Unsupervised Layered Image Decomposition into Object Prototypes

Tom Monnier[1] Elliot Vincent[1,2] Jean Ponce[2,3] Mathieu Aubry[1]
[1] LIGM, Ecole des ponts, Univ Gustave Eiffel, CNRS [2] Inria and DIENS (ENS-PSL, CNRS, Inria) [3] Center for Data Science, NYU

http://imagine.enpc.fr/~monniert/DTI-Sprites

Pytorch code

2021 ICCV OCTOBER 11-17 VIRTUAL

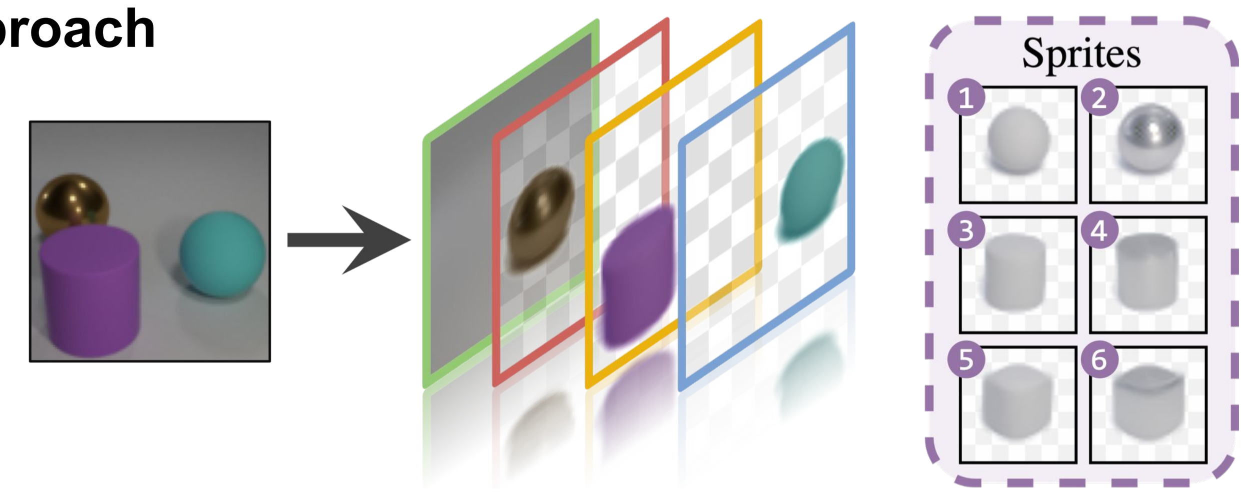
Motivation

Goal → decompose images into object layers without supervision

Previous works → no object types & mostly synthetic images

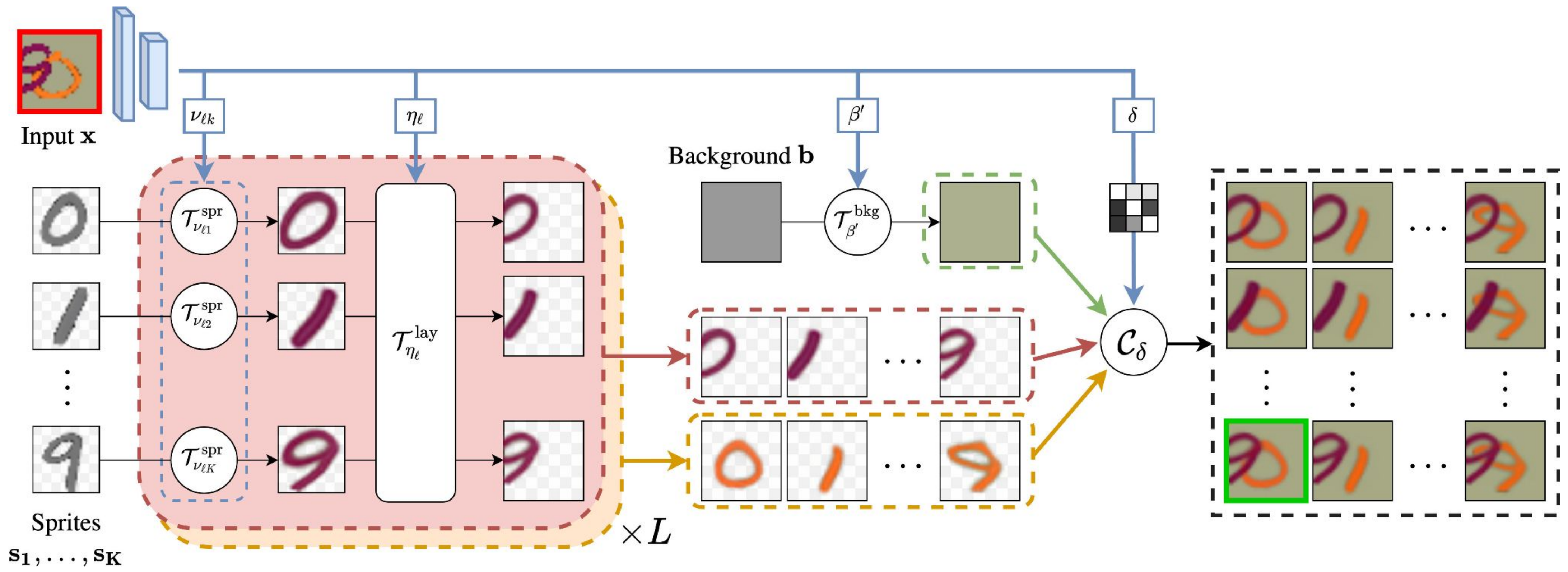
1. Spatial mixture models: MONet [1], IODINE [2], Slot Att. [3], etc.
2. Spatial attention based: AIR [4], SQAIR [5], SPACE [6], etc.

Our approach



1. New image formation model defined as a layered composition of transformed 2D prototypes
2. New unsupervised learning framework for image decomposition
3. Strong results on multi-object benchmarks in both *instance* and *semantic* segmentation + application to real images

Method



Unsupervised learning loss

$$\mathcal{L}(\mathbf{s}_{1:K}, \phi_{1:L}, \psi) = \sum_{i=1}^{N} \min_{k_1,\dots,k_L} \Big[\lambda \sum_{j=1}^{L} \mathbb{1}_{[k_j \neq 0]} + \Big\| \mathbf{x}_i - \mathcal{C}_{\psi(\mathbf{x}_i)}\Big(\mathcal{T}_{\phi_1(\mathbf{x}_i)}(\mathbf{s}_{k_1}), \dots, \mathcal{T}_{\phi_L(\mathbf{x}_i)}(\mathbf{s}_{k_L})\Big)\Big\|_2^2 \Big]$$

Image formation model

L object layers $\{\mathbf{o}_1, \dots, \mathbf{o}_L\}$ where $\mathbf{o}_\ell = (\mathbf{o}_\ell^\alpha, \mathbf{o}_\ell^c)$

Layers are composed following a predicted occlusion matrix δ:

$$\mathcal{C}_\delta(\mathbf{o}_1, \dots, \mathbf{o}_L) = \sum_{\ell=1}^{L} \Big(\prod_{j=1}^{L} (\mathbf{1} - \delta_{j\ell} \mathbf{o}_j^\alpha) \Big) \odot \mathbf{o}_\ell^\alpha \odot \mathbf{o}_\ell^c$$

K learnable sprites $\{\mathbf{s}_1, \dots, \mathbf{s}_K\} \cup \{\mathbf{s}_0 = \mathbf{0}\}$

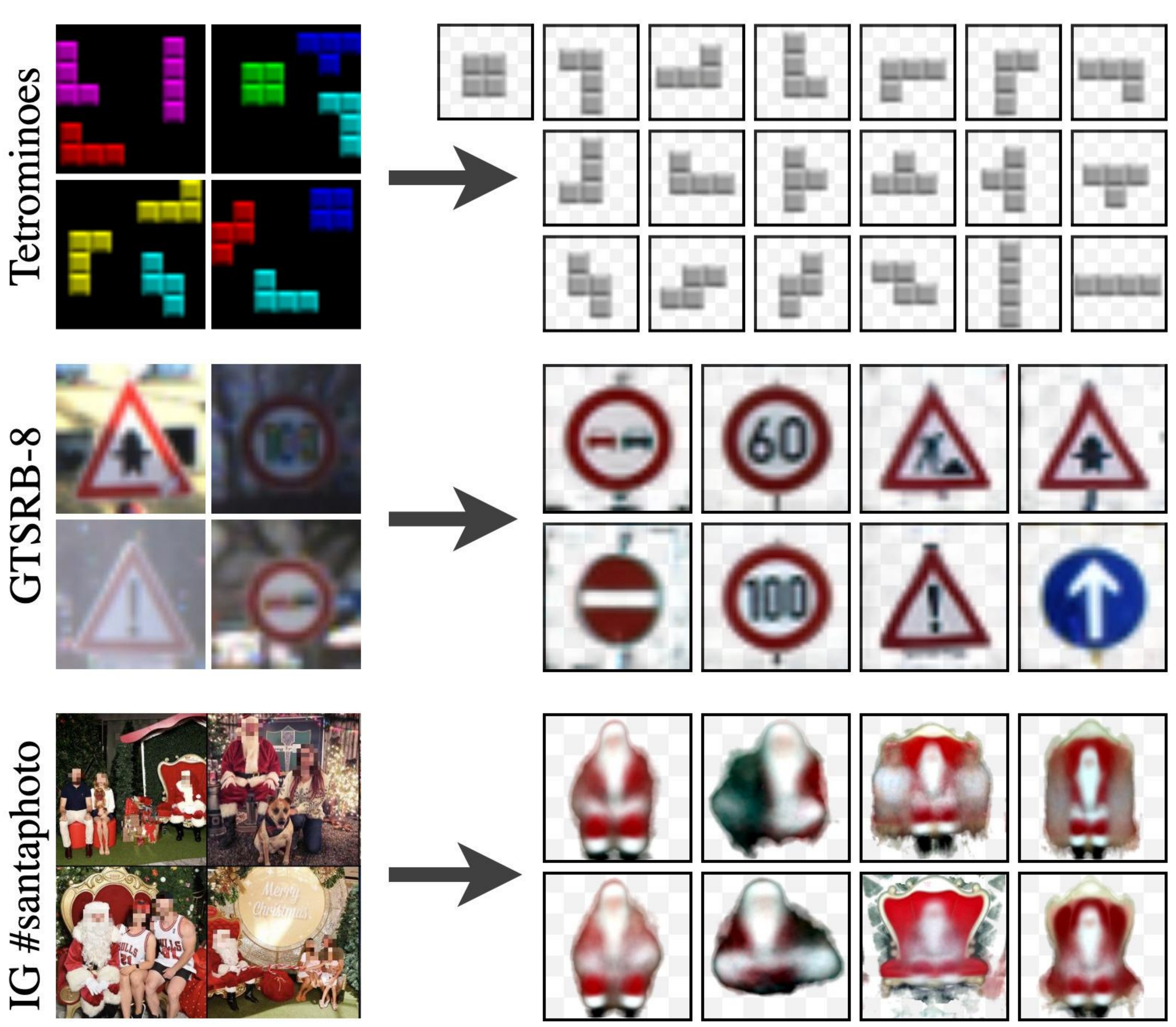
Sprite-specific and/or shared parametric transformations $\mathcal{T}_\beta$

Each layer is modeled as a transformation of one of the K sprites: $\mathbf{o}_\ell = \mathcal{T}_\beta(\mathbf{s}_{k_\ell})$

$$\Rightarrow \mathbf{x} = \mathcal{C}_\delta\Big(\mathcal{T}_{\beta_1}(\mathbf{s}_{k_1}), \dots, \mathcal{T}_{\beta_L}(\mathbf{s}_{k_L})\Big)$$

Results

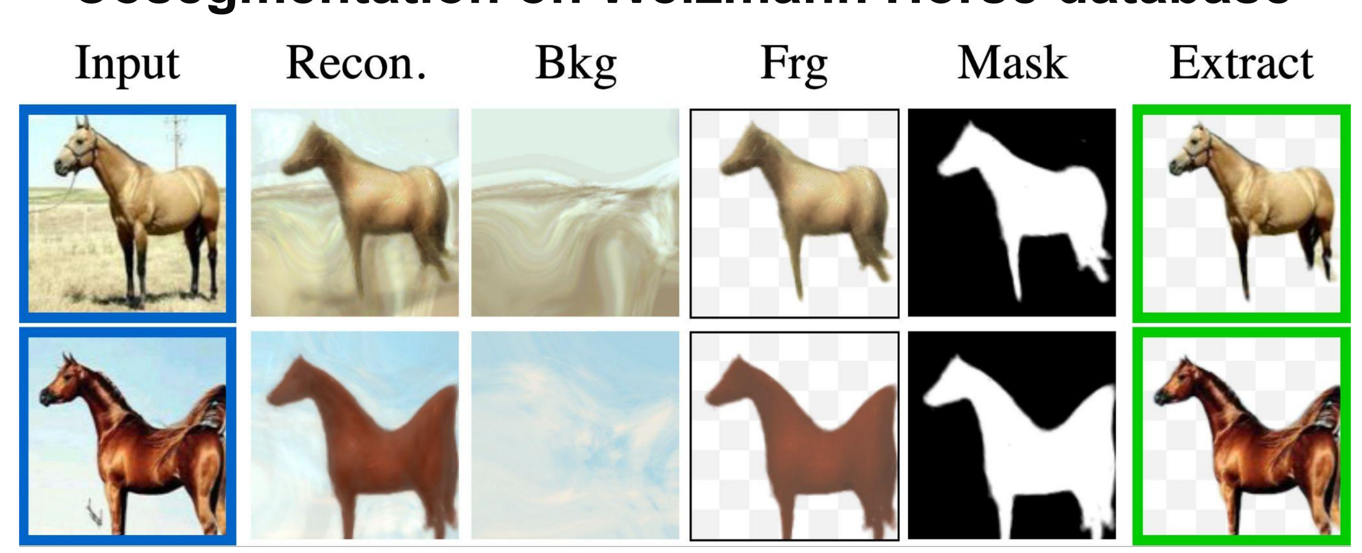
Examples of automatically discovered sprites



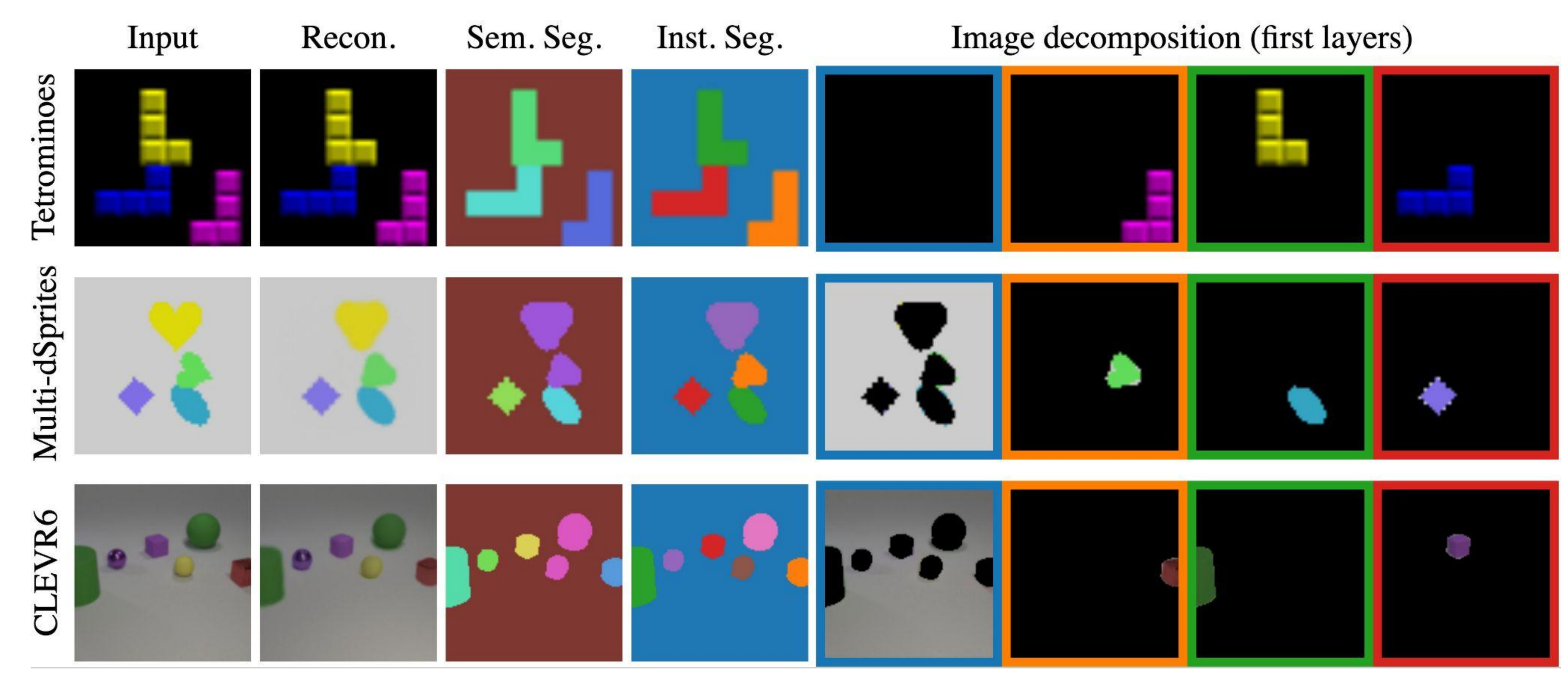
Multi-object discovery comparisons

	Metric	Tetrominoes	Multi-dSprites	CLEVR6
MONet [1]	ARI-FG	-	90.4 ± 0.8	96.2 ± 0.6
IODINE [2]	ARI-FG	99.2 ± 0.4	76.7 ± 5.6	$\mathbf{98.8 \pm 0.0}$
Slot Att [3]	ARI-FG	$\underline{99.5 \pm 0.2}$	$\underline{91.3 \pm 0.3}$	$\mathbf{98.8 \pm 0.3}$
Ours	ARI-FG	$\mathbf{99.6 \pm 0.2}$	$\mathbf{92.5 \pm 0.3}$	$\underline{97.2 \pm 0.2}$
Ours	ARI	99.8 ± 0.1	95.1 ± 0.1	90.7 ± 0.1

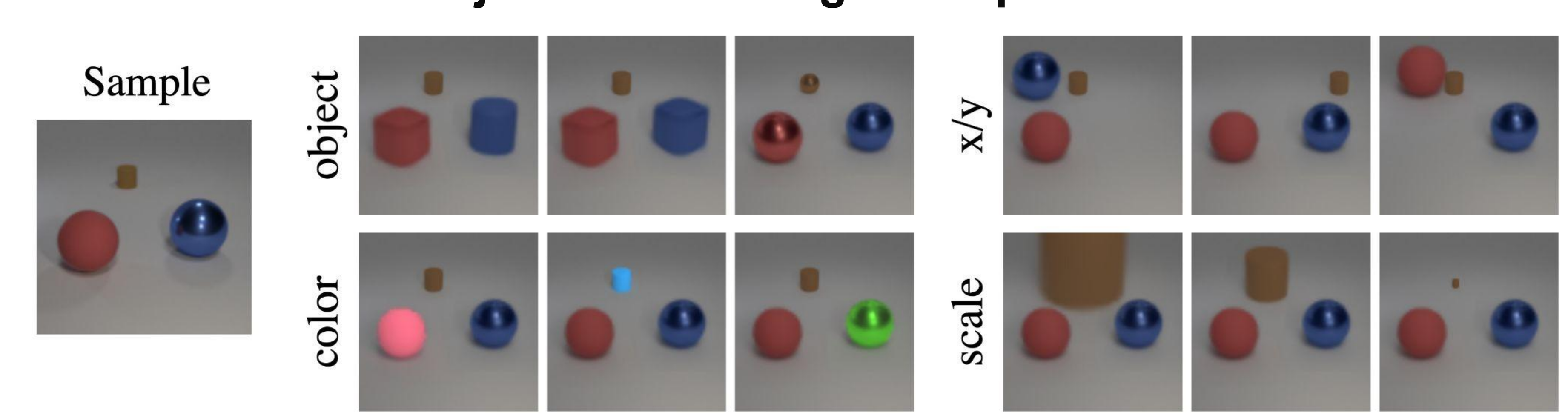
Cosegmentation on Weizmann Horse database



Unsupervised segmentation and image decomposition



Object-centric image manipulation



[1] MONet: Unsupervised Scene Decomposition and Representation, Burgess et al., arXiv 2019
[2] Multi-Object Representation Learning with Iterative Variational Inference, Greff et al., ICML 2019
[3] Object-Centric Learning with Slot Attention, Locatello et al., NeurIPS 2020
[4] Attend, Infer, Repeat: Fast Scene Understanding with Generative Models, Eslami et al., NIPS 2016
[5] Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects, Kosiorek et al., NIPS 2018
[6] Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition, Lin et al., ICLR 2020