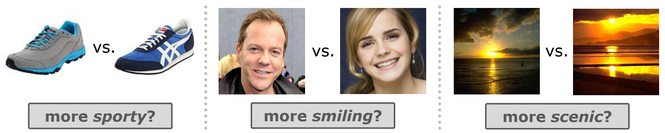
Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images

Aron Yu Kristen Grauman
University of Texas at Austin

Fine-Grained Visual Comparisons



Existing Approaches

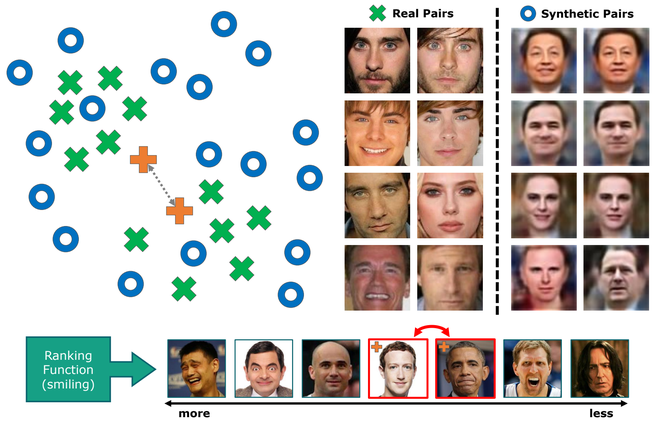
- focus on improving the ranking algorithms
  [Yang et al. '16, Souri et al. '16, Singh & Lee '16, Yu & Grauman '14, Li et al. '12, ...]
- existing datasets contain *insufficient representation of fine-grained differences* using real images

Problem: Sparsity of Supervision

1) pairwise supervision → quadratic # of potential pairs (label availability)
2) lack a direct way to curate the "right" data for *optimal coverage* of the attribute space (image availability)

Our Idea

Densify the attribute space using *synthetic image pairs* to improve supervision for fine-grained learning.



Densifying Supervision

Generate synthetic images exhibiting *subtle differences*

- "fill in" the *sparsely sampled regions* to enhance fine-grained supervision
- pre-trained **Attribute2Image** [Yan et al. '16] image generation engine
- attribute-conditioned generation of synthetic identities $I_j = (y_j, z_j)$ (attributes, latent factors)

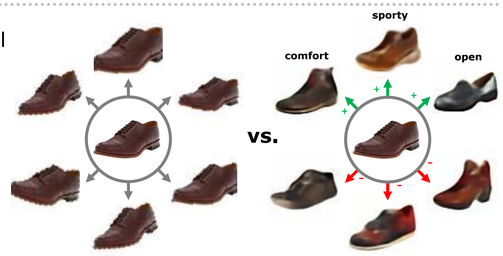
Semantic "jittering" to augment real training images

- high-level changes that modify underlying meaning

Ranker Independent!

RankSVM [Yu & Grauman '14]

DeepSTN [Singh & Lee '16]

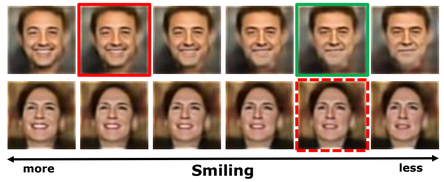


vs.

Hybrid Real + Synthetic Training

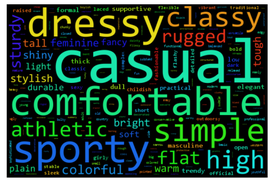
Create *synthetic complement* for real image datasets in each domain (shoes & faces)

- generate synthetic identities
- sample both intra- and inter-pairs from resulting spectrum
- collect pairwise labels from *human annotators* as well



Expand upon our UT-Zap50K dataset [Yu & Grauman '14]

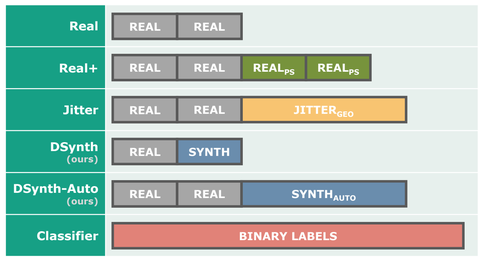
- crowdsource a new *fine-grained attribute lexicon* based on visual subtleties
- collect large new set of pairwise labels, **more than 3 times** that of the original (largest to date)



Experimental Results

* real face images from LFW-10 [Sandeep et al. '14]
* real shoe images from UT-Zap50K [Yu & Grauman '14]

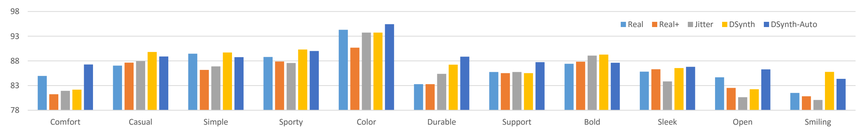
Observation: The synthetic image pairs successfully densify the supervision. Given a novel pair, the nearest neighbors consist of both real and synthetic pairs, suggesting their *combined importance*.



Zap50K-1 (coarse)	Open	Sporty	Comfort
[Parikh & Grauman '11]	88.33	89.33	91.33
[Yu & Grauman '14]	90.67	91.33	93.67
[Singh & Lee '16]	93.00	93.67	94.33
DSynth-Auto (Ours)	**95.00**	**96.33**	**95.00**

Zap50K-2 (fine-grained)	Open	Sporty	Comfort
[Parikh & Grauman '11]	60.36	65.65	62.82
[Yu & Grauman '14]	69.36	66.39	63.84
[Singh & Lee '16]	70.73	67.49	66.09
DSynth-Auto (Ours)	**72.18**	**68.70**	**67.72**

* all methods are trained and tested on 64 x 64 images for fair comparison



Observation: Even with 2x the real data, the state-of-the-art models fail to predict fine-grained differences as well as when our synthetic data are added, demonstrating the importance of *having a dense set of training data*. (Note: All methods use the same amount of human supervision.)

Conclusion

- semantic data augmentation approach to tackle the sparsity of supervision
- data *density* ≠ data *quantity*
- positive evaluation over two domains using two state-of-the-art ranking models demonstrates generalizability, even when using auto labels