pm$0.5 | **42.6**$\pm$0.5 | **43.9**$\pm$0.5 | **61.8**$\pm$0.5 | **43.7**$\pm$0.5 | **63.3**$\pm$0.5 | **63.5**$\pm$0.5 | **43.2**$\pm$0.5 | **62.5**$\pm$0.5 | **43.7**$\pm$0.5 |
| SoSN+SigmE | 30 | 60.6$\pm$0.6 | 40.1$\pm$0.7 | 58.3$\pm$0.4 | 34.5$\pm$0.5 | 35.1$\pm$0.6 | 54.2$\pm$0.6 | 36.8$\pm$0.6 | 58.6$\pm$0.7 | 56.6$\pm$0.7 | 35.9$\pm$0.7 | 57.1$\pm$0.7 | 37.1$\pm$0.6 |
| SoSN+SigmE+224x224 | | 61.7$\pm$0.7 | 46.6$\pm$0.6 | 64.1$\pm$0.6 | 41.4$\pm$0.6 | 40.9$\pm$0.6 | 60.3$\pm$0.6 | 41.6$\pm$0.6 | 61.0$\pm$0.7 | 60.0$\pm$0.6 | 42.4$\pm$0.6 | 61.2$\pm$0.6 | 41.4$\pm$0.6 |
| SoSN+SigmE | 45 | 53.3$\pm$0.5 | 37.3$\pm$0.5 | 54.6$\pm$0.5 | 30.8$\pm$0.4 | 32.4$\pm$0.5 | 52.4$\pm$0.5 | 32.1$\pm$0.5 | 54.2$\pm$0.5 | 51.1$\pm$0.5 | 30.5$\pm$0.4 | 51.9$\pm$0.5 | 33.4$\pm$0.5 |
| SoSN+SigmE+224x224 | | 59.7$\pm$0.5 | 40.5$\pm$0.5 | 57.9$\pm$0.5 | 36.5$\pm$0.5 | 38.2$\pm$0.5 | 55.7$\pm$0.5 | 39.5$\pm$0.5 | 56.6$\pm$0.4 | 56.0$\pm$0.5 | 37.4$\pm$0.5 | 55.5$\pm$0.5 | 38.5$\pm$0.5 |
| SoSN+SigmE | 60 | 51.2$\pm$0.4 | 34.6$\pm$0.4 | 49.1$\pm$0.5 | 28.4$\pm$0.4 | 31.1$\pm$0.4 | 48.2$\pm$0.5 | 30.1$\pm$0.4 | 50.0$\pm$0.4 | 48.3$\pm$0.5 | 30.0$\pm$0.4 | 49.2$\pm$0.5 | 30.6$\pm$0.4 |
| SoSN+SigmE+224x224 | | 48.2$\pm$0.6 | 36.0$\pm$0.5 | 54.4$\pm$0.5 | 30.7$\pm$0.4 | 32.4$\pm$0.5 | 52.2$\pm$0.5 | 32.35$\pm$0.4 | 51.0$\pm$0.5 | 51.6$\pm$0.5 | 32.7$\pm$0.5 | 53.6$\pm$0.5 | 35.7$\pm$0.4 |
| SoSN+SigmE | 90 | 45.6$\pm$0.3 | 29.7$\pm$0.3 | 45.5$\pm$0.4 | 24.5$\pm$0.3 | 26.3$\pm$0.3 | 43.6$\pm$0.3 | 26.4$\pm$0.3 | 44.2$\pm$0.3 | 43.2$\pm$0.3 | 25.5$\pm$0.3 | 46.0$\pm$0.3 | 27.5$\pm$0.3 |
| SoSN+SigmE+224x224 | | 47.3$\pm$0.3 | 33.4$\pm$0.3 | 49.8$\pm$0.3 | 25.3$\pm$0.4 | 27.1$\pm$0.4 | 47.0$\pm$0.4 | 27.1$\pm$0.4 | 45.7$\pm$0.4 | 48.9$\pm$0.5 | 28.1$\pm$0.3 | 46.7$\pm$0.5 | 31.6$\pm$0.3 |

p1: shn+hon+clv, p2: clk+gls+scl, p3: sci+nat, p4: shx+rlc. Notation $x\rightarrow y$ means training on exhibition $x$ and testing on $y$.

- One-shot classification (realistic one-shot scenario, task-shift only). We go up to 90-way (typically 5- or 20-way protocols used on *mini*-ImageNet not exciting).
- As $L$-way number increases, we see that few-shot learning has some way to go (some results reach only $\sim$25% accuracy).
- Relation Net [F. Sung et al., CVPR'18], SoSN: our Second-order Similarity Network, SoSN+SigmE: SoSN+Power Normalization, $224\times224$: image resolution (typically few-shot uses $84\times84$).

# Experiments

## Evaluations on the Open MIC dataset for Protocol II (asterisk $^*L'$ indicates splits with the number of classes $L' < L$).

| Model | $L$ | *shn* | *hon* | *clv* | *clk* | *gls* | *scl* | *sci* | *nat* | *shx* | *rlc* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Relation Net | 5 | 43.2±1.0 | 49.6±1.0 | 49.8±1.0 | 62.1±1.1 | 59.3±1.0 | 51.5±1.0 | 45.9±1.0 | 54.8±1.0 | 71.1±1.0 | 72.0±1.0 |
| SoSN | | 60.3±1.1 | 62.6±1.1 | 60.5±1.1 | 72.9±1.1 | 74.3±1.1 | 72.3±1.0 | 53.4±1.1 | 68.0±1.1 | 77.0±1.0 | 78.4±1.0 |
| SoSN+SigmE | | 61.5±1.1 | 63.6±1.1 | 61.7±1.1 | 74.5±1.2 | 74.9±1.1 | 72.9±1.0 | 54.2±1.0 | 68.9±1.1 | 78.0±1.0 | 79.1±1.0 |
| Relation Net | 20 | 20.8±0.4 | 25.7±0.4 | 26.1±0.4 | 34.3±0.4 | 35.5±0.5 | 18.4±0.3 | 18.6±0.3 | 32.8±0.5 | 51.8±0.5 | 48.2±0.5 |
| SoSN | | 36.3±0.5 | 36.4±0.5 | 33.3±0.4 | 48.5±0.5 | 54.3±0.5 | 54.1±0.5 | 24.8±0.4 | 44.0±0.5 | 59.5±0.5 | 54.2±0.5 |
| SoSN+SigmE | | 37.4±0.5 | 37.5±0.5 | 34.9±0.4 | 49.6±0.5 | 55.2±0.5 | 55.5±0.5 | 25.1±0.4 | 45.3±0.5 | 61.9±0.5 | 56.6±0.5 |
| Relation Net | 30 | 18.1±0.3 | 21.1±0.3 | 23.2±0.3 | 27.0±0.3 | 31.8±0.4 | 12.8±0.2 | 12.4±0.2 | 27.1±0.3 | 40.6±0.4 | 41.0±0.4 |
| SoSN | | 34.2±0.4 | 35.2±0.4 | 32.7±0.3 | 46.7±0.4 | 51.0±0.4 | 52.2±0.4 | 20.3±0.3 | 39.9±0.4 | 56.7±0.4 | 51.0±0.4 |
| SoSN+SigmE | | 35.5±0.4 | 36.0±0.4 | 33.5±0.3 | 47.7±0.5 | 52.3±0.4 | 53.0±0.3 | 21.1±0.3 | 40.8±0.4 | 58.3±0.4 | 52.7±0.5 |
| SoSN+SigmE+224x224 | | 41.4±0.6 | 39.4±0.7 | 37.2±0.6 | 51.3±0.7 | 53.4±0.7 | 59.0±0.6 | 23.3±0.5 | 46.7±0.7 | 59.8±0.6 | 55.4±0.6 |
| SoSN+SigmE | 45 | 34.1±0.5 | 33.4±0.4 (*39) | 29.2±0.5 | 45.2±0.5 | 48.5±0.5 | 49.6±0.5 (*42) | 19.2±0.4 (*36) | 38.0±0.5 | 54.1±0.6 | 49.3±0.5 |
| SoSN+SigmE+224x224 | | 34.9±0.4 | 34.5±0.4 | 30.7±0.5 | 50.5±0.5 | 39.9±0.6 | 50.6±0.5 | 20.1±0.4 | 41.9±0.5 | 54.6±0.5 | 52.1±0.5 |
| SoSN+SigmE | 60 | 30.0±0.4 | - | 25.5±0.4 | 42.6±0.5 | 46.6±0.4 | - | - | 37.5±0.4 | 51.3±0.5 | 46.6±0.4 |
| SoSN+SigmE+224x224 | | 34.5±0.4 | | 28.3±0.4 | 47.9±0.5 | 47.4±0.5 | | | 37.9±0.3 | 52.0±0.4 | 47.4±0.4 |
| SoSN+SigmE | 90 | 26.4±0.3 (*78) | - | 24.6±0.3 (*80) | 41.8±0.3 | 39.2±0.3 | - | - | 33.0±0.3 | 49.4±0.5 | 39.5±0.3 |
| SoSN+SigmE+224x224 | | 33.2±0.3 | | 27.5±0.3 | 44.5±0.3 | 40.2±0.3 | | | 34.6±0.3 | 50.4±0.6 | 42.6±0.3 |

Training on source images and testing on target images for every exhibition, respectively.

- The goal of this protocol it to test how few-shot learning algorithms deal with the domain shift.
- Even for low $L$-way number *e.g.*, 30, Relation Net scores only ∼12–20%. SoSN is more robust (∼40–50% accuracy) but there is still some way to go to reach 100%.

# Conclusions (Thank You)

- New challenging dataset for domain adaptation and few-shot learning (Open MIC)
- We have interesting evaluation protocols for DA: supervised/unsupervised DA, per-exhibition and combined protocols, breakdowns w.r.t. factors impairing recognition, even one-shot learning protocol.
- We have interesting evaluation protocols for few-shot learning: within-domain protocol using target combined splits (generalization from task to task), between-domain protocol using original exhibitions (generalization from domain to domain), between-task between-domain protocol III (we are evaluating it now).
- We plan to extend this dataset to detection, segmentation, saliency detection, deblurring, *etc.*
- Our dataset is available for the academic use on claret.wikidot.com or http://users.cecs.anu.edu.au/∼koniusz.