Fine-Grained Visual Comparisons with Local Learning

Aron Yu Kristen Grauman

University of Texas at Austin

Visual Comparisons

Which shoe is more *sporty*?

Problem:

Fine-grained visual comparisons require accounting for *subtle visual differences* specific to each comparison pair.

Status Quo: Learning a Global Ranking Function

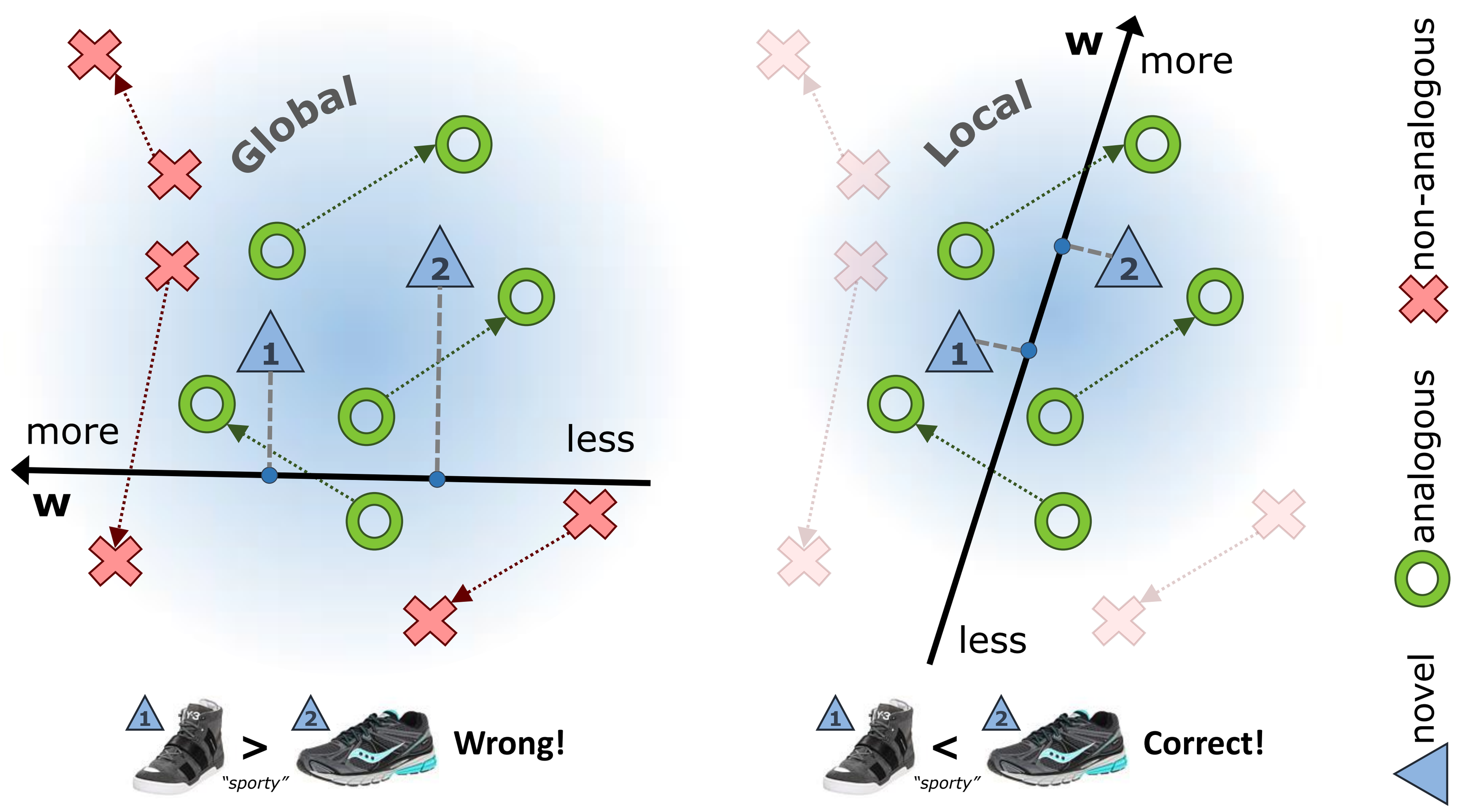
[Parikh & Grauman 11, Datta et al. 11, Li et al. 12, Kovashka et al. 12, ...]

- fails to account for subtle differences among closely related images
- each comparison pair exhibits unique visual cues/rationales
- visual comparisons need not be *transitive*

Our Approach

Project webpage here ➔

We propose a **local learning** approach for fine-grained comparisons.



- learn attribute-specific distance metrics
- identify top K *analogous* neighboring pairs w.r.t. each novel pair
- train local function that tailors to the neighborhood statistics

Key Idea: having the *right* data > having *more* data

Analogous Neighboring Pairs

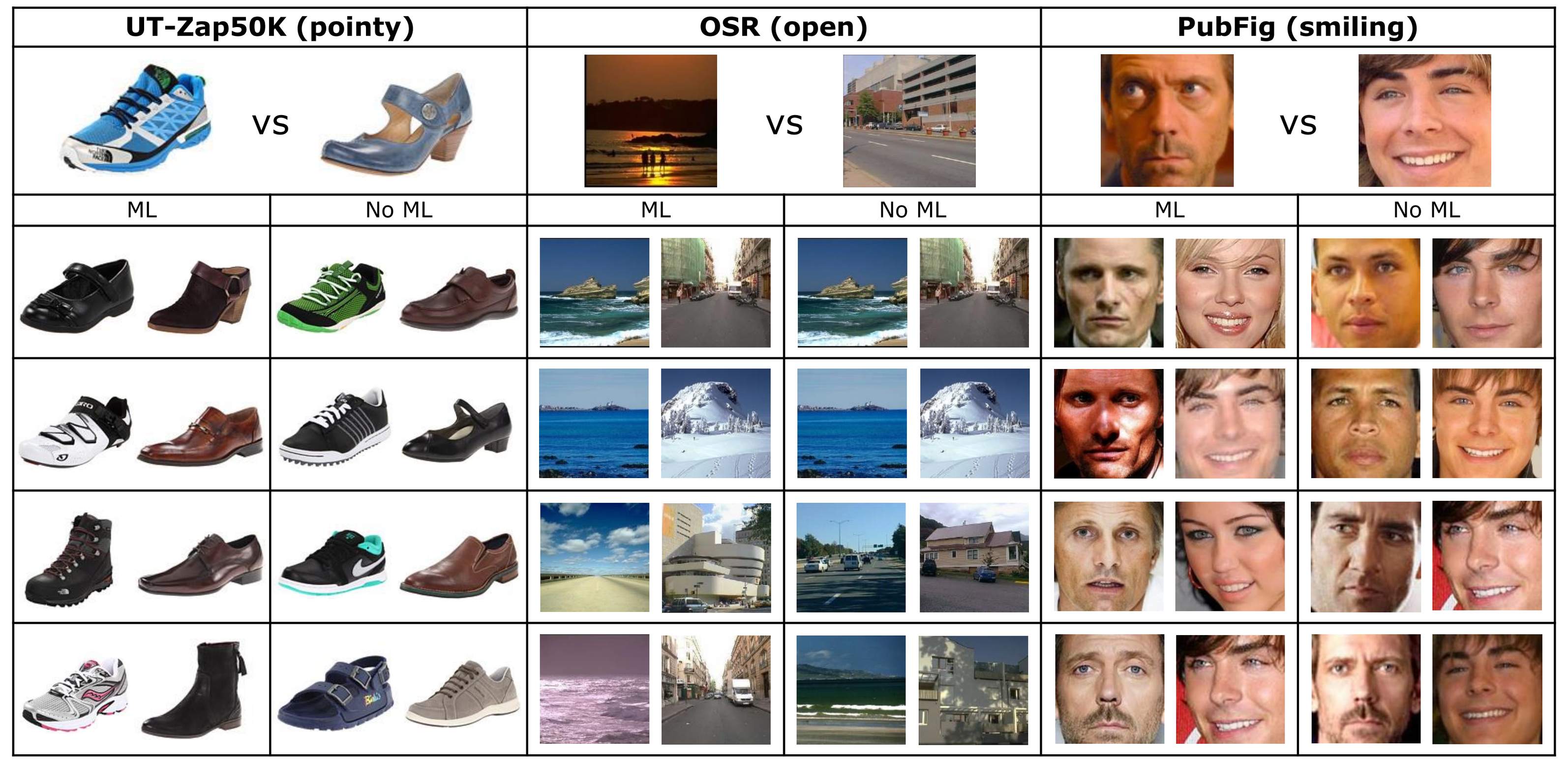
Detect analogous pairs based on *individual similarity* & *paired contrast*.

- select neighboring pairs that accentuate fine-grained differences
- take *product* of pairwise distances of individual members
- i.e. highly analogous if both query-training couplings are similar

Learned Attribute Distance

Learn a **Mahalanobis** metric per attribute (similarity computation).

- attribute similarity doesn't rely equally on each dim of feature space
- constraints ➔ similar images be close, dissimilar images be far



Observation: Nearest *analogous* pairs most suited for local learning need not be those closest in raw feature space.

UT Zappos50K Dataset

Get dataset here ➔

We introduce a new large shoe dataset **UT-Zap50K**, consisting of **50,025** catalog images from Zappos.com.

- 4 relative attributes (open, pointy, sporty, comfort)
- high confidence pairwise labels from mTurk workers
- 6,751 ordered labels + 4,612 "equal" labels
- 4,334 twice-labeled *fine-grained* labels (no "equal" option)

Results: UT-Zap50K

- **FG-LocalPair:** our proposed fine-grained approach
- **LocalPair:** our approach w/o the learned metric
- **RandPair:** local approach with random neighbors
- **Global**[Parikh & Grauman 11]**:** status quo of learning a single global ranking function per attribute
- **RelTree**[Li et al. 12]**:** non-linear relative attribute approach

Accuracy Comparison *(10 iterations @ K=100)*

- coarser comparisons

	Open	Pointy	Sporty	Comfort
Global	87.77	89.37	91.20	89.93
RandPair	82.53	83.70	86.30	84.77
LocalPair	88.53	88.87	92.20	90.90
FG-LocalPair	**90.67**	**90.83**	**92.67**	**92.37**

- fine-grained comparisons

	Open	Pointy	Sporty	Comfort
Global	60.18	59.56	62.70	**64.04**
RandPair	61.00	53.41	58.26	59.24
LocalPair	71.64	59.56	61.22	59.75
FG-LocalPair	**74.91**	**63.74**	**64.54**	62.51

1

Ours ✓ – Global ✗

We detect subtle changes while global relies only on the overall shape and color, often leading to incorrect decisions for fine-grained comparisons.

2

Ours ✗ – Global ✓

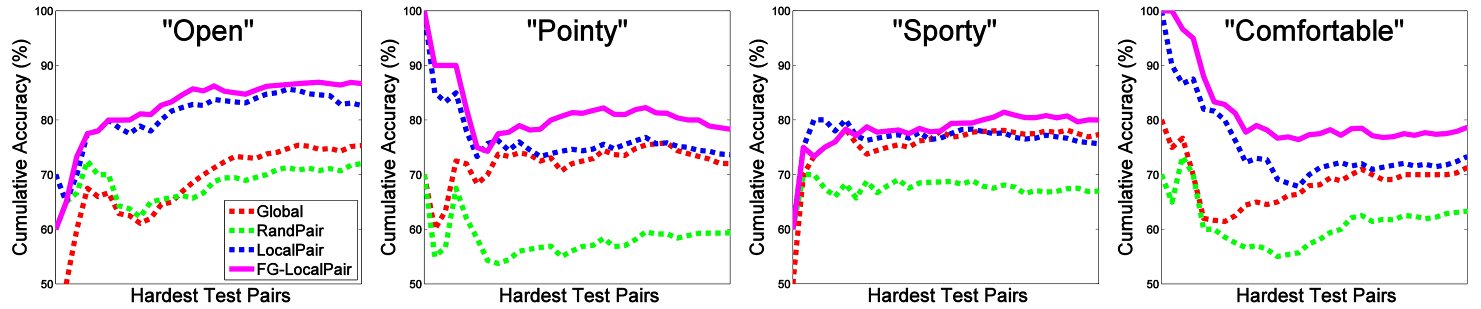
These coarser differences are sufficiently captured by a global model w/o the need for fine-grained details.

3

Ours ✗ – Global ✗

Such pairs are so fine-grained that they are difficult even for humans to make a firm consistent decision.

- accuracy for the **30 hardest** test pairs (according to learned metrics)

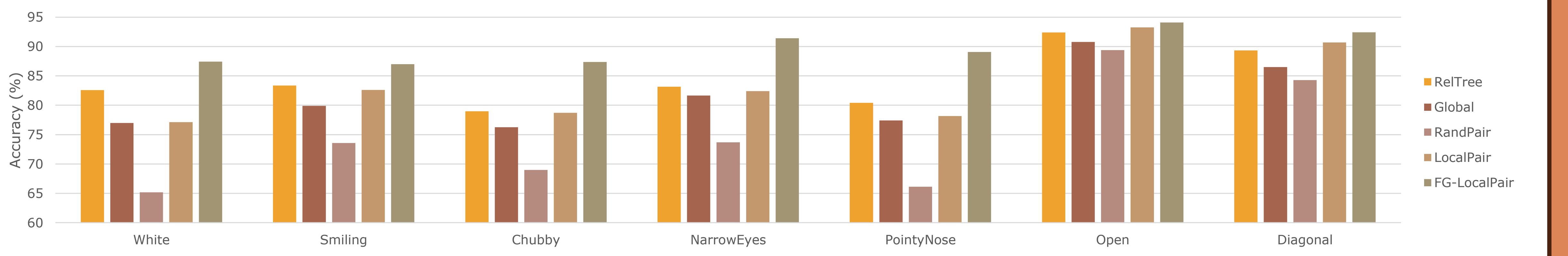


Observation:
We outperform all baselines, demonstrating strong advantage for detecting subtle differences on the harder comparisons (~20% more).

Results: PubFig & Scenes

We form supervision pairs using the category-wise comparisons ➔ avg. 20,000 ordered labels / attribute.

- **Public Figures Face (PubFig):** 772 images w/ 11 attributes
- **Outdoor Scene Recognition (OSR):** 2,688 images w/ 6 attributes



Observation: We outperform the current state of the art on 2 popular relative attribute datasets. Our gains are especially dominant on localizable attributes due to the learned metrics.