Promising or Elusive? Unsupervised Object Segmentation from Real-world Single Images

Anonymous Author(s)
Affiliation
email

Abstract

In this paper, we study the problem of unsupervised object segmentation from single images. We do not introduce a new algorithm, but systematically investigate the effectiveness of existing unsupervised models on challenging real-world images. We firstly introduce four complexity factors to quantitatively measure the distributions of object- and scene-level biases in appearance and geometry for datasets with human annotations. With the aid of these factors, we empirically find that, not surprisingly, existing unsupervised models catastrophically fail to segment generic objects in real-world images, although they can easily achieve excellent performance on numerous simple synthetic datasets, due to the vast gap in objectness biases between synthetic and real images. By conducting extensive experiments on multiple groups of ablated real-world datasets, we ultimately find that the key factors underlying the colossal failure of existing unsupervised models on real-world images is the challenging distributions of object- and scene-level biases in appearance and geometry. Because of this, the inductive biases introduced in existing unsupervised models can hardly capture the diverse object distributions. Our research results suggest that future work should exploit more explicit objectness biases in the network design.

1 Introduction

The capability of automatically identifying individual objects from complex visual observations is a central aspect of human intelligence [52]. It serves as the key building block for higher-level cognition tasks such as planning and reasoning [28]. In last years, a plethora of models have been proposed to segment objects from single static images in an unsupervised fashion: from the early AIR [22] and MONet [8] to the recent SPACE [39], SlotAtt [40], GENESIS-V2 [20], *etc*. They jointly learn to represent and segment multiple objects from a single image, without needing any human annotations in training. This process is often called perceptual grouping/binding or object-centric learning. These methods and their variants have achieved impressive segmentation results on numerous synthetic scene datasets such as dSprites [42] and CLEVR [33]. Such advances come with great expectations that the unsupervised techniques would likely close the gap with fully-supervised methods for real-world visual understanding. However, few work has systematically investigated the true potential of the emerging unsupervised object segmentation models on complex real-world images such as COCO dataset [38]. This naturally raises an essential question:

Is it promising or even possible to segment generic objects from real-world single images using (existing) unsupervised methods?

What is an object? To answer the above question involves another fundamental question: what is an object? Exactly 100 years ago in Gestalt psychology, Wertheimer [60] first introduced a set of Gestalt principles such as proximity, similarity and continuation to heuristically define visual data as objects. However, these factors are highly subjective, whilst the real-world generic objects

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Figure 1: The failure of SlotAtt [40] on three real-world images (right-hand side), although it can perfectly segment simple objects on three synthetic images (left-hand side).

are far more complex with extremely diverse appearances and shapes. Therefore, it is practically impossible to quantitatively define what is an object, *i.e.*, the objectness, from visual inputs (*e.g.*, a set of image pixels). Nevertheless, to thoroughly understand whether unsupervised methods can truly learn objectness akin to the psychological process of humans, it is vital to investigate the underlying factors that potentially facilitate or otherwise hinder the ability of unsupervised models. In this regard, by drawing on Gestalt principles, we instead define a series of new factors to quantitatively measure the complexity of objects and scenes in Section 2. By taking into account both the appearance and geometry of objects and scenes, our complexity factors explicitly assess the difficulty of segmenting a specific object. For example, a chair with colorful textures tend to have higher complexity than a single-color ball for unsupervised methods. With the aid of these factors, we extensively study whether and how existing unsupervised models can discover objects in Section 4.

What is the problem of unsupervised object segmentation from single images? A large number of models [62] aim to tackle the problem of unsupervised object segmentation from single images. They share several key problem settings: 1) all training images do not have any human annotations; 2) every single image has multiple objects; 3) each image is treated as a static data point without any dynamic or temporal information; 4) all models are trained from scratch without requiring any pretrained networks on additional datasets. Ultimately, the goal of these models is to segment all individual objects as accurate as the ground truth human annotations. In this paper, we regard these settings as the basic and necessary part of unsupervised object segmentation from single images, and empirically evaluate how successfully the existing models can exhibit on real-world images.

Contributions and findings. This paper addresses the essential question regarding the potential of unsupervised segmentation of generic objects from real-world single images. Our contributions are:

- We firstly introduce 4 complexity factors to quantitatively measure the difficulty of objects and scenes. These factors are key to investigate the true potential of existing unsupervised models.
- We extensively evaluate current unsupervised approaches in a large-scale experimental study. We implement 4 representative methods and train more than 130 models on 6 curated datasets from scratch. The datasets, code and pretrained models are available at `https://github.com/xxx`.
- We analyze our experimental results and find that: 1) existing unsupervised object segmentation models cannot discover generic objects from single real-world images, although they can achieve outstanding performance on synthetic datasets, as qualitatively illustrated in Figure 1; 2) the challenging distributions of both object- and scene-level biases in appearance and geometry from real-world images are the key factors incurring the failure of existing models; 3) the inductive biases introduced in existing unsupervised models are fundamentally not matched with the objectness biases exhibited in real-world images, and therefore fail to discover the real objectness.

Related Work. Recently, ClevrTex [35] and the concurrent work [43] also study unsupervised object segmentation on single images. Through evaluation on (complex) synthetic datasets only, both works focus on benchmarking the effectiveness of particular network designs of baselines. By comparison, our paper aims to explore what and how the objectness distribution gaps between synthetic and real-world images incur the failure of existing models. The recent work [59] which investigates video object discovery is orthogonal to our work as the motion signals do not exist in single images.

Scope of this research. This paper does not investigate unsupervised object discovery on saliency maps [58], static multi-views or dynamic videos [62]. Recent methods [10, 30] requiring pretrained models on monolithic object images such as ImageNet [47] are not evaluated as well.

2 Complexity Factors

As illustrated in the top row of Figure 2, an individual object, represented by a set of color pixels painted within a mask, can vary significantly given different types of appearance and geometric shape. A specific scene, represented by a set of objects placed within an image, can also differ vastly given different types of relative appearance and geometric layout between objects, as illustrated in the bottom row. Unarguably, such variation and complexity of appearance and geometry in both object level and scene level directly affects human's ability to precisely separate all objects. Naturally, the performance of unsupervised segmentation models are also expected to be influenced by the variation. In this regard, we carefully define the following two groups of factors to quantitatively describe the complexity of different datasets.

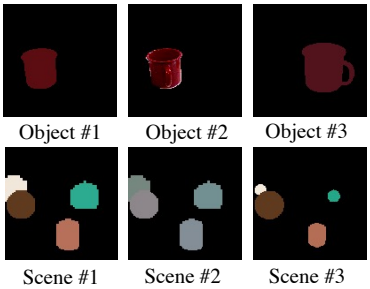

Figure 2: Complexity in appearance and geometry for objects and scenes.

2.1 Object-level Complexity Factors

As to a specific object, all its information can be described by appearance and geometry. Therefore we define the below two factors to measure the complexity of appearance and geometry respectively. Notably, both factors are nicely invariant to the object scale.

- **Object Color Gradient:** This factor aims to calculate how frequently the appearance changes within the object mask. In particular, given the RGB image and mask of an object, we firstly convert RGB into grayscale and then apply Sobel filter [51] to compute the gradients horizontally and vertically for each pixel within the mask. The final gradient value is obtained by averaging out all object pixels. Note that, the object boundary pixels are removed to avoid the interference of background. Numerically, the higher this factor is, the more complex texture and/or lighting effect the object has, and therefore it is likely harder to segment.
- **Object Shape Concavity:** This factor is designed to evaluate how irregular the object boundary is. Particularly, given an object (binary) mask, denoted as $\boldsymbol{M}_{obj} \in \mathbb{R}^{H \times W}$, we firstly find the smallest convex polygon mask ($\boldsymbol{M}_{cvx} \in \mathbb{R}^{H \times W}$) that surrounds the object mask using an existing algorithm [19], and then the object shape concavity value is computed as: $1 - \sum \boldsymbol{M}_{obj} / \sum \boldsymbol{M}_{cvx}$. Clearly, the higher this factor is, the more irregular object shape is, and segmentation is more tricky.

2.2 Scene-level Complexity Factors

As to a specific image, in addition to the object-level complexity, the spatial and appearance relationships between all objects can also incur extra difficulty for segmentation. We define the following two factors to quantify the complexity of relative appearance and geometry between objects in an image.

- **Inter-object Color Similarity:** This factor intends to assess the appearance similarity between all objects in the same image. Specifically, we firstly calculate the average color for each object, and then compute the pair-wise Euclidean distances of object colors, obtaining a $K \times K$ matrix where K represents the object number. The final inter-object color similarity is calculated by averaging the matrix excluding diagonal entries. Intuitively, the higher this factor is, the more similar all objects appear to be, the less distinctive each object is, and it is harder to separate each object.
- **Inter-object Shape Variation:** This factor aims to measure the relative geometry diversity between all objects in the image. We firstly calculate the diagonal length of each object bounding box, and then compute the pair-wise absolute differences for all object diagonal lengths, obtaining a $K \times K$ matrix. The final inter-object shape variation is the average of the matrix excluding diagonal entries. If this factor is higher, it means that objects within the image have more diverse and imbalanced sizes, and therefore segmenting both gigantic and tiny objects is likely more challenging.

By capturing the appearance and geometry in both object and scene levels, the four factors are designed to quantify the complexity of objects and images. For illustration, Figure 3 shows sample images for the four factors at different values. The higher the values, the more complex the objects and scenes. In fact, these factors are carefully selected from more than 10 candidates because they are

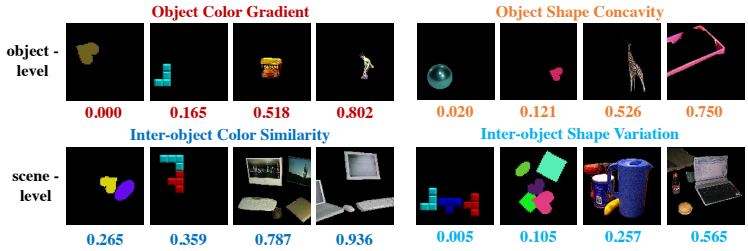

Figure 3: Sample objects and scenes for the four factors at different complexity values.

empirically more suitable to differentiate the gaps between synthetic and real-world images, and they eventually serve as key indicators to diagnose existing unsupervised models in Section 4. Calculation details of the four factors and other candidates are in appendix.

3 Experimental Design

3.1 Considered Methods

A range of works have explored unsupervised object segmentation in recent years. They are typically formulated as (variational) autoencoders (AE/VAE) [36] or generative adversarial networks (GAN) [24]. GAN based models [13, 3, 7, 55, 5, 57, 1] are usually limited to identifying a single foreground object and can hardly discover multiple objects due to the training instabilities, therefore not considered in this paper. As shown in Table 1, the majority of existing models are based on AE/VAE and can be generally divided into two groups according to the object representation:

- **Factor-based models**: Each object is represented by explicit factors such as size, position, appearance, *etc*., and the whole image is a spatial organization of multiple objects. Basically, such representation explicitly enforces objects to be bounded within particular regions.
- **Layer-based models**: Each object is represented by an image layer, *i.e*., a binary mask, and the whole image is a spatial mixture of multiple object layers. Intuitively, this representation does not have strict spatial constrains, and instead is more flexible to cluster similar pixels as objects.

In order to decompose the input images into objects, these approaches introduce different types of network architecture, loss functions, and regularization terms as inductive biases. These biases broadly include: 1) variational encoding which encourages the disentanglement of latent variables; 2) iterative inference which likely ends up with better scene representations over occlusions; 3) object relationship regularization such as depth estimation and autoregressive prior which aims at capturing the dependency of multiple objects; and many other biases. With different combinations of these biases, many methods have shown outstanding performance in synthetic datasets. Among them, we select 4 representative models for our investigation: 1) AIR [22], 2) MONet [8], 3) IODINE [25], and 4) SlotAtt [40]. We also add the fully-supervised Mask R-CNN [29] as an additional baseline for comprehensive comparison. Implementation details are provided in appendix.

3.2 Considered Datasets

We consider two groups of datasets for extensive benchmarking and analysis: 1) three commonly-used synthetic datasets: dSprites [42], Tetris [34] and CLEVR [33], 2) three real-world datasets: YCB [9], ScanNet [17], and COCO [38], representing the small-scale, indoor- and outdoor-level real scenes respectively. Naturally, objects and scenes in different datasets tend to have very different types of biases. For example, the objects in dSprites tend to have the single-color bias, while COCO does not. Generally, the object-level biases can be divided as: 1) appearance biases including different textures and lighting effects, and 2) geometry biases including the object shape and occlusions. Similarly, the scene-level biases include: 1) appearance biases such as the color similarity between all objects, and 2) geometry biases such as the diversity of all object shapes. In fact, our complexity factors introduced in Section 2 are designed to well capture these biases. Table 2 qualitatively summarizes the biases of selected datasets. We may hypothesize that the large gaps of biases between synthetic and real-world datasets would have a huge impact on the effectiveness of existing models.

Table 1: Existing unsupervised models for object segmentation on single images. Each model includes different inductive biases, such as variational autoencoding (VAE), iterative inference (Iter), object relationship regularization (Rel), *etc*.

Factor-based Models		Inductive Biases			Layer-based Models		Inductive Biases		
		VAE	Iter	Rel			VAE	Iter	Rel
CST-VAE [31]	ICLRW'16	✓			Tagger [26]	NIPS'16		✓	
AIR [22]	NIPS'16	✓			RC [26]	ICLRW'16		✓	
SPAIR [16]	AAAI'19	✓			NEM [27]	NIPS'17		✓	
SuPAIR [53]	ICML'19	✓			MONet [8]	arXiv'19	✓		
GMIO [63]	ICML'19	✓	✓		IODINE [25]	ICML'19	✓	✓	
ASR [61]	NeurIPS'19	✓			ECON [56]	ICLRW'20	✓		✓
SPACE [39]	ICLR'20	✓			GENESIS [21]	ICLR'20	✓		✓
GNM [32]	NeurIPS'20	✓		✓	SlotAtt [40]	NeurIPS'20		✓	
SPLIT [11]	arXiv'20	✓			GENESIS-V2 [20]	NeurIPS'21	✓		✓
OCIC [2]	arXiv'20	✓		✓	R-MONet [50]	arXiv'21	✓		✓
GSGN [18]	ICLR'21	✓		✓	CAE [41]	arXiv'22			✓

Table 2: The object- and scene-level biases in appearance and geometry of the considered datasets.

		Synthetic Datasets			Real-world Datasets		
		dSprites [42]	Tetris [34]	CLEVR [33]	YCB [9]	ScanNet [17]	COCO [38]
Object-level Biases							
Appearance	Texture:	simple	simple	simple	diverse	simple	diverse
	Lighting:	no	no	synthetic	real	real	real
Geometry	Shape:	simple	simple	simple	simple	diverse	diverse
	Occlusion:	minor	no	minor	severe	severe	severe
Scene-level Biases							
Appearance	Similarity:	low	low	high	high	high	high
Geometry	Diversity:	low	low	low	high	high	high

To guarantee the fairness and consistency of all experiments, we carefully prepare all six datasets using the following same protocols. Preparation details for each dataset are provided in appendix.

- All images are rerendered or cropped with the same resolution of 128×128.
- Each image has about 2 to 6 solid objects with a blank background.
- Each dataset has about 10000 images for training, 2000 images for testing.

3.3 Considered Metrics

Having the six representative datasets and four existing unsupervised methods at hand, we choose the following metrics to evaluate the object segmentation performance: 1) AP score which is widely used for object detection and segmentation [23], 2) PQ score which is used to measure non-overlap panoptic segmentation [37], and 3) Precision and Recall scores. A predicted mask is considered correct if its IoU against a ground truth mask is above 0.5. All objects are treated as a single class. The blank background is not taken into account for fair comparison. To compute AP, we simply treat the mean value of the soft object mask as the object confidence score. Note that, the alternative metrics ARI [45] and segmentation covering (SC) [4] are not considered as they can be easily saturated.

4 Key Experimental Results

4.1 Can current unsupervised models succeed on real-world datasets?

First of all, we evaluate all baselines on our six datasets separately. In particular, we train each model from scratch on each dataset separately. For fair evaluation, we carefully tune the hyperparameters of each model on every dataset and fully optimize the networks until convergence. Figure 4 compares the quantitative results. It can be seen that all methods demonstrate satisfactory segmentation results on synthetic datasets, especially the recent strong baselines IODINE and SlotAtt. However, not surprisingly, all unsupervised methods fail catastrophically on the three real-world datasets.

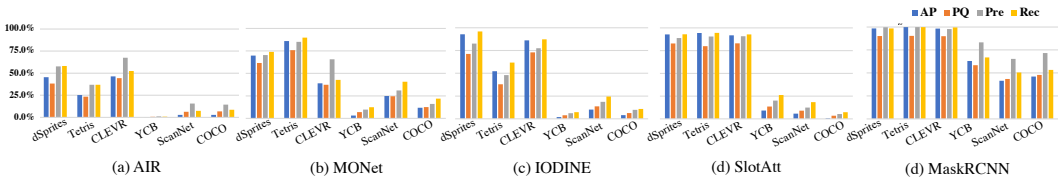

Figure 4: Quantitative results of object segmentation from the five methods on six datasets.

Preliminary Diagnosis: In order to diagnose the colossal failure, we hypothesize that it is because of the huge gaps in objectness biases between two types of datasets. In this regard, we quantitatively compute the distributions of our four complexity factors on the six datasets. In particular, the two object-level factors, *i.e.*, Object Color Gradient and Object Shape Concavity, are computed for each object of the six training splits. The two scene-level factors, *i.e.*, Inter-object Color Similarity and Inter-object Shape Variation, are computed for each image of the six training splits.

From the distributions shown in the top row of Figure 5, we can see that, 1) for the two object-level factors (Subfigs 1-a and 1-c), the three synthetic datasets tend to have extremely lower scores than the real-world datasets, which means that the synthetic objects are more likely have uniform colors and convex shapes; 2) for the two scene-level factors (Subfigs 1-b and 1-d), the images in synthetic datasets tend to include less similar objects in terms of color, which means that multiple objects in real-world scenes are less distinctive in appearance. In addition, the multiple objects in synthetic scenes tend to have similar sizes, whereas real-world scenes usually have diverse object scales in single images. To validate whether these distribution biases are the true reasons incurring the failure, we conduct extensive experiments in Sections 4.2, 4.3 and 4.4.

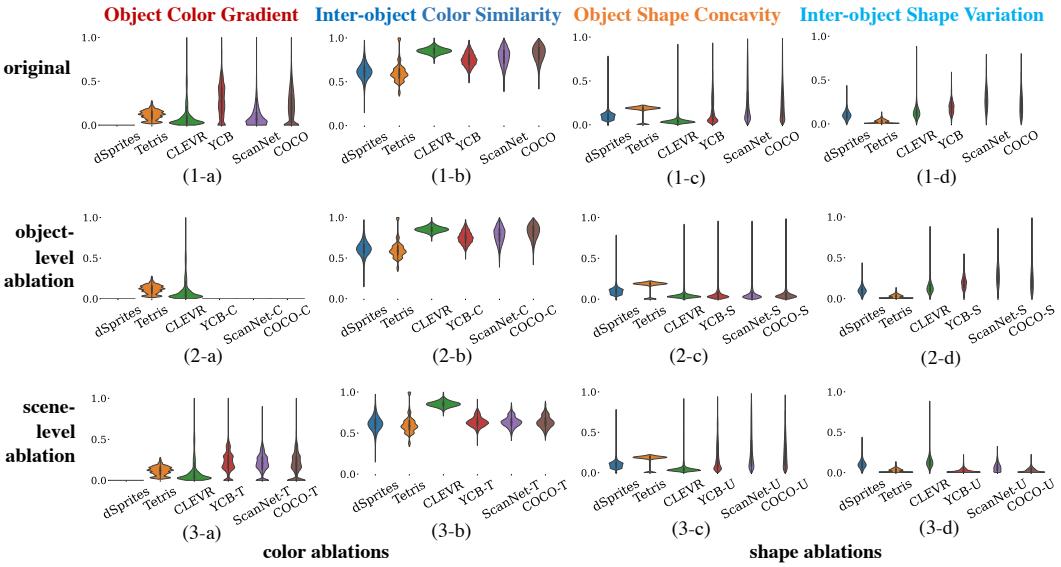

Figure 5: Distributions of the four complexity factors. The *top row* shows the distributions of original synthetic/real-world datasets in Sec 4.1. The *2nd row* shows the distributions of object-level ablated datasets in Sec 4.2. The *3rd row* shows distributions of scene-level ablated datasets in Sec 4.3.

4.2 How do object-level factors affect current models?

In this section, we aim to verify to what extent the distributions of object-level factors affect the segmentation performance. In particular, we conduct the following three ablative experiments.

- *Ablation of Object Color Gradient*: For each object of the three real-world datasets, we only replace all pixel colors by its average color rgb value within each object mask, without touching the object shapes. In this way, the **c**olor gradients of each object are totally erased, thus removing the potential impact of Object Color Gradient. The three ablated datasets are: YCB-C / ScanNet-C / COCO-C.
- *Ablation of Object Shape Concavity*: For each object in the real-world datasets, we find the smallest convex hull [19] for its object mask and then fill the empty pixels by shifting the original object pixels. Basically, this ablation aims to only reduce the irregularity of object **s**hapes, yet retaining the distributions of color gradients. The ablated datasets are: YCB-S / ScanNet-S / COCO-S.

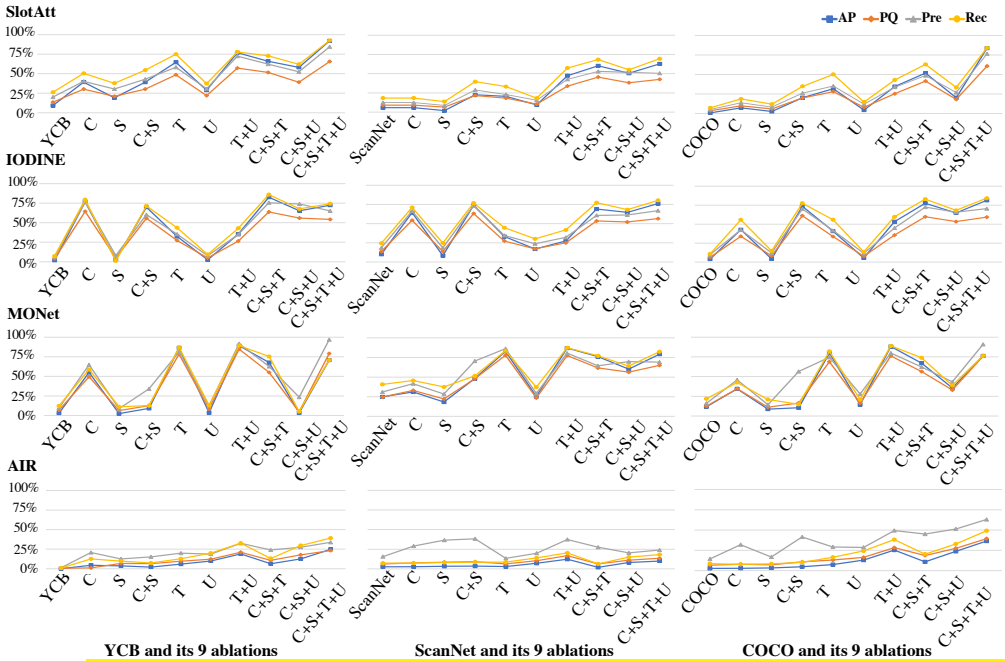

Figure 6: Quantitative results of baselines on the three real-world datasets and their variants. The letters C/S/C+S represent the three ablated datasets in Sec 4.2; T/U/T+U represent the three ablated datasets in Sec 4.3; C+S+T/C+S+U/C+S+T+U represent the three ablated datasets in Sec 4.4.

- *Ablation of both Object Color Gradient and Shape Concavity*: We simply combine the above two ablation for each real-world object, getting datasets: YCB-C+S / ScanNet-C+S / COCO-C+S.

In the second row of Figure 5 (Subfigs 2-a and 2-b), we calculate the new distributions of both Object Color Gradient and Inter-object Color Similarity on the ablated datasets YCB-C/ScanNet-C/COCO-C. It can be seen that the object-level gradients become all zeros in Subfig 2-a, even simpler than the three synthetic datasets. Yet, the distributions of Inter-object Color Similarity are almost the same as the original YCB/ScanNet/COCO, *i.e.*, Subfig 2-b is similar to Subfig 1-b.

Similarly, the second row of of Figure 5 (Subfigs 2-c and 2-d) shows that the distributions of Object Shape Concavity of ablated datasets now become similar to the synthetic datasets, while the distributions of Inter-object Shape Variation keep the same, *i.e.*, Subfig 2-d is similar to Subfig 1-d. Note that, for the ablation of both Object Color Gradient and Shape Concavity, the distributions will be the same as shown in second row of Figure 5. Having the three groups of object-level ablated real-world datasets, we then evaluate segmentation performance of the baselines separately.

Brief Analysis: As shown in Figure 6, we can see that: 1) Once the pixels of real-world objects are replaced by its mean color, *i.e.*, without any color gradients, the object segmentation performance has been significantly improved for almost all methods. 2) Reducing the irregularity of real-world objects can also improve the object segmentation, although not significantly. 3) Overall, these results show that existing methods are more likely to learn the objectness represented by uniform colors and/or regular objects. However, comparing with Figure 4, the results of current ablated datasets in Figure 6 still lag behind the synthetic datasets. This means that there must be some other factors that also potentially affect the object segmentation of existing models.

4.3 How do scene-level factors affect current models?

In this section, we turn to investigate to what extent the distributions of scene-level factors affect the segmentation performance. Particularly, we conduct the following three ablative experiments.

- *Ablation of Inter-object Color Similarity*: In each image of the three real-world datasets, we replace all object textures by a set of new distinctive textures from the existing DTD database [15]. In this way, the multiple objects look more distinctive in appearance, while the per-object texture gradients are roughly preserved. The ablated datasets are denoted: YCB-T / ScanNet-T / COCO-T.

- *Ablation of Inter-object Shape Variation*: In each image of the real-world datasets, we normalize the scales of multiple objects by shrinking or expanding the diagonal length of their bounding boxes, such that the new object sizes tend to be **u**niform. For each object, its shape and texture are linearly scaled up or down. Basically, this aims to remove the diversity of object sizes within single images. The ablated datasets are denoted as: YCB-U / ScanNet-U / COCO-U.
- *Ablation of both Inter-object Color Similarity and Shape Variation*: We simply combine the above two ablation strategies for each real-world image. Ablation details are in appendix. The ablated datasets are denoted as: YCB-T+U / ScanNet-T+U / COCO-T+U.

In the third row of Figure 5 (Subfigs 3-a and 3-b), we calculate the new distributions of both Object Color Gradient and Inter-object Color Similarity on the ablated datasets YCB-T/ScanNet-T/COCO-T. It can be seen that the distributions of scene-level color similarity become more similar to synthetic datasets in Subfig 3-b, while the distributions of object color gradients are still similar as original real-world datasets YCB/ScanNet/COCO, *i.e.*, Subfig 3-a is similar to Subfig 1-a.

Similarly, the third row of Figure 5 shows that the distributions of Inter-object Shape Variation of ablated datasets now become similar to the synthetic datasets in Subfig 3-d, whereas the distributions of Object Shape Concavity are still the same as original real-world datasets YCB/ScanNet/COCO, *i.e.*, Subfig 3-c is similar to Subfig 1-c. Note that, for the ablation of both Inter-object Color Similarity and Shape Variation, the distributions will be the same as the third row of Figure 5. Having these three groups of scene-level ablated real-world datasets, we then separately evaluate each baseline.

Brief Analysis: As shown in Figure 6, we can see that: 1) Once the textures of real-world images are replaced by more distinctive textures, *i.e.*, with lower similarity between object appearance, the object segmentation performance has been surprisingly boosted remarkably for almost all methods. 2) Normalizing object sizes over images can also reasonably improve the object segmentation performance. 3) Overall, these results clearly show that existing unsupervised models significantly favor objectness with distinctive appearance in single images. However, compared with Figure 4, the results on current scene-level ablated datasets are still inferior to the synthetic datasets, meaning that the scene-level factors alone are not enough to explain the performance gap.

4.4 How do object- and scene-level factors jointly affect current models?

In this section, we aim to study how the object- and scene-level factors jointly affect the segmentation performance. In particular, we conduct the following three ablative experiments.

- *Ablation of Object Color Gradient, Object Shape Concavity and Inter-object Color Similarity*: In each image of the three real-world datasets, we replace the object color by averaging all pixels of the distinctive texture, and also replace the object shape with a simple convex hull. The ablated datasets are denoted: YCB-C+S+T / ScanNet-C+S+T / COCO-C+S+T.
- *Ablation of Object Color Gradient, Object Shape Concavity and Inter-object Shape Variation*: In each image of the three real-world datasets, we replace the object color by averaging its own texture, and modify the object shape as convex hull following by size normalization. The ablated datasets are denoted as: YCB-C+S+U / ScanNet-C+S+U / COCO-C+S+U.
- *Ablation of all four factors*: We aggressively combine all four ablation strategies and we get datasets: YCB-C+S+T+U / ScanNet-C+S+T+U / COCO-C+S+T+U.

Since these ablations are conducted independently, the new distributions of four complexity factors on current jointly ablated datasets are exactly the same as the second and third rows of Figure 5.

Brief Analysis: As shown in Figure 6, we can see that: 1) Combining the two object-level factors and either of scene-level factors to ablate the real-world datasets, the segmentation performance can be improved as we expected, especially for IODINE and SlotAtt. 2) If the challenging real-world objects and images are ablated in both object- and scene-level, the segmentation performance of all unsupervised models achieves the same level with three synthetic datasets as shown in Figure 4. 3) Overall, these three groups of experiments demonstrate that the colossal failure of unsupervised models on real-world images involves both object- and scene-level dataset biases.

4.5 Why do current unsupervised models fail on real-world datasets?

As demonstrated in Sections 4.2/4.3/4.4, once the complexity factors are removed from the challenging real-world datasets, existing unsupervised models can perform as excellent as on the synthetic datasets, as qualitatively illustrated in Figure 7. From this, we can safely conclude that the inductive

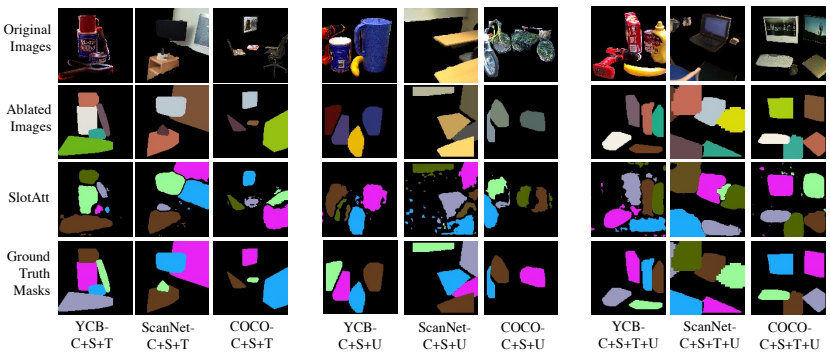

Figure 7: Results of SlotAtt on ablated real-world datasets with both object- and scene-level factors.

biases designed in existing unsupervised models are far from able to match with and fully capture the true and complex objectness biases exhibited in real-world images. Nevertheless, from Figure 6, each baseline tends to favor different objectness biases. In particular,

- AIR [22]: As a factor based model, AIR has a strong spatial-locality bias. Despite its poor segmentation performance across all datasets, there is a notable improvement when inter-object shape variation is ablated from real-world datasets (U / T+U / C+S+U / C+S+T+U). More convincingly, even when all other three factors are ablated (C+S+T), it can be hardly improved. These observations show that object shape variation is a significant factor for AIR.
- MONet [8]: MONet is more sensitive to color-related factors than shape-related factors. The ablations of object color gradient and inter-object color similarity significantly improve its performance, while ablations of object shape concavity and inter-object shape variation make little differences. For the two color-related factors, the scene-level one is more important than the object-level factor. From this, we can see that MONet has a strong dependency on color. Similar colors tend to be grouped together while different colors are separated apart. Furthermore, the ablation on object color gradient alleviates over-segmentation whereas the ablation on inter-object color similarity alleviates under-segmentation. We conjecture that under-segmentation can be a more severe issue for MONet on real-world datasets, leading to a larger sensitivity on the scene-level color factor.
- IODINE [25]: IODINE also has a heavy dependency on both object- and scene-level color-related factors. However, different from MONet, the ablation on object color gradient brings better performance than inter-object color similarity. We speculate it is because the regularization on shape latent alleviates under-segmentation by bias towards more regular shapes. In this way, over-segmentation is the key issue, making the object color gradient a dominant factor.
- SlotAtt [40]: The ablations on all four factors increase the performance of SlotAtt at different levels, among which object- and scene-level color-related factors are more significant. We conjecture that it is because the feature embeddings used by Slot Attention module are learnt from both pixel colors and coordinates, which contributes to its sensitivity to both shape and color factors.

5 Conclusions

We systematically show that existing unsupervised methods are practically impossible to segment generic objects from single real-world images, and investigate the underlying factors that incur the catastrophic failure. With the aid of our carefully designed four object- and scene-level complexity factors, we conduct extensive experiments on multiple groups of ablated real-world objects and images, and safely conclude that the distributions of both object- and scene-level biases in appearance and geometry of real-world datasets are particularly diverse and indiscriminative, such that current unsupervised models cannot segment real objects. Based on this finding, we suggest two main directions for future study: 1) To exploit more discriminative objectness biases such as object motions which expressively describe the ownership of visual pixels as recently explored in [54, 12, 6]. 2) To leverage pretrained features from single-object-dominant datasets which explicitly regard each image as an object as recently studied in [10, 30], although such settings may not be purely unsupervised.

Broader Impact: One limitation of this work is the lack of study about foreground-background segmentation, but our work can still have a positive impact for researchers to design more practically useful models. A potential negative impact could be using our findings to attack existing models.

References

[1] R. Abdal, P. Zhu, N. Mitra, and P. Wonka. Labels4Free: Unsupervised Segmentation using StyleGAN. *ICCV*, 2021.

[2] T. Anciukevicius, C. H. Lampert, and P. Henderson. Object-Centric Image Generation with Factored Depths, Locations, and Appearances. *arXiv:2004.00642*, 2020.

[3] R. Arandjelovic and A. Zisserman. Object Discovery with a Copy-Pasting GAN. *arXiv:1905.11369*, 2019.

[4] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hierarchical Image Segmentation. *TPAMI*, 33(5):898–916, 2011.

[5] S. Azadi, D. Pathak, S. Ebrahimi, and T. Darrell. Compositional GAN: Learning Image-Conditional Binary Composition. *IJCV*, 2020.

[6] D. M. Bear, C. Fan, D. Mrowca, Y. Li, S. Alter, A. Nayebi, J. Schwartz, L. Fei-Fei, J. Wu, J. B. Tenenbaum, and D. L. Yamins. Learning Physical Graph Representations from Visual Scenes. *NeurIPS*, 2020.

[7] A. Bielski and P. Favaro. Emergence of object segmentation in perturbed generative models. *NeurIPS*, 2019.

[8] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner. MONet: Unsupervised Scene Decomposition and Representation. *arXiv:1901.11390*, 2019.

[9] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar. Yale-CMU-Berkeley dataset for robotic manipulation research. *IJRR*, 36(3):261 – 268, 2017.

[10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. *ICCV*, 2021.

[11] R. Charakorn, Y. Thawornwattana, S. Itthipuripat, N. Pawlowski, and N. Dilokthanakul. An Explicit Local and Global Representation Disentanglement Framework with Applications in Deep Clustering and Unsupervised Object Detection. *arXiv:2001.08957*, 2020.

[12] H. Chen, R. Venkatesh, Y. Friedman, J. Wu, J. B. Tenenbaum, D. L. K. Yamins, and D. M. Bear. Unsupervised Segmentation in Real-World Images via Spelke Object Inference. *arXiv:2205.08515*, 2022.

[13] M. Chen, T. Artières, and L. Denoyer. Unsupervised object segmentation by redrawing. *NeurIPS*, 2019.

[14] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. *CVPR*, 2021.

[15] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing Textures in the Wild. *CVPR*, 2014.

[16] E. Crawford and J. Pineau. Spatially Invariant Unsupervised Object Detection with Convolutional Neural Networks. *AAAI*, 2019.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. *CVPR*, 2017.

[18] F. Deng, Z. Zhi, D. Lee, and S. Ahn. Generative Scene Graph Networks. *ICLR*, 2021.

[19] S. Eddins. Binary image convex hull. *https://blogs.mathworks.com/steve/2011/10/04/binary-image-convex-hull-algorithm-notes/*, 2011.

[20] M. Engelcke, O. P. Jones, and I. Posner. GENESIS-V2: Inferring Unordered Object Representations without Iterative Refinement. *NeurIPS*, 2021.

[21] M. Engelcke, A. R. Kosiorek, O. P. Jones, and I. Posner. GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations. *ICLR*, 2020.

[22] S. M. A. Eslami, N. Heess, T. Weber, Y. Tassa, K. Kavukcuoglu, and G. E. Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. *NIPS*, 2016.

[23] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge: A Retrospective. *IJCV*, 111:98–136, 2015.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. *NIPS*, 2014.

[25] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner. Multi-object representation learning with iterative variational inference. *ICML*, 2019.

[26] K. Greff, A. Rasmus, M. Berglund, T. H. Hao, J. Schmidhuber, and H. Valpola. Tagger: Deep unsupervised perceptual grouping. *NIPS*, 2016.

[27] K. Greff, S. Van Steenkiste, and J. Schmidhuber. Neural Expectation Maximization. *NIPS*, 2017.

[28] K. Greff, S. van Steenkiste, and J. Schmidhuber. On the binding problem in artificial neural networks. *arXiv:2012.05208*, 2020.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. *ICCV*, 2017.

[30] O. J. Hénaff, S. Koppula, E. Shelhamer, D. Zoran, A. Jaegle, A. Zisserman, J. Carreira, and R. Arandjelović. *ECCV*, 2022.

[31] J. Huang and K. Murphy. Efficient inference in occlusion-aware generative models of images. *ICLR Workshops*, 2016.

[32] J. Jiang and S. Ahn. Generative neurosymbolic machines. *NeurIPS*, 2020.

[33] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *CVPR*, 2017.

[34] R. Kabra, C. Burgess, L. Matthey, R. L. Kaufman, K. Greff, M. Reynolds, and A. Lerchner. Multi-Object Datasets. *https://github.com/deepmind/multi-object-datasets/*, 2019.

[35] L. Karazija, I. Laina, and C. Rupprecht. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. *NeurIPS*, 2021.

[36] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *ICLR*, 2014.

[37] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic Segmentation. *CVPR*, 2019.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. *ECCV*, 2014.

[39] Z. Lin, Y.-F. Wu, S. V. Peri, W. Sun, G. Singh, F. Deng, J. Jiang, and S. Ahn. SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition. *ICLR*, 2020.

[40] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-Centric Learning with Slot Attention. *NeurIPS*, 2020.

[41] S. Löwe, P. Lippe, M. Rudolph, and M. Welling. Complex-Valued Autoencoders for Object Discovery. *arXiv:2204.02075*, 2022.

[42] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dSprites - Disentanglement testing Sprites dataset. *https://github.com/deepmind/dsprites-dataset/*, 2017.

[43] S. Papa, O. Winther, and A. Dittadi. Inductive Biases for Object-Centric Representations in the Presence of Complex Textures. *arXiv:2204.08479*, 2022.

[44] D. D. Polsby and R. D. Popper. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale L. & Pol'y Rev.*, 9, 1991.

[45] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[46] D. J. Rezende and F. Viola. Taming vaes. *arXiv:1810.00597*, 2018.

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, (115):211–252, 2015.

[48] J. E. Schwartzberg. Reapportionment, gerrymanders, and the notion of compactness. *Minn. L. Rev.*, 50:443, 1965.

[49] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[50] Shengxin Qian. R-MONet: Region-Based Unsupervised Scene Decomposition and Representation via Consistency of Object Representations. *OpenReview.Net*, 2021.

[51] I. Sobel and G. Feldman. A 3x3 Isotropic Gradient Operator for Image Processing. *Pattern Classification and Scene Analysis*, pages 271–272, 1973.

[52] E. S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson. Origins of knowledge. *Psychological Review*, 99(4):605–632, 1992.

[53] K. Stelzner, R. Peharz, and K. Kersting. Faster attend-infer-repeat with tractable probabilistic models. *ICML*, 2019.

[54] M. Tangemann, S. Schneider, J. von Kügelgen, F. Locatello, P. Gehler, T. Brox, M. Kümmerer, M. Bethge, and B. Schölkopf. Unsupervised Object Learning via Common Fate. *arXiv:2110.06562*, 2021.

[55] S. van Steenkiste, K. Kurach, J. Schmidhuber, and S. Gelly. Investigating object compositionality in Generative Adversarial Networks. *Neural Networks*, 2020.

[56] J. von Kügelgen, I. Ustyuzhaninov, P. Gehler, M. Bethge, and B. Schölkopf. Towards causal generative scene models via competition of experts. *ICLR Workshops*, 2020.

[57] A. Voynov, S. Morozov, and A. Babenko. Object Segmentation Without Labels with Large-Scale Generative Models. *ICML*, 2021.

[58] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Salient Object Detection in the Deep Learning Era: An In-Depth Survey. *TPAMI*, 2021.

[59] M. A. Weis, K. Chitta, Y. Sharma, W. Brendel, M. Bethge, A. Geiger, and A. S. Ecker. Benchmarking Unsupervised Object Representations for Video Sequences. *JMLR*, 2021.

[60] M. Wertheimer. Untersuchungen zur Lehre yon der Gestalt. *Psychologische Forschung*, (4):301–350, 1923.

[61] K. Xu, C. Li, J. Zhu, and B. Zhang. Multi-object generation with amortized structural regularization. *NeurIPS*, 2019.

[62] J. Yuan, T. Chen, B. Li, and X. Xue. Compositional Scene Representation Learning via Reconstruction: A Survey. *arXiv:2202.07135*, 2022.

[63] J. Yuan, B. Li, and X. Xue. Generative modeling of infinite occluded objects for compositional scene representation. *ICML*, 2019.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Our main contribution is the investigation and analysis on existing approaches.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] On the one hand, it would be too computationally expensive. On the other hand, it has been studies extensively in the methods we are investigating.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]