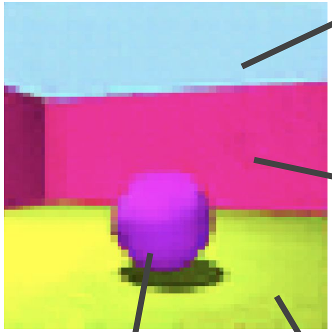
Weakly Supervised Disentanglement with Guarantees

Rui Shu

Joint work with Yining Chen, Abhishek Kumar, Stefano Ermon, Ben Poole

Why

Decompose data into a set of underlying **human-interpretable** factors of variation



Explainable models

What is in the scene?

Controllable generation

Generate a red ball instead

How: Fully-Supervised

Strategy: Label everything

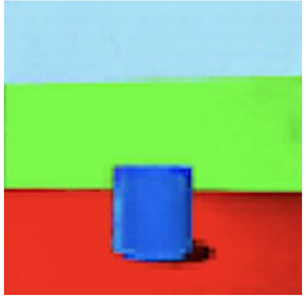
{dark blue wall, green floor, green oval}

{green wall, red floor, green cylinder}

{red wall, green floor, pink ball}

Controllable generation as **label-conditional generative modeling**

green wall, red floor, blue cylinder



How: Fully-Supervised

Problem: Some things are hard to label

What kind of hairstyle?

What kind of glasses?

Generate this guy with this hair

How: Unsupervised?

Strategy: Exploit statistical independence assumption + neural net magic

Swivel the chair

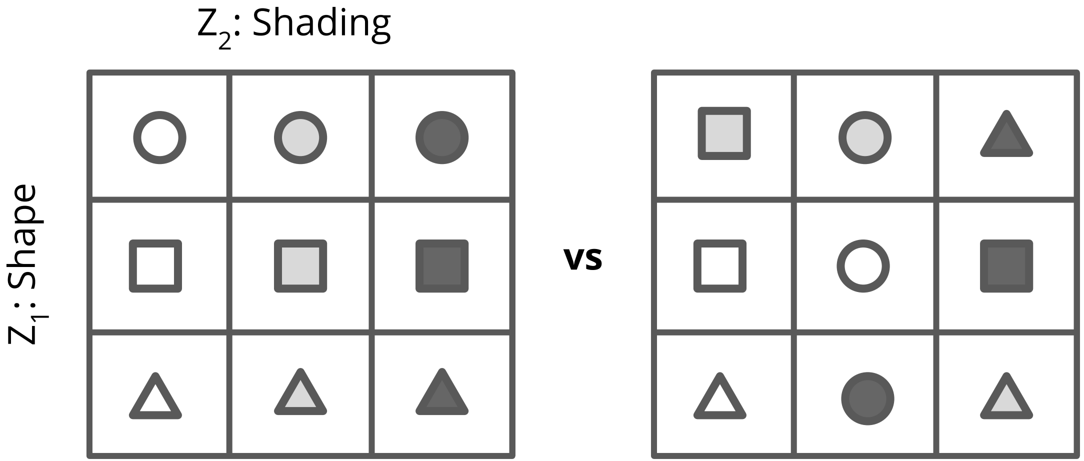
Beta-VAE

TC-VAE

FactorVAE

How: Unsupervised?

Problem: Is statistical independence assumption + neural net magic enough?



> mean) are correlated. (ii) We do not find any evidence that the considered models can be used to reliably learn disentangled representations in an *unsupervised* manner as random seeds and hyperparameters seem to matter more than the model choice. Furthermore, good trained models seemingly cannot be identified without access to ground-truth labels even if we are allowed to transfer good hyperparameter values across data sets. (iii) For

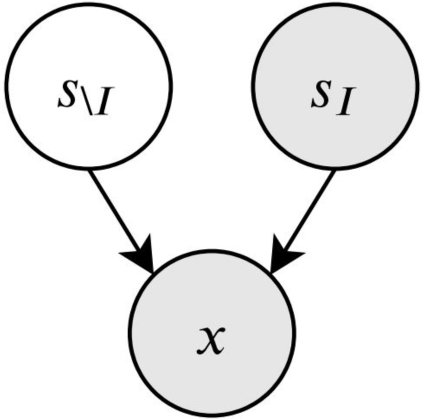
Locatello, et al. *Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations*, ICML 2019.

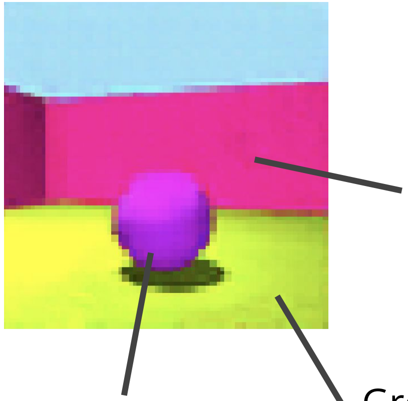
How: Weakly Supervised

Strategy: Leverage “weak” supervision when possible

How: Weakly Supervised

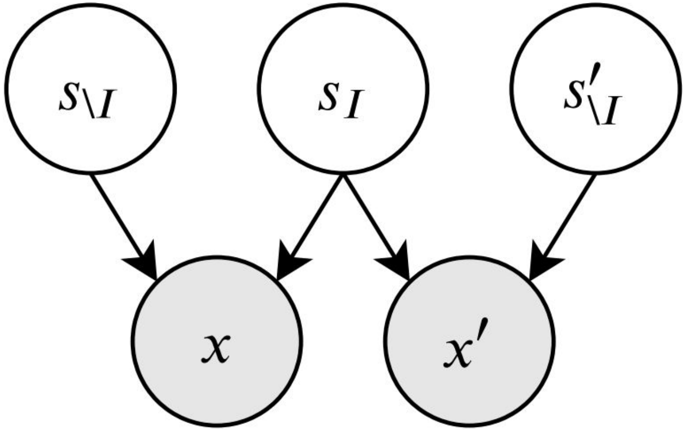
Restricted Labeling: Label what we can

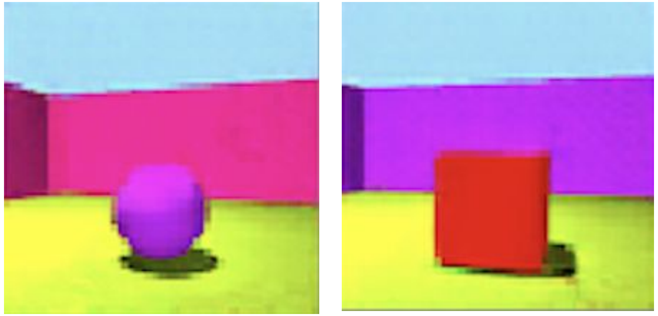




How: Weakly Supervised

Match Pairing: Find pairs with known similarities



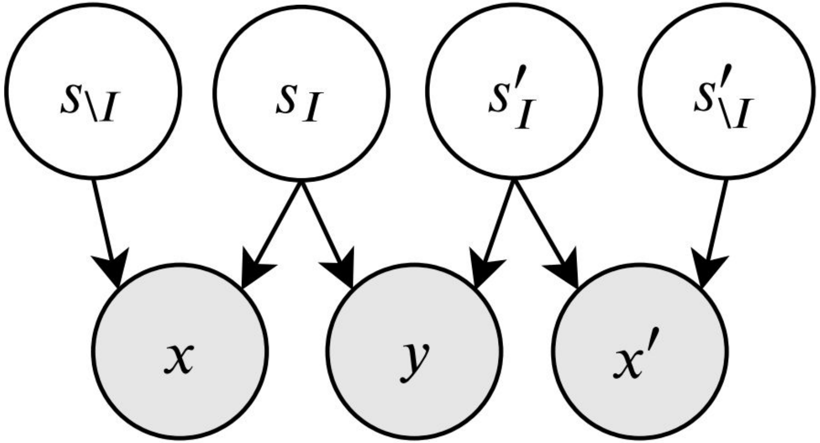


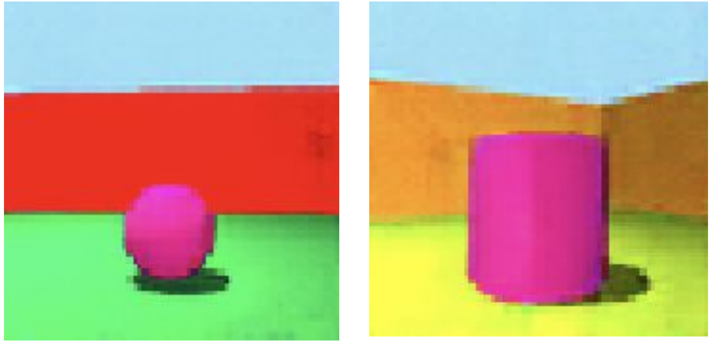
Same ground color

Real world data: direct intervention to share / change certain factors

How: Weakly Supervised

Rank Pairing: Compare pairs

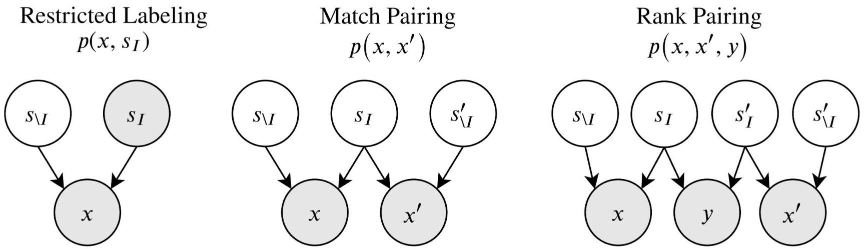




Which is bigger?

The Plan

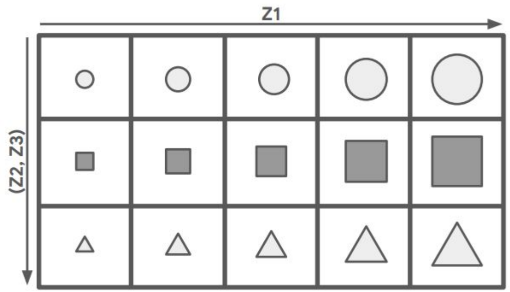
1. **Definitions**: Decompose disentanglement into:
   a. Consistency
   b. Restrictiveness
2. **Guarantees:** Prove whether weak supervision guarantees consistency, restrictiveness, or both



Departure from existing literature: no end-to-end theoretical framework of disentanglement

Definitions

Disentangle: What does it mean when I say Z1 disentangles *size?*

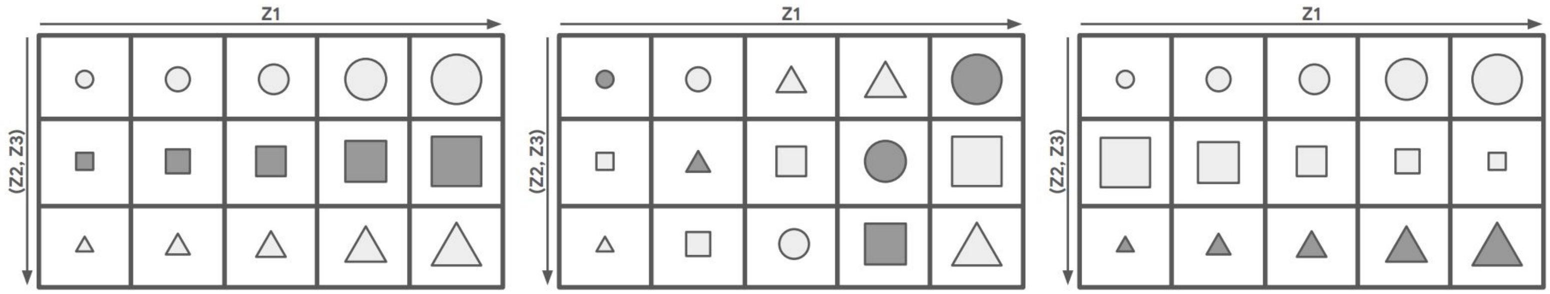


(a) Disentanglement

1. When z_1 is fixed, is size fixed?
2. When we only change z_1, does only size change?

Definitions

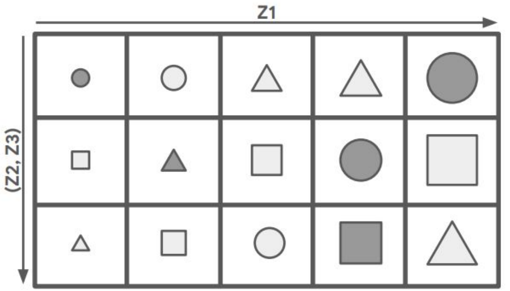
Disentangle: What does it mean when I say Z1 disentangles *size?*



(a) Disentanglement (b) Consistency (c) Restrictiveness

1. When z_1 is fixed, is size fixed? (**Consistency**)
2. When we only change z_1, does only size change? (**Restrictiveness**)

Definitions: Consistency



(b) Consistency

When Z_I is fixed, S_I is fixed

Oracle encoder

Generative model

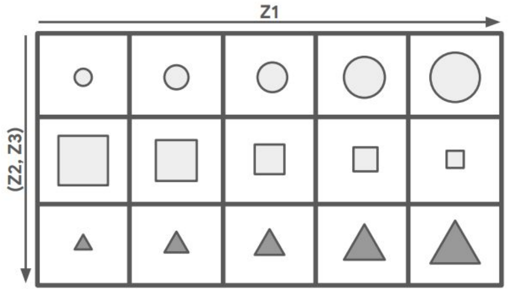
$$\mathbb{E}_{p_I} \| e_I^* \circ g(z_I, z_{\setminus I}) - e_I^* \circ g(z_I, z'_{\setminus I}) \|^2 = 0$$

$$z_I \sim p(z_I)$$

$$z_{\setminus I}, z'_{\setminus I} \overset{\text{iid}}{\sim} p(z_{\setminus I} \mid z_I).$$

Perturbation-based generation

Definitions: Restrictiveness



(c) Restrictiveness

When only Z_I is changed, only S_I is changed

Equivalently: when $Z_{\setminus I}$ is fixed, $S_{\setminus I}$ is fixed

Oracle encoder

Generative model

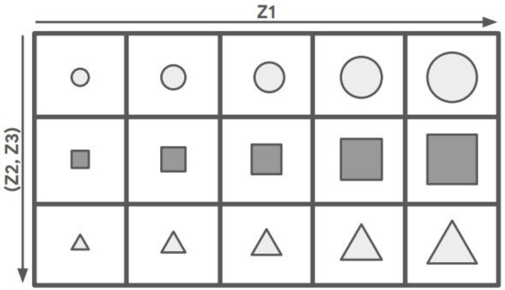
$$\mathbb{E}_{p_{\setminus I}} \| e^*_{\setminus I} \circ g(z_I, z_{\setminus I}) - e^*_{\setminus I} \circ g(z'_I, z_{\setminus I}) \|^2 = 0$$

$$z_{\setminus I} \sim p(z_{\setminus I})$$

$$z_I, z'_I \overset{\text{iid}}{\sim} p(z_I \mid z_{\setminus I}).$$

Perturbation-based generation

Definitions: Disentanglement

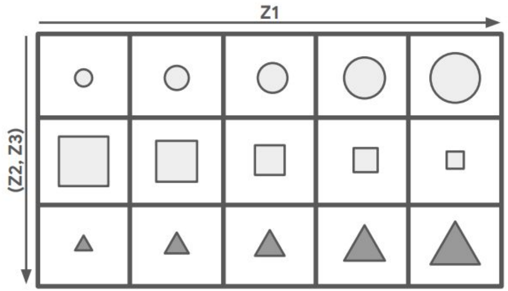


(a) Disentanglement

$$D(I) := C(I) \wedge R(I)$$

Z_I is consistent **and** restricted to S_I

Consistency versus Restrictiveness



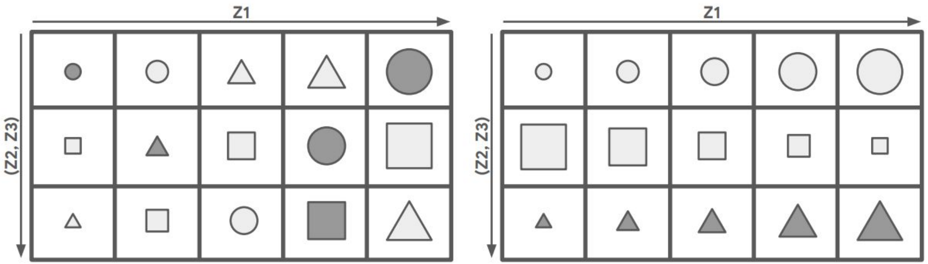
(c) Restrictiveness

When only Z_I is changed, only S_I is changed

Equivalently: when $Z_{\setminus I}$ is fixed, $S_{\setminus I}$ is fixed

$$C(I) \iff R(\setminus I)$$

Consistency versus Restrictiveness



(b) Consistency (c) Restrictiveness

$$R(I) \not\Rightarrow C(I)$$

$$C(I) \not\Rightarrow R(I)$$

Union Rules

Consistency Union:

If fixing Z_I fixes S_I
and fixing Z_J fixes S_J
then fixing (Z_I, Z_J) fixes (S_I, S_J)

$$C(I) \wedge C(J) \implies C(I \cup J)$$

Restrictiveness Union:

If changing Z_I changes only S_I
and changing Z_J changes only S_J
then changing (Z_I, Z_J) changes only (S_I, S_J)

$$R(I) \wedge R(J) \implies R(I \cup J)$$

Intersection Rules

Consistency Intersection:

If fixing Z_I *fixes* S_I
and fixing Z_J *fixes* S_J
then fixing Z_V *fixes* S_V

$$C(I) \wedge C(J) \implies C(I \cap J)$$

Restrictiveness Intersection:

If changing Z_I *changes only* S_I
and changing Z_J *changes only* S_J
then changing Z_V *changes only* S_V

$$R(I) \wedge R(J) \implies R(I \cap J)$$

Disentanglement Rule

Disentanglement via Consistency

Consistency on all factors implies disentanglement on all factors

$$\bigwedge_{i=1}^{n} C(i) \iff \bigwedge_{i=1}^{n} D(i)$$

Disentanglement via Restrictiveness

Restrictiveness on all factors implies disentanglement on all factors

$$\bigwedge_{i=1}^{n} R(i) \iff \bigwedge_{i=1}^{n} D(i)$$

Summary of Rules

Consistency and Restrictiveness

$$C(I) \not\Rightarrow R(I) \qquad R(I) \not\Rightarrow C(I) \qquad C(I) \Longleftrightarrow R(\backslash I)$$

Union Rules

$$C(I) \wedge C(J) \implies C(I \cup J) \qquad R(I) \wedge R(J) \implies R(I \cup J)$$

Intersection Rules

$$C(I) \wedge C(J) \implies C(I \cap J) \qquad R(I) \wedge R(J) \implies R(I \cap J)$$

Full Disentanglement

$$\bigwedge_{i=1}^{n} C(i) \Longleftrightarrow \bigwedge_{i=1}^{n} D(i) \qquad \bigwedge_{i=1}^{n} R(i) \Longleftrightarrow \bigwedge_{i=1}^{n} D(i)$$

Summary of Rules

Consistency and Restrictiveness

$$C(I) \;\not\Longrightarrow\; R(I) \qquad\qquad R(I) \;\not\Longrightarrow\; C(I) \qquad\qquad C(I) \Longleftrightarrow R(\backslash I)$$

Union Rules

$$C(I) \wedge C(J) \implies C(I \cup J) \qquad\qquad R(I) \wedge R(J) \implies R(I \cup J)$$

Intersection Rules

$$C(I) \wedge C(J) \implies C(I \cap J) \qquad\qquad R(I) \wedge R(J) \implies R(I \cap J)$$

Full Disentanglement

$$\textstyle\bigwedge_{i=1}^{n} C(i) \Longleftrightarrow \bigwedge_{i=1}^{n} D(i) \qquad\qquad \bigwedge_{i=1}^{n} R(i) \Longleftrightarrow \bigwedge_{i=1}^{n} D(i)$$

Strategy for Disentanglement

Dataset 1 → C(1)

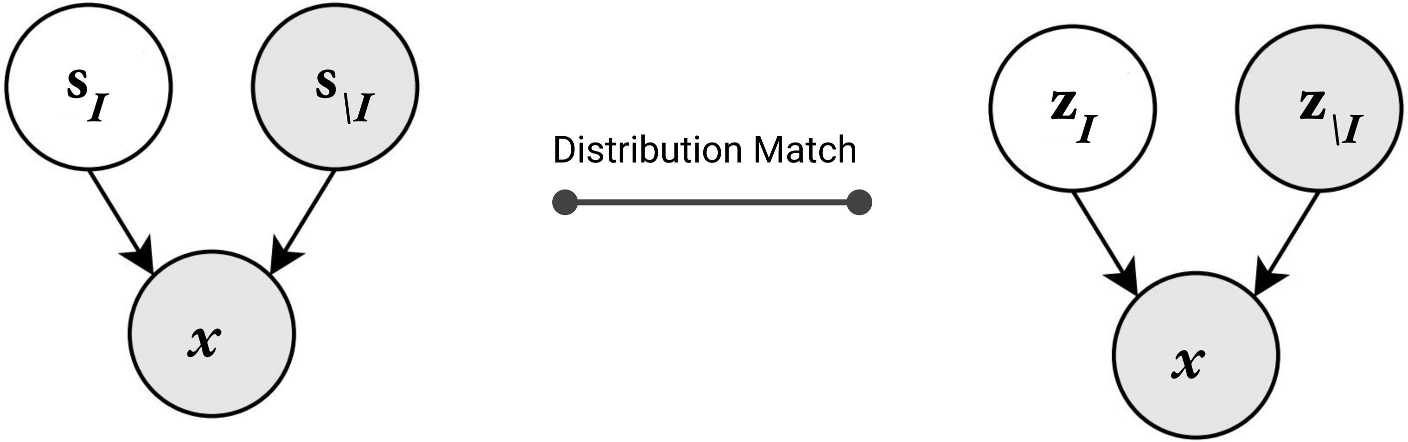
Dataset 2 → C(2)

…

Dataset n → C(n)

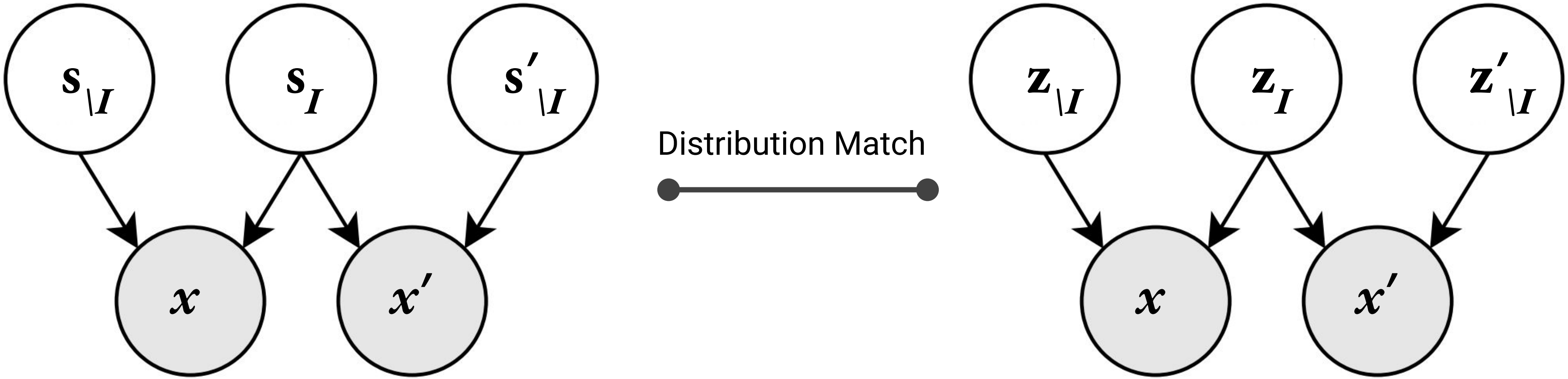
Using datasets together (+ right algorithm) guarantees full disentanglement

Restricted Labeling Guarantees Consistency



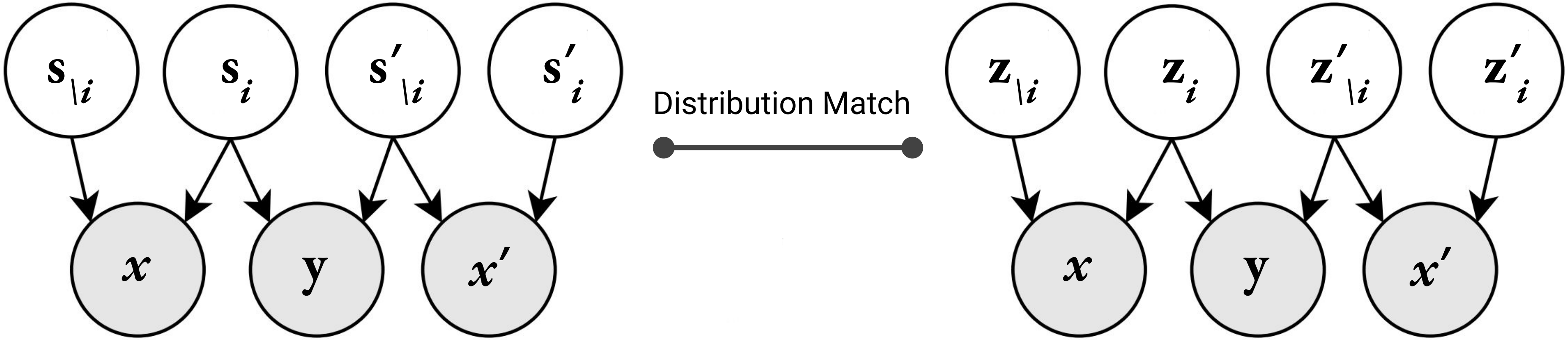
Z_I will be consistent with S_I

Match Pairing Guarantees Consistency



Z_I will be consistent with S_I

Rank Pairing Guarantees Consistency



Z_I will be consistent with S_I

Summary of Guarantees

Theorem 1. *Given any oracle* $(p^*(s), g^*, e^*) \in \mathcal{H}$*, consider the distribution-matching algorithm* $\mathcal{A}$ *that selects a model* $(p(z), g, e) \in \mathcal{H}$ *such that:*

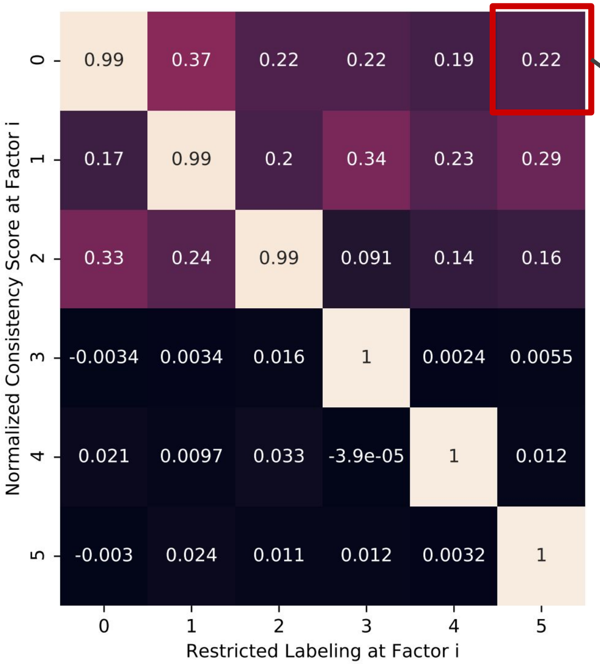
1. $(g^*(S), S_I) \stackrel{d}{=} (g(Z), Z_I)$ ***(Restricted Labeling)****; or*

2. $\left(g^*(S_I, S_{\setminus I}), g^*(S_I, S'_{\setminus I})\right) \stackrel{d}{=} \left(g(Z_I, Z_{\setminus I}), g(Z_I, Z'_{\setminus I})\right)$ ***(Match Pairing)****; or*

3. $(g^*(S), g^*(S'), \mathbf{1}\{S_I \leq S'_I\}) \stackrel{d}{=} (g(Z), g(Z'), \mathbf{1}\{Z_I \leq Z'_I\})$ ***(Rank Pairing)****.*

Then the latent variable Z_I *from the learned generative model* $(p(z), g)$ *will be consistent with the factor of variation* S_I*.*

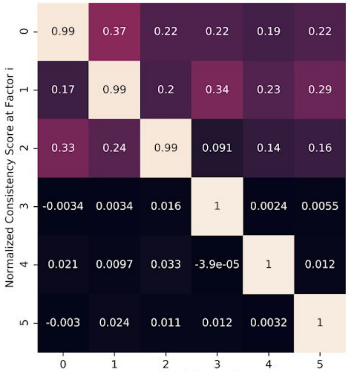
Targeted Consistency / Restrictiveness



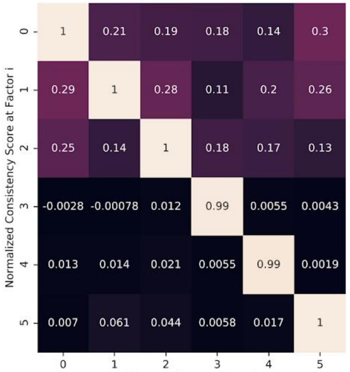
Generative model trained via restricted labeling at S_5

Evaluated model on consistency of Z_0 vs S_0

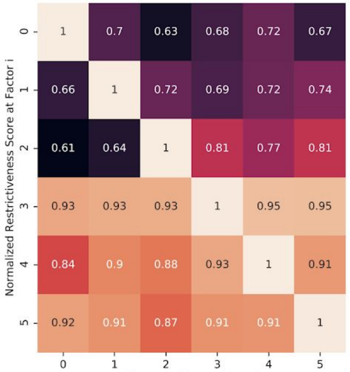
Targeted Consistency / Restrictiveness



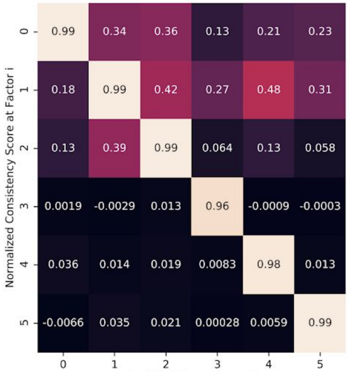
Consistency:
Restricted Labeling



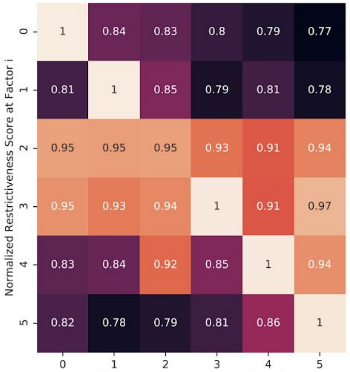
Consistency:
Match Pairing
(Share 1 factor)



Restrictiveness:
Match Pairing
(Change 1 factor)

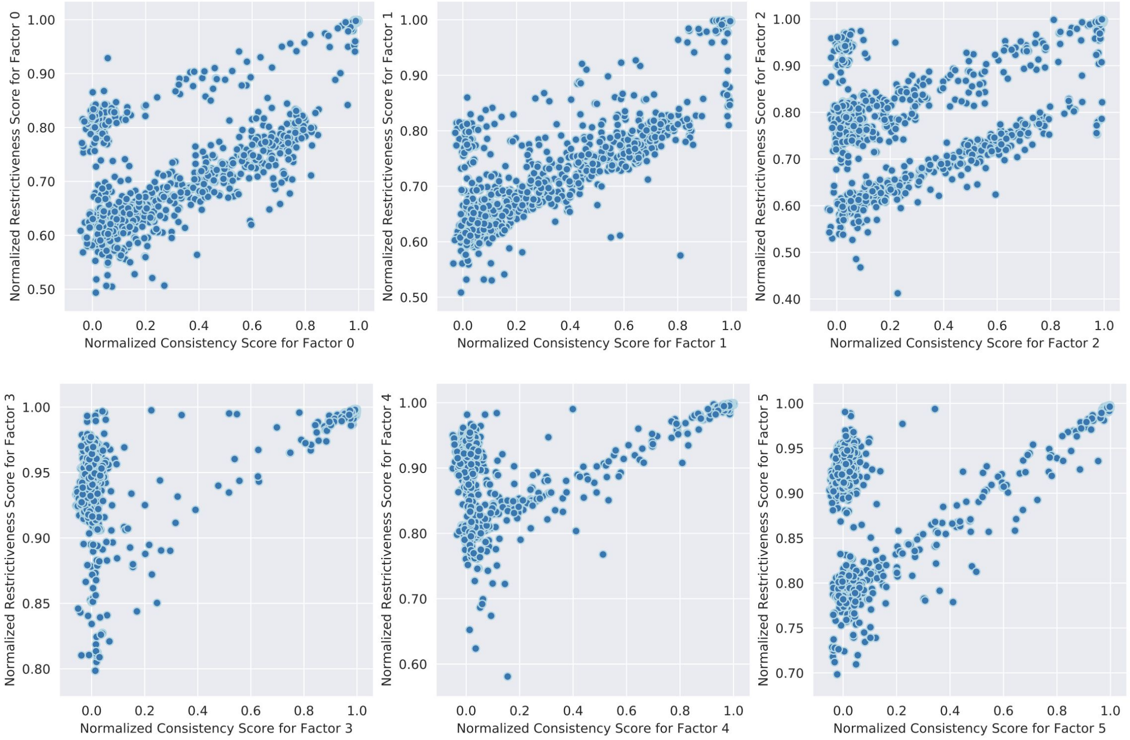


Consistency:
Rank pairing



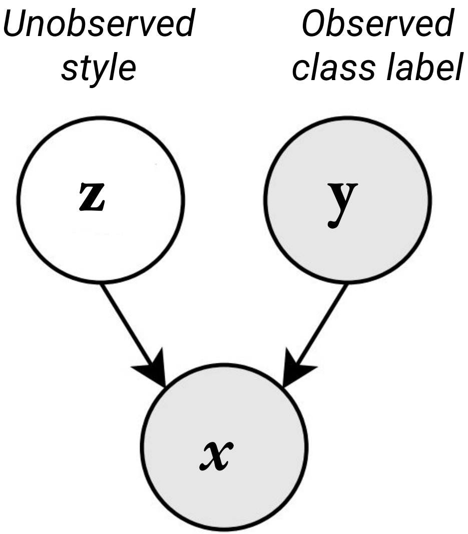
Restrictiveness:
Intersection

Consistency versus Restrictiveness



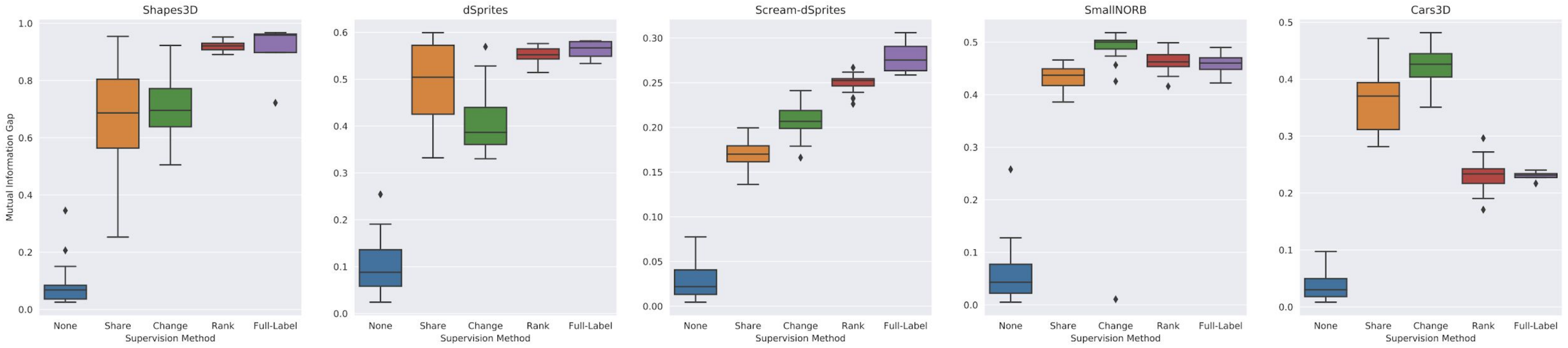
- *Models trained to guarantee only consistency or restrictiveness of one factor*
- *Strong correlation of consistency vs restrictiveness*

Digression: Style-Content Disentanglement

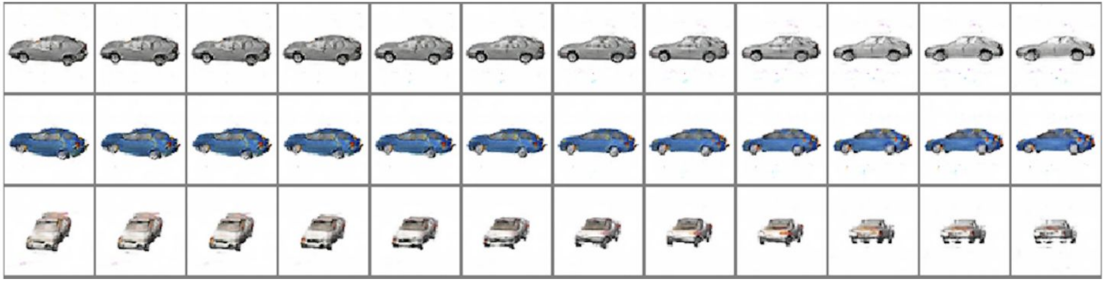


Only content-consistency is guaranteed
Style-content disentanglement not guaranteed (but due to neural net magic)

Full Disentanglement



Full Disentanglement: Visualizations

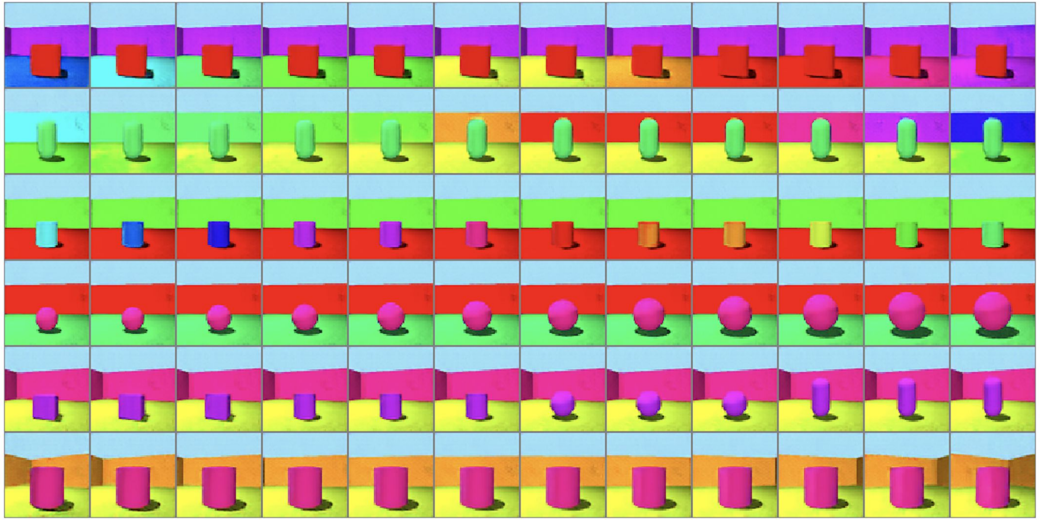


Elevation

Azimuth

- *Visualize multiple rows of single-factor ablation*
- *Check for consistency **and** restrictiveness*

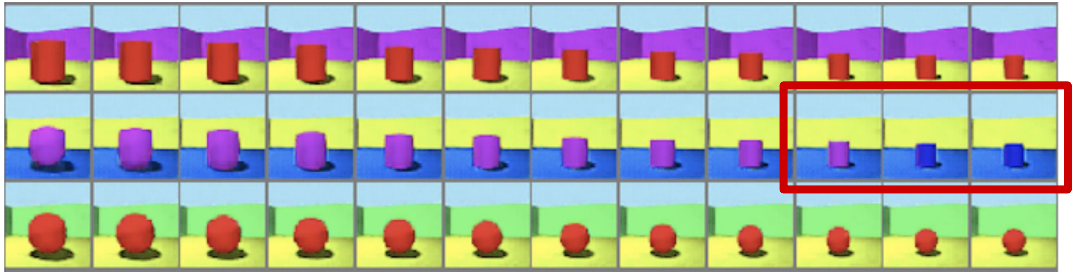
Full Disentanglement: Visualizations



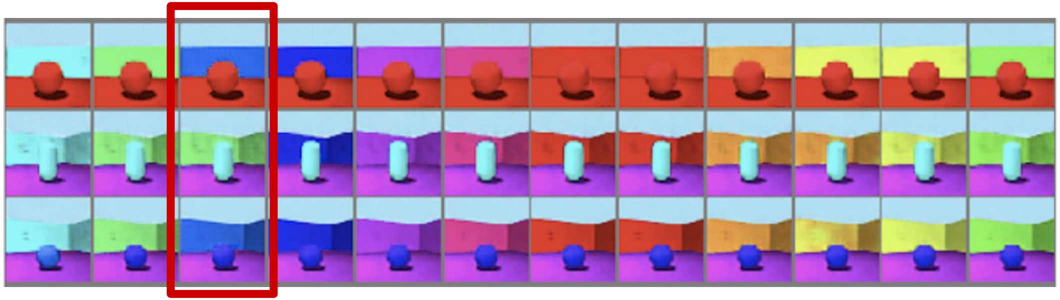
Ground truth factors: floor color, wall color, object color, object size, object type, and azimuth.

- *Visualize multiple rows of single-factor ablation*
- *Check for consistency **and** restrictiveness*

Full Disentanglement: Visualizations



Ground truth factor: object size



Ground truth factor: wall color

- *Visualize multiple rows of single-factor ablation*
- *Check for consistency **and** restrictiveness*

Conclusions

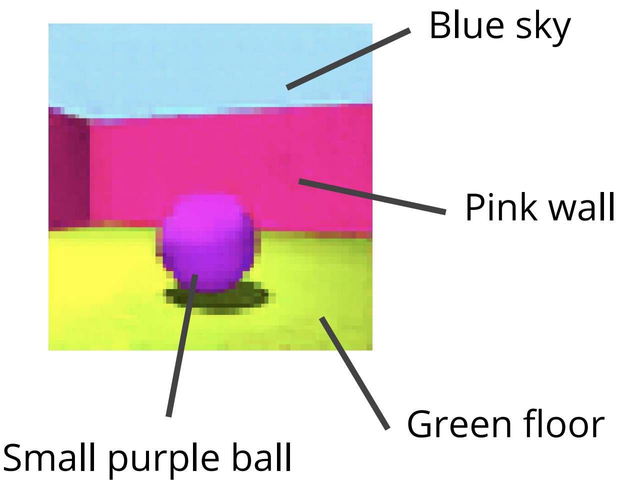
- Definitions for disentanglement
- A calculus of disentanglement
- Analyzed weak supervision methods
- Demonstrated guarantees empirically

Conclusions

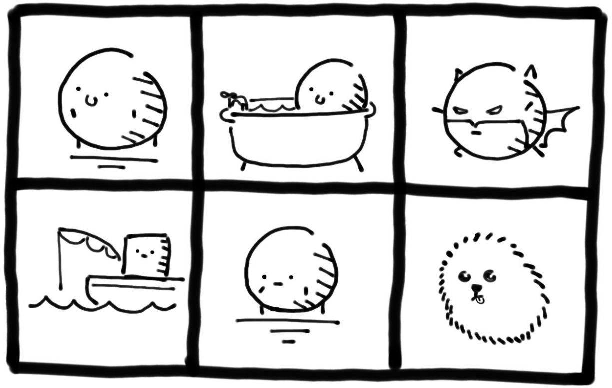
- Definitions for disentanglement
- A calculus of disentanglement
- Analyzed weak supervision methods
- Demonstrated guarantees empirically

- *Better definitions?*
- *Do new definitions preserve calculus?*
- *Analyze other weak supervision methods?*
- *Cost of weak supervision in real world?*

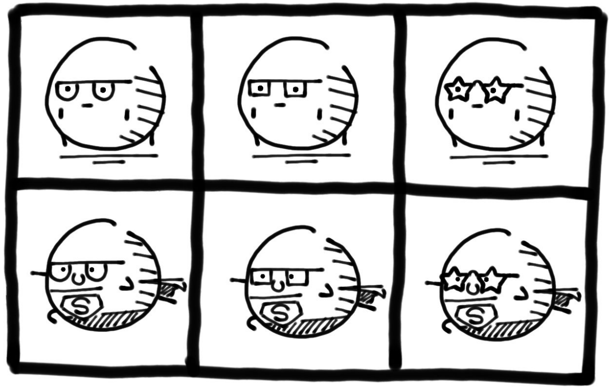
Assumption: $X \rightarrow S$ is deterministic



Questions?



Entangled



Disentangled

ruishu@stanford.edu
@_smileyball
@smiley._.ball