Seeing Eye to AI? Applying Deep-Feature-Based Similarity Metrics to Information Visualization

Sheng Long
Northwestern University

Angelos Chatzimparmpas
Utretcht University

Emma Alexander
Northwestern University

Matthew Kay
Northwestern University

Jessica Hullman
Northwestern University

arxiv

INTRODUCTION

- Similarity is a fundamental construct in human cognition.
- Human-perceived similarity has been studied extensively and has been leveraged in applications such as image retrieval and human-in-the-loop categorization.
- Recent studies in computer vision has shown that deep-feature-based similarity metrics correlate well with perceptual judgments of image similarity. For example, see ↓

The success of these applications in computer vision rasies an interesting question:

🤔 Can similar approaches be effectively applied in the domain of information visualization?

RESEARCH CONTRIBUTION

1. We implement a domain-independent transfer-learning technique from computer vision to information visualization.
2. We extend prior work on deep-feature-based similarity metrics using weights trained on Stylized ImageNet, a modified ImageNet-1K dataset where images are artistically stylized while presering their original content and labels.
3. We conceptually replicate two prior experiments:
   1. **Scatterplot experiment:** When using certain deep-learning networks, DF-based similarity metrics achieve ***better clustering alignment*** with human judgments of scatterplot similarity than traditional computer vision metrics whose parameters are gradient-descent-tuned on the same set of scatterplots and human judgments.
   2. **Visual channel experiment:** For visual channels like *color* and *shape*, DF-based metrics struggle to capture what humans perceive as similar. However, they perform well when assessing the visual channel of *size*.

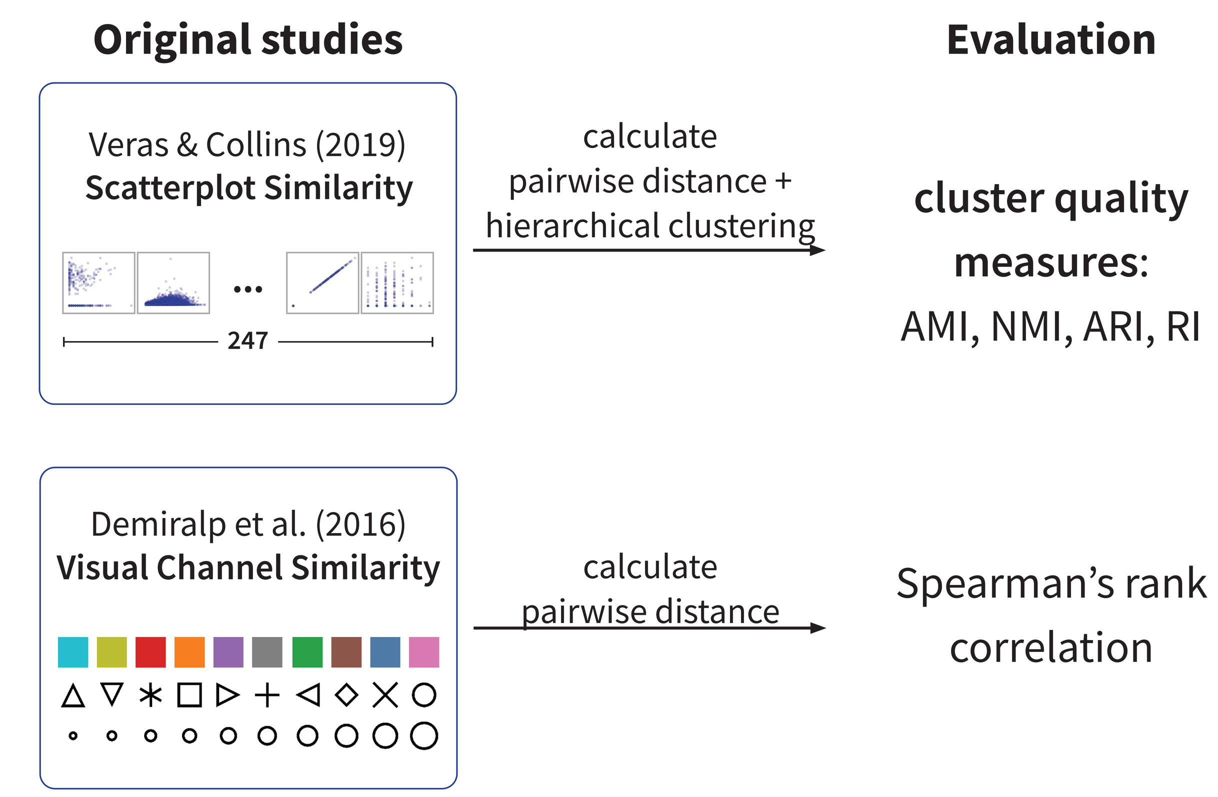
Patches were taken from Zhang et al.

METHODS

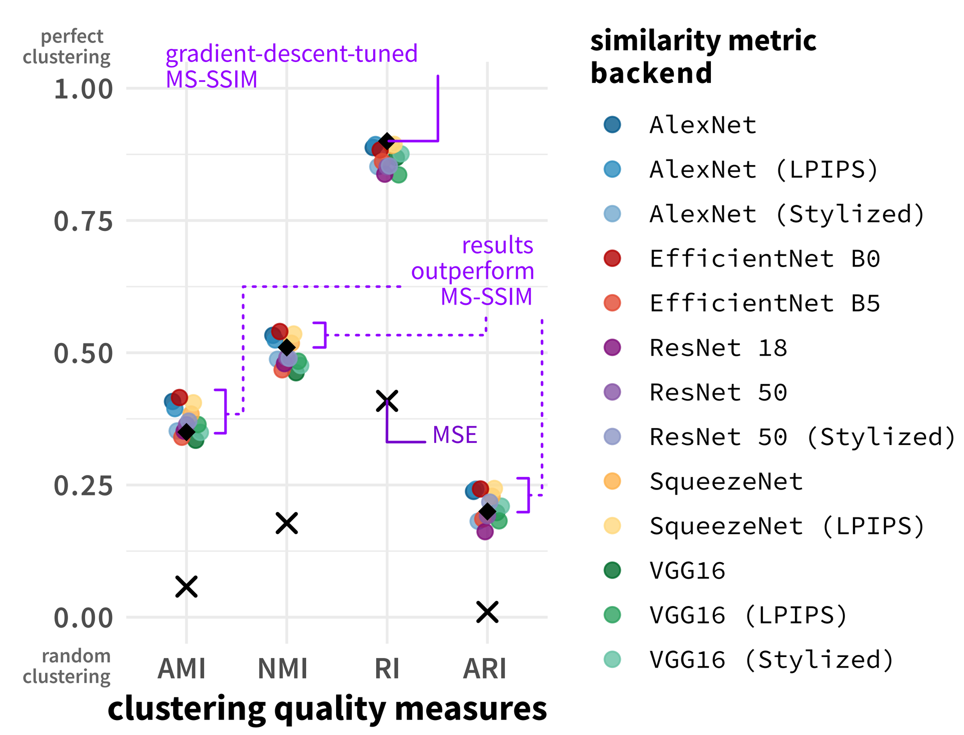
Given two images $x, y \in \mathbb{R}^{H \times W \times C}$ and a network $\mathscr{F}$, the *"perceptual" distance* between x and y is the weighted sum of squared differences between feature activations of x and y across multiple layers and spatial positions. Formally,

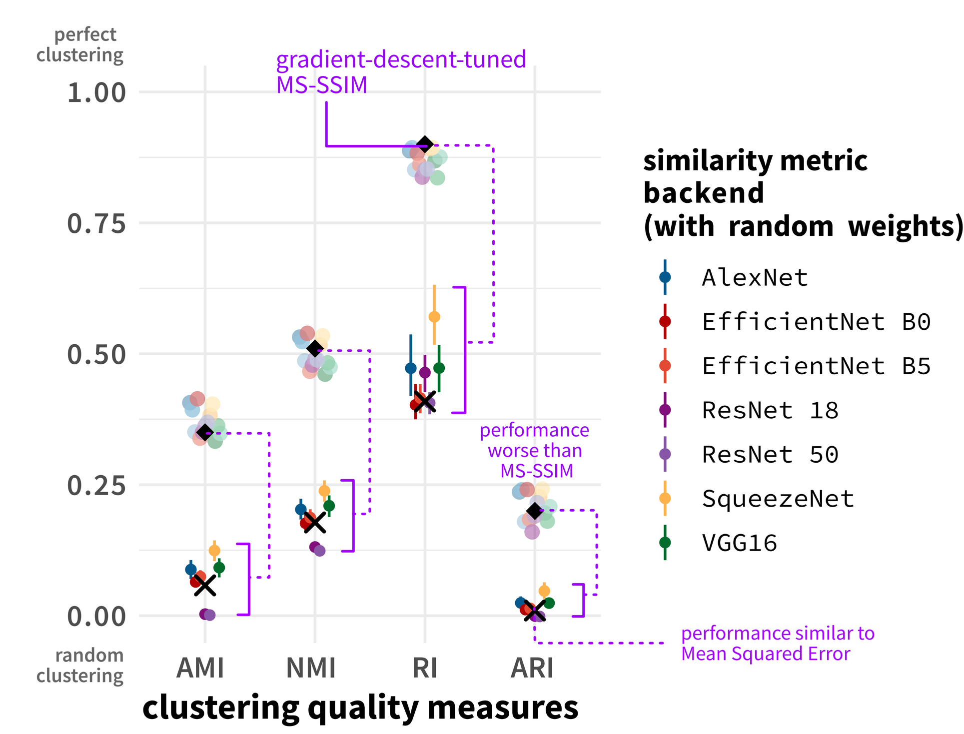
$$d(x,y) = \sum_{l \in \mathscr{L}} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{x}^l_{h,w} - \hat{y}^l_{h,w} \right) \right\|_2^2$$

where $\mathscr{L}$ is the set of feature extraction layers in network $\mathscr{F}$, $\hat{x}^l, \hat{y}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ are the unit-normalized deep feature maps extracted by $\mathscr{F}$ at layer l, and vector $w_l \in \mathbb{R}^{C_l}$ is a channel-wise scaling vector for the difference between unit-normalized feature maps $\hat{x}^l$ and $\hat{y}^l$ at spatial location (h, w).



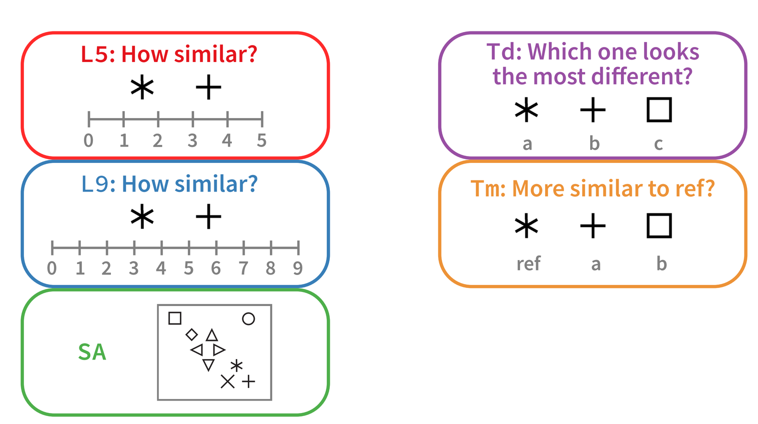
RESULTS: SCATTERPLOT SIMILARITY (VERAS & COLLINS 2019)



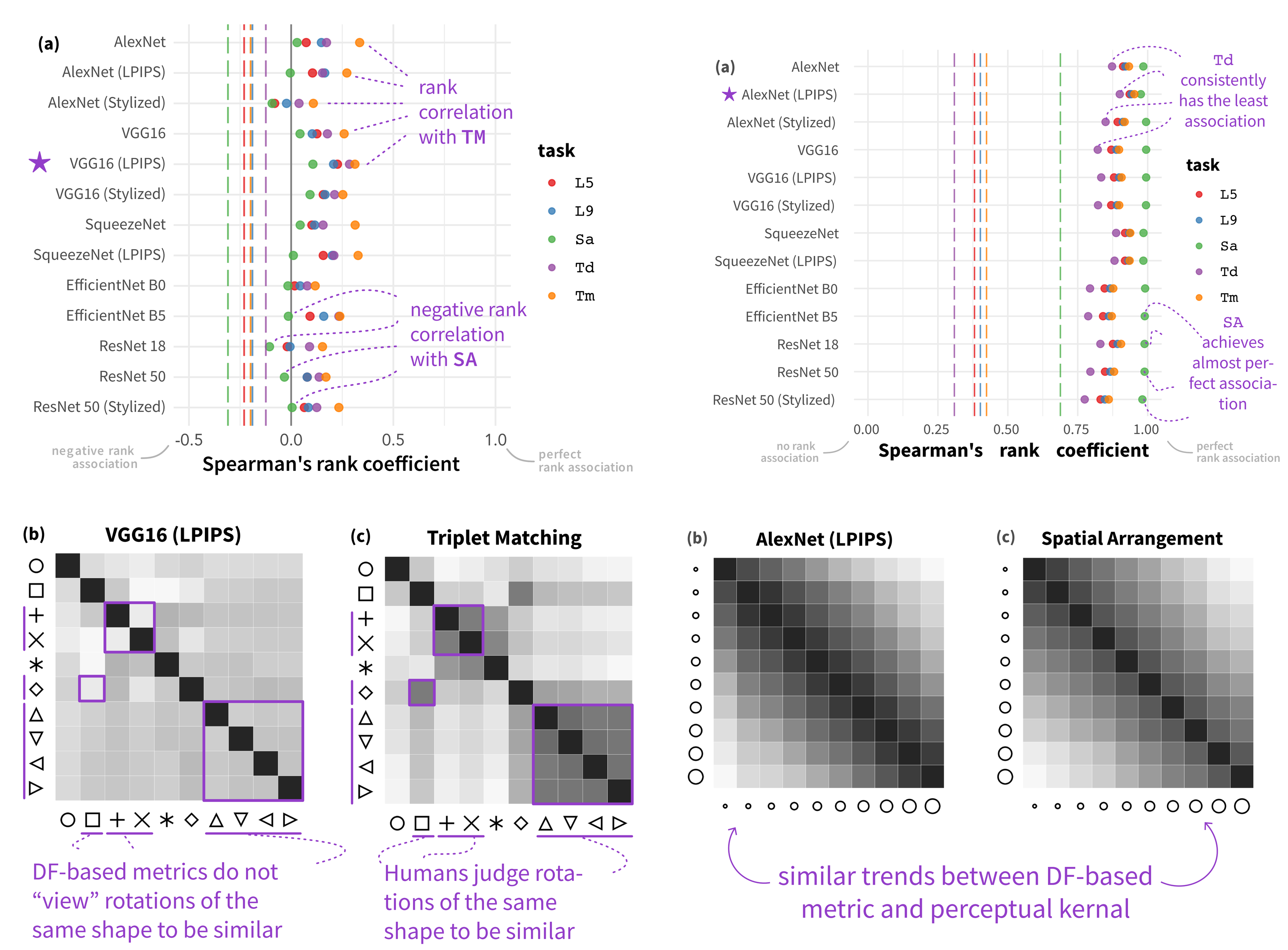


RESULTS: VISUAL CHANNEL SIMILARITY (DEMIRALP ET AL. 2016)

There are different ways to elicit similarity judgments:



For simple visual stimuli without complex patterns or textures, the feature map difference primarily reflects how well the stimuli spatially align/ structurally correspond 👉



DISCUSSIONS & FUTURE WORK

- Performance is consistent across neural network architectures but varies with pre-trained weight complexity.
- When judging multi-channel visual stimuli, participants prioritize color before size similarity. Such perceptual hierarchies are not captured by DF-based similarity metrics.
- Different similarity judgment tasks impose varying cognitive constraints, which are then encoded in different outcome variables and turned into different inferred representations. We need to think more about ways to model similarity.
- Nevertheless, DF-based metrics show promise for pre-screening visualization designs before costly human studies.

CONCLUSIONS

- We explore DF-based similarity metrics for information visualization through replicating two well-established prior studies.
- Deep features trained on diverse, large-scale natural images (e.g., ImageNet-1k) transfer remarkably well to visualizations like scatterplots, where ***spatial distribution is key.***
- Limitations emerge when applying deep features to abstract visual primitives (glyph shapes, colors), likely because such judgments extend beyond purely perceptual processes.

ACKNOWLEDGMENTS

We thank Noah Shen for initiating the preliminary work and members of the Mu Collective for their continued support. We thank Enrico Bertini, Rafael Veras, and Cagaty Demiralp for making their data and analysis publicly available. This work was partially supported by NSF IIS-1930642.

REFERENCES

- Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- Veras, Rafael, and Christopher Collins. "Discriminability tests for visualization effectiveness and scalability." IEEE transactions on visualization and computer graphics 26.1 (2019): 749-758.
- Demiralp, Çağatay, Michael S. Bernstein, and Jeffrey Heer. "Learning perceptual kernels for visualization design." IEEE transactions on visualization and computer graphics 20.12 (2014): 1933-1942.