Top-Down Guidance for Learning Object-Centric Representations

Junhong Zou[1,2], **Xiangyu Zhu**[1,2*], **Zhaoxiang Zhang**[1,2,3] and **Zhen Lei**[1,2,3,4]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAIR, HKSIS, Chinese Academy of Sciences, Hong Kong, China
[4]School of Computer Science and Engineering, the Faculty of Innovation Engineering, M.U.S.T, Macau, China
{zoujunhong2022, xiangyu.zhu, zhaoxiang.zhang, zhen.lei}@ia.ac.cn

Abstract

Humans' innate ability to decompose scenes into objects allows for efficient understanding, predicting, and planning. In light of this, Object-Centric Learning (OCL) attempts to endow networks with similar capabilities, learning to represent scenes with the composition of objects. However, existing OCL models only learn through reconstructing the input images, which does not assist the model in distinguishing objects, resulting in suboptimal object-centric representations. This flaw limits current object-centric models to relatively simple downstream tasks. To address this issue, we draw on humans' top-down vision pathway and propose Top-Down Guided Network (TDGNet), which includes a top-down pathway to improve object-centric representations. During training, the top-down pathway constructs guidance with high-level object-centric representations to optimize low-level grid features output by the backbone. While during inference, it refines object-centric representations by detecting and solving conflicts between low- and high-level features. We show that TDGNet outperforms current object-centric models on multiple datasets of varying complexity. In addition, we expand the downstream task scope of object-centric representations by applying TDGNet to the field of robotics, validating its effectiveness in downstream tasks including video prediction and visual planning. Code will be available at https://github.com/zoujunhong/RHGNet.

1 Introduction

Humans are skilled at decomposing visual scenes into the compositions of objects [Kahneman *et al.*, 1992], which is crucial for humans' efficient understanding, predicting, and planning. Inspired by this property, Object-Centric Learning (OCL) seeks to achieve human-like representations in neural networks. Specifically, models are trained in a self-supervised manner to represent visual signals, such as images or videos with a set of latent vectors which are referred to as 'slots' [Greff *et al.*, 2019; Locatello *et al.*, 2020]. Ideally, each slot corresponds to an object in the scene.

Previous methods [Locatello *et al.*, 2020; Jia *et al.*, 2023; Singh *et al.*, 2021; Jiang *et al.*, 2023] have achieved considerable progress in OCL. However, we observe that the performance of existing methods is highly unstable across different scenarios. For example, in Figure 1, we present the cases where current models show sup-optimal object understanding and decompose scenes in an inferior manner: some objects are missed or split into parts. This expresses a concern that current models highly rely on the inductive biases of model structures and visual scenes to learn object-centric representations, posing challenges for adapting to various scenarios and tackling downstream tasks.

We attribute this issue to the fact that current OCL models typically adopt an auto-encoding paradigm that learns object-centric representations by encoding images into slots and using these slots to reconstruct images. However, reconstruction loss does not tell apart objects: models do not need to decompose the scene according to objects. As a result, these models fail to learn distinguishable features at the backbone. For instance, in Figure 1 and 6, we visualize the output features of the backbone when the object-centric models are trained solely with reconstruction loss. It is observed that these features are blurry and fail to align with the edges of objects. Moreover, the features of small objects are particularly challenging to distinguish from the background. Such indistinguishable feature makes it difficult for the model to decide how to assign features to slots, resulting in suboptimal object-centric representations.

To address this issue, we propose to guide the model by an additional top-down pathway. This coincides with humans' perceptual learning process, where theories about human vision [Hochstein and Ahissar, 2002; Wolfe, 2021] argue that humans first learn concepts at high-level consciousness and then use the high-level perception to guide the learning of low-level neurons. Drawing on this mechanism, we propose **Top-Down Guided Network** (**TDGNet**) that introduces the top-down pathway to use the high-level representations (i.e., the slots) to guide the low-level features output by the backbone. Specifically, we obtain a high-level guidance by weighted summing the slots according to their masks, and then introduce a projection network to predict this guidance signal using low-level features. In this way, the model tends

*Corresponding author

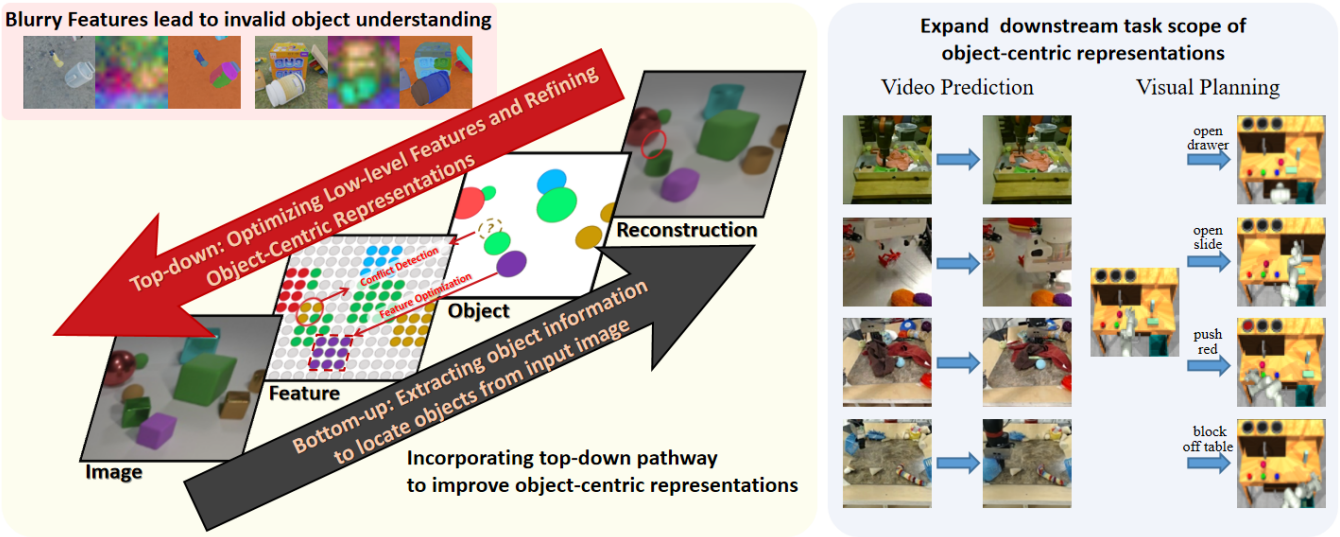


Figure 1: **Overview**. Observing that existing object-centric models fail to learn distinguishable features, limiting the models' object understanding ability, we propose TDGNet that introduces a top-down pathway to help optimize the low-level features output by the backbone and refine object-centric representation. Furthermore, we apply the refined representations to the field of robotics to demonstrate that TDGNet can adapt to a broad range of tasks and scenarios.

to cluster low-level features belonging to the same slots, and otherwise keep them apart, thus making low-level features more distinguishable. Moreover, we extend this concept to the inference phase, introducing a conflict detection method designed to refine object-centric representations during inference: when a feature's prediction is far from all existing slots or close to multiple slots, it may represent a suboptimal perception such as missing objects or splitting objects into parts, which we call a conflict. We detect and solve such conflicts by adding or merging slots, thus refining the object-centric representations.

We evaluate TDGNet and compare it with current SOTA models on multiple tasks. We first introduce CLEVR-Tex [Karazija *et al.*, 2021], MOVi-C [Greff *et al.*, 2022] and COCO to evaluate the object-centric representations, where TDGNet outperforms current SOTA models in terms of common object discovery metrics. Furthermore, we expand the downstream task scope of TDGNet by applying it to the field of robotics. We introduce RoboNet [Dasari *et al.*, 2020] and VP^2 [Tian *et al.*, 2023], to evaluate TDGNet with downstream tasks including video prediction and visual planning, demonstrating that TDGNet adapts well to these tasks.

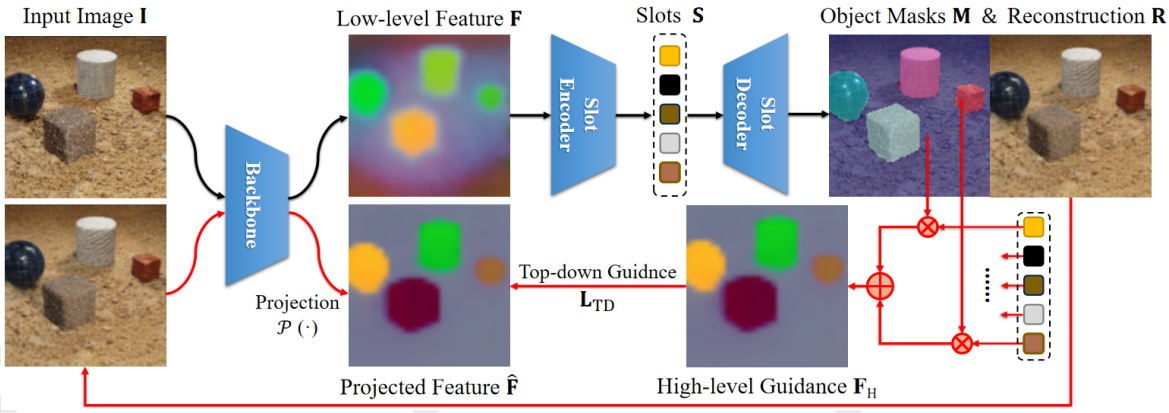
To sum up, our contributions are summarized as follows:

- Drawing on the top-down visual pathway of humans, we propose TDGNet, which incorporates a top-down pathway that constructs high-level guidance to optimize the low-level features, thereby improving object-centric representations.
- Based on the top-down pathway, we propose conflict detection to discover perceptual errors and further refine object-centric representations during inference.
- We demonstrate the SOTA performance of TDGNet in object-centric representation tasks. Besides, we introduce it into more complex robotic scenarios and verify that it works effectively in the video prediction and visual planning task.

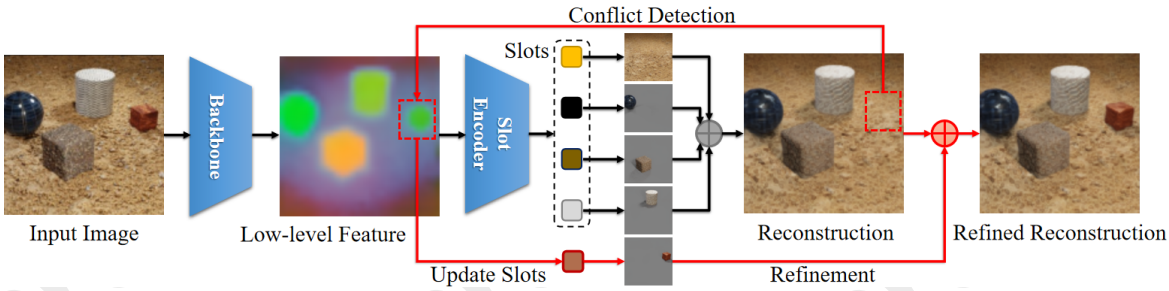
2 Related Work

Object-Centric Learning. Most current OCL methods follow an auto-encoding paradigm that first encodes input signals into several slots and reconstructs the original signal with these slots. Earlier works, including IODINE [Greff *et al.*, 2019], MONet [Burgess *et al.*, 2019] and GENESIS [Engelcke *et al.*, 2019], accomplish this task by using multiple encoder-decoder structures. Slot-Attention [Locatello *et al.*, 2020] proposed an iterative attention method that allows slots to compete for input image segments and conduct segmentation. A critical issue of current OCL methods is how to generalize to more complex scenes. BO-QSA [Jia *et al.*, 2023], I-SA [Chang *et al.*, 2022] and InvariantSA [Biza *et al.*, 2023] focus on query optimization, which uses learnable parameters to initialize slots. SLATE [Singh *et al.*, 2021] and LSD [Jiang *et al.*, 2023] attempt to improve the decoder structure, introducing transformer-based and diffusion-based decoders to enhance the model's reconstruction ability. DINOSAUR [Seitzer *et al.*, 2022] proposes that the simple reconstruction task is insufficient to distinguish objects and replaces the reconstruction objective with the output feature of DINO [Caron *et al.*, 2021].

Top-down connections in human vision. Human brain transmits high-level semantic information to low-level neurons [Beck and Kastner, 2009], resulting in a biased competition between different objects to control attention. In addition, the brain can receive task-relevant information and inhibit irrelevant neurons to improve the efficiency of completing tasks [Karimi-Rouzbahani *et al.*, 2017]. Reverse hierarchy theory [Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004] generalizes neural connections in the human brain into two functional pathways: the bottom-up path-



(a) Top-down Guidance (TDG) during training.



(b) Conflict detection (CD) for refining representations during inference.

Figure 2: **Architecture of Top-Down Guided Network (TDGNet)**. The proposed TDGNet acquires the initial perception through the bottom-up pathway (black arrow) and refines its object-centric representations with the top-down pathway (red arrow). **(a)** During training, TDGNet uses slots and object masks to guide the backbone to produce distinguishable low-level features. **(b)** During Inference, TDGNet refines the slots by detecting conflicts between its slots and the low-level features.

way works implicitly, acquiring the gist of the scene rapidly, and the top-down pathway returns to lower-level neurons to bring detailed information into consciousness.

Top-down connections in Neural Networks. It has long been explored to incorporate top-down feedback connections into neural networks. [Liang and Hu, 2015] introduce recurrent connections into convolutional networks. [Wen *et al.*, 2018] propose to achieve predictive coding through a network with feedback connections. Recent works have utilized a top-down pathway to solve different visual or multi-modal tasks, including semantic segmentation [Yin *et al.*, 2022; Liu *et al.*, 2024], visual saliency [Ramanishka *et al.*, 2017] and vision question & answering [Anderson *et al.*, 2018]. Most of these models integrate features from multiple layers through a parameterized network module, introducing additional depth into the network through feedback connections. However, there is no evidence that they achieve the visual functions of human feedback connections.

3 Method

3.1 Preliminary: Auto-encoding-based Object-Centric Models

The architecture of TDGNet is shown in Figure 2. A typical auto-encoding-based object-centric model serves as our **bottom-up pathway**: the backbone $\mathcal{E}_B$ first extract the low-level features $\mathbf{F} \in \mathbb{R}^{C_f \times H/s \times W/s}$ from the input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$. Then a slot encoder $\mathcal{E}_S$ encodes $\mathbf{F}$ into K slots $\mathbf{S} \in \mathbb{R}^{K\times C_s}$, which is regarded as the high-level representations. Finally, a slot decoder $\mathcal{D}_S$ decodes slots into reconstructions $\mathbf{R} \in \mathbb{R}^{3\times H\times W}$ and object masks $\mathbf{M} \in \mathbb{R}^{K\times H\times W}$. Formally, the bottom-up pathway is described as:

$$\begin{cases} \mathbf{F} = \mathcal{E}_B(\mathbf{I}), \\ \mathbf{S} = \mathcal{E}_S(\mathbf{F}), \\ \mathbf{R}, \mathbf{M} = \mathcal{D}_S(\mathbf{S}). \end{cases} \tag{1}$$

The bottom-up pathway is optimized by reconstructing input images. Here we use a combination of L1 loss and perceptual loss (LPIPS) [Zhang *et al.*, 2018] for optimization. The reconstruction loss is written as:

$$\mathbf{L}_{\text{rec}} := \|\mathbf{R} - \mathbf{I}\|_1 + \text{LPIPS}(\mathbf{R}, \mathbf{I}) \tag{2}$$

3.2 Learning Object-Centric Representations with Top-down Pathway

We introduce a top-down pathway to refine the object-centric representations, which works in two ways: during training, it constructs guidance with high-level slots to optimize low-level features output by the backbone; during inference, it refines the slots by detecting and solving conflicts between slots and low-level features.

Model	CLEVRTex			CAMO			OOD			MOVi-C			
	↑ARI-FG	↑mIoU	↓MSE	↑ARI-FG	↑mIoU	↓MSE	↑ARI-FG	↑mIoU	↓MSE	↑ARI-FG	↑mIoU	↑mBO	↓MSE
SLATE [Singh *et al.*, 2021]	45.4	49.5	498	43.5	37.7	349	46.5	35.4	550	49.5	37.8	39.4	526
LSD [Jiang *et al.*, 2023]	64.4	62.5	237	62.6	60.8	245	58.9	56.4	492	52.3	44.1	45.6	661
BO-QSA [Jia *et al.*, 2023]	80.5	46.7	268	72.6	41.5	246	72.5	37.1	805	52.9	33.1	36.4	157
InvariantSA [Biza *et al.*, 2023]	92.9	72.4	177	86.2	65.6	196	84.4	66.7	578	35.7	26.0	26.9	484
DINOSAUR [Seitzer *et al.*, 2022]	88.9	52.6	-	83.5	51.3	-	83.1	51.9	-	67.8	31.2	38.2	-
TDGNet	94.2	80.3	65	88.9	76.3	82	84.1	69.6	302	61.2	52.9	53.5	151
+CD	**94.8**	**80.5**	**63**	**89.5**	**77.0**	**74**	**84.8**	**71.9**	**291**	**68.5**	**55.6**	**57.1**	**148**

Table 1: Model performance comparison on CLEVRTex and MOVi-C. CAMO and OOD represent CLEVRTex-CAMO and CLEVRTex-OOD where models trained on CLEVRTex are directly evaluated without finetune.

COCO	ARI-FG	$\mathrm{mBO}^{\mathrm{i}}$	mIoU
MLP-based methods			
SA [Locatello *et al.*, 2020]	17.5	18.2	12.2
BO-QSA [Jia *et al.*, 2023]	35.7	26.0	26.9
DINOSAUR-mlp [Seitzer *et al.*, 2022]	40.5	27.7	26.4
DINOSAUR-mlp + DINOv2	42.9	28.9	27.3
TDGNet (ours)	**45.0**	**29.6**	**28.5**
Transformer/Diffusion-based methods			
SLATE [Singh *et al.*, 2021]	23.2	20.2	19.3
LSD [Jiang *et al.*, 2023]	37.0	34.8	32.2
DINOSAUR-tf [Seitzer *et al.*, 2022]	32.3	32.0	30.0
SPOT [Kakogeorgiou *et al.*, 2024]	37.0	35.0	33.0
TDGNet (ours)	**37.3**	**35.6**	**33.2**

Table 2: Unsupervised object discovery result on COCO. Higher is better for all the metrics.

Top-down Guidance during Training

We first introduce how the top-down pathway constructs the **Top-Down Guidance** (**TDG**) to optimize low-level features during training. As shown in Figure 2(a), the bottom-up pathway provides initial object-centric representations with slots. The top-down pathway utilizes these slots to construct guidance for training. Ideally, it makes features from the same slot more similar than those from different ones.

The guidance is constructed through high-level representations including object masks $\mathbf{M}$ and slots $\mathbf{S}$. Formally, the high-level guidance $\mathbf{F}_{\mathrm{H}}$ is obtained through the sum of $\mathbf{S}$, weighted by $\mathbf{M}$ at each spatial location:

$$\mathbf{F}_{\mathrm{H}} = \mathcal{SG}(\mathrm{Sum}(\mathbf{S} * \mathbf{M}, \mathrm{axis} = \mathrm{slots})). \tag{3}$$

Here $\mathcal{SG}$ represents the stop-gradient operation. We stop the gradient of high-level signals (namely $\mathbf{M}$ and $\mathbf{S}$) so that the guidance only works on the low-level features.

Subsequently, we introduce a projection network $\mathcal{P}$ that uses low-level features to predict the high-level guidance $\mathbf{F}_{\mathrm{H}}$. In this way, the low-level features are required to predict their corresponding slots. Considering that $\mathbf{F}_{\mathrm{H}}$ are produced with the slots $\mathbf{S}$, it provides more accurate object regions for the reconstructed images $\mathbf{R}$ than the input images $\mathbf{I}$. Therefore, we let $\mathcal{P}$ project the low-level features extracted from $\mathbf{R}$ for more accurate guidance. Formally, we re-input $\mathbf{R}$ into the backbone $\mathcal{E}_B$ to extract its low-level features $\widehat{\mathbf{F}}$:

$$\widehat{\mathbf{F}} = \mathcal{E}_B(\mathcal{SG}(\mathbf{R})). \tag{4}$$

Finally, we use the projection network $\mathcal{P}$ to predict $\mathbf{F}_{\mathrm{H}}$ with $\widehat{\mathbf{F}}$. In our method, slots $\mathbf{S}$ are normalized to unit length, and the distance is measured through cosine similarity CosSim. The top-down guidance loss $\mathbf{L}_{\mathrm{TD}}$ is formulated as below:

$$\mathbf{L}_{\mathrm{TD}} := 1 - \mathrm{CosSim}(\mathcal{P}(\widehat{\mathbf{F}}), \mathbf{F}_{\mathrm{H}}). \tag{5}$$

Overall, TDGNet is trained by $\mathbf{L}$, the weighted sum of the reconstruction loss and the top-down guidance loss:

$$\mathbf{L} = \mathbf{L}_{\mathrm{rec}} + \lambda_{\mathrm{TD}}\mathbf{L}_{\mathrm{TD}}. \tag{6}$$

Conflict Detection during Inference

During training, we require low-level features to predict their corresponding slots with the projection network $\mathcal{P}$, which provides a method for the model to refine the slots during inference, which we call **conflict detection (CD)**: Ideally, the prediction of a low-level feature should be close to its corresponding slot. This indicates two facts: (i) When a feature's prediction is far away from all the slots, it represents a wrong perception, such as an undiscovered object, and (ii) When a feature's prediction is close to multiple slots, these slots may represent a single object divided into multiple parts. We resolve such conflicts by adding and merging slots, thereby improving the object-centric representations.

Specifically, we first solve the first issue. After the model extracts low-level features $\mathbf{F}$ and slots $\mathbf{S}$ from the images with the bottom-up pathway, we use the projection network $\mathcal{P}(\cdot)$ to produce a prediction $\mathcal{P}(\mathbf{F})$ and computes the cosine distance between $\mathbf{S}$ and $\mathcal{P}(\mathbf{F})$ and acquire the conflict $\mathbf{C}$. Here we define $\mathbf{C}$ as the distance from $\mathcal{P}(\mathbf{F})$ to their nearest slot:

$$\mathbf{C} := \min(1 - \mathrm{CosSim}(\mathcal{P}(\mathbf{F}), \mathbf{S})). \tag{7}$$

We set a threshold th to determine whether a conflict is large or not. Repetitively, we select $\hat{f}$ from $\mathbf{F}$ that has the largest conflict. If the conflict of $\hat{f}$ exceeds th, we add $\mathcal{P}(\hat{f})$ to the slot set. This process is repeated until all the conflicts are lower than th. After that, we use an agglomerative clustering algorithm to merge slots with less cosine distance than th, thus mitigating the situation where an object is divided into multiple parts. As for the choice of th, we propose a heuristic method that for each trained model, we calculate the average distance between slots and set th as half of this distance.

4 Experiments

In this section, we first evaluate the object-centric representations with the object discovery task, demonstrating TDGNet's superior object discovery performance across multiple scenes of varying complexity. We then evaluate the performance of the learned object-centric representations in downstream tasks such as unconditional/conditional video prediction and video prediction for visual planning. Specifically, for conditional video prediction and visual planning, we introduce robotics benchmarks [Dasari *et al.*, 2020; Tian *et al.*, 2023] to demonstrate that TDGNet's object-centric representations can be applied in a variety of scenarios.

4.1 Unsupervised Object Discovery

Setup. We first introduce the object discovery task for evaluating object-centric representations. We use three common datasets: CLEVRTex [Karazija *et al.*, 2021], MOVi-C [Greff *et al.*, 2022], and COCO [Caesar *et al.*, 2018]. The model's generalization ability is also evaluated using CLEVRTex-OOD and -CAMO, two out-of-distribution test sets. The complexity of the datasets varies. Objects in CLEVRTex have regular shapes, but the challenge is that it adds texture maps to objects and backgrounds, increasing the appearance complexity of objects. MOVi-C takes a further step to use realistic, richly textured objects from the GSO dataset [Downs *et al.*, 2022] to create multi-object scenes. COCO contains a large number of natural scene images, and the variability of object appearance has increased significantly. Following previous works [Karazija *et al.*, 2021], we use the foreground adjusted rand index (ARI-FG) [Rand, 1971], mean IoU (mIoU), and mean square error (MSE) to evaluate the models' ability to discover objects and reconstruct the images. For MOVi-C and COCO, we follow [Jiang *et al.*, 2023; Seitzer *et al.*, 2022] to include mean best overlapping (mBO) for evaluation.

Result. Table 1 displays quantitative comparison results on CLEVRTex and MOVi-C. In the comparison, we adopt two inference modes of TDGNet: one that only initiates the bottom-up pathway and another that applies the conflict detection (CD) operation to acquire a refined perception. TDGNet with only the bottom-up pathway, outperforms most current models. Conflict detection provides a further performance boost. TDGNet also generalizes well on CLEVRTex-OOD and -CAMO, outperforming current SOTA models by a large margin.

On MOVi-C, SOTA models, such as LSD and DINOSAUR, rely on reconstructing features from pre-training backbones to generalize to complex scenarios. However, these models show significant biases in different evaluation metrics. For instance, DINOSAUR offers an ARI-FG far superior to other models but provides lower mIoU and mBO. On the other hand, the performance of LSD is just the opposite, with lower ARI-FG and higher mIoU and mBO. Our TDGNet, by contrast, outperforms previous models in all the metrics, demonstrating a more comprehensive object discovery capability. Figure 3 gives examples to illustrate the improvement. DINOSAUR fails to fit the edges and cannot segment the background holistically on MOVi-C. LSD tends to divide objects into multiple parts as is circled. TDGNet, instead, succeeds in segmenting backgrounds, as well as showing better object understanding where the object masks fit the outlines of objects well and do not divide objects into parts, thus achieving higher scores on all metrics.

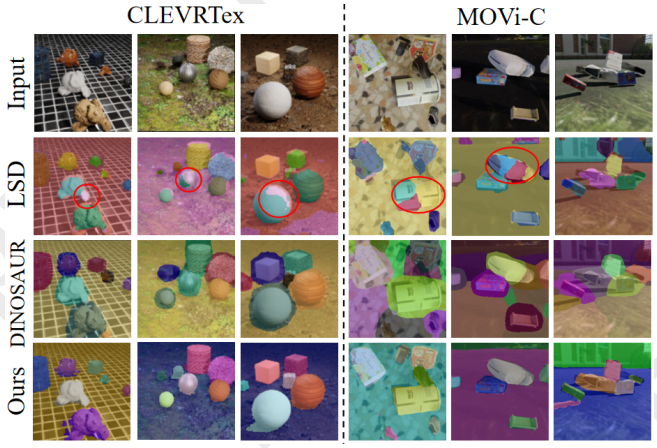


Figure 3: Segmentation results on CLEVRTex and MOVi. TDGNet successfully distinguishes the image's background while segmenting objects with more correct boundaries.

We further provide the comparison result on the COCO dataset in Table 2. Following [Seitzer *et al.*, 2022], we discuss the situations when different types of decoders are used. For the MLP-based decoder, we adopt DINOSAUR as the baseline and use DINO_v2 as the backbone. Compared with DINOSAUR trained only through reconstruction, we have improved by 2.1, 0.7, and 1.2 respectively, according to ARI-FG, mBO, and mIoU. For the transformer-based decoder, we choose SPOT as the baseline and combine it with TDGNet, achieving performance improvements of 0.3, 0.6, and 0.2 in terms of the three metrics.

4.2 Exploring Object-Centric Representation in Predicting the World's State

We believe that object-centric learning, which uses "objects" as the basic unit to represent visual scenes, is worth exploring in world modeling. Here we explore the field of robotics, introducing video prediction and visual planning tasks to evaluate whether the object-centric representations from TDGNet benefit the prediction of the world's states. In video prediction tasks, after training TDGNet, we additionally train an auto-regressive Transformer to predict the future slots, and then decode these slots to acquire future frames. Furthermore, the visual planning task evaluates the video prediction model by using the prediction results in control tasks.

Video Prediction

Setup. We conduct video prediction on the MOVi-C and RoboNet datasets. MOVi-C, a widely used multi-object dataset, contains videos of the interactions of moving objects. Following previous works [Song *et al.*, 2024], the model predicts 8 future frames from 6 context frames with no conditions. We primarily compare TDGNet to existing Object-Centric models in MOVi-C to ensure that the slots provided

MOVi-C	↑PSNR	↑SSIM	↓LPIPS
SlotFormer [Wu *et al.*, 2023]	19.5	45.6	53.4
OCVP [Villar-Corrales *et al.*, 2023]	19.9	50.2	45.0
OCK [Song *et al.*, 2024]	<u>21.0</u>	<u>59.3</u>	<u>37.0</u>
TDGNet (ours)	**22.6**	**66.3**	**25.9**

Table 3: Unconditional video prediction on MOVi-C. LPIPS and SSIM scores are scaled by 100 for convenient display.

RoboNet	↓FVD	↑PSNR	↑SSIM	↓LPIPS
MaskViT [Gupta *et al.*, 2022]	211.7	20.4	67.1	17.0
iVideoGPT [Wu *et al.*, 2024]	<u>197.9</u>	<u>23.8</u>	**80.8**	<u>14.7</u>
TDGNet (ours)	**187.3**	**23.9**	<u>79.7</u>	**14.2**

Table 4: Conditional video prediction results on RoboNet. LPIPS and SSIM scores are scaled by 100 for convenient display.

Dataset	Components				Metrics		
	L1	LP	TDG	CD	ARI-FG	mBO	mIoU
CLEVRTex	✓				88.3	74.2	73.1
	✓	✓			90.9	78.4	77.6
	✓	✓	✓		94.2	80.4	80.3
	✓	✓	✓	✓	**94.8**	**80.8**	**80.5**
MOVi-C	✓				52.9	36.4	33.1
	✓	✓			58.6	46.8	44.9
	✓	✓	✓		61.2	53.5	52.9
	✓	✓	✓	✓	**68.5**	**57.1**	**55.6**

Table 5: Ablation on the components of TDGNet. 'L1' and 'LP' represent the L1 and LPIPS loss used for reconstruction, while 'TDG' and 'CD' represent top-down guidance and conflict detection.

by TDGNet outperform existing models. RoboNet is a large-scale robot dataset that contains a large number of videos of object manipulation with robotic arms. The models are required to predict 10 future frames given 2 context frames and the robotic arm's action as conditions. We compare TDGNet to existing conditional video prediction models on RoboNet. Following previous work, we use SSIM, PSNR, and LPIPS for evaluation on MOVi-C, and additionally introduce FVD for RoboNet.

Result. We list the comparison results in Table 3 and 4. For MOVi-C, using the slots extracted by TDGNet for prediction significantly improves the performance, outperforming other models by a large margin. For RoboNet, we provide results on 256 resolution. We observe that our model outperforms iVideoGPT, the current SOTA model, in three out of the four evaluation metrics. In Figure 4, we provide the visualization results on RoboNet, demonstrating that our model accurately simulates the interaction between robots and objects.

Visual Planning

Setup. As is discussed in previous works [Tian *et al.*, 2023], the video prediction performance sometimes cannot reflect whether models correctly predict the world's state. Therefore, we use their proposed VP^2 benchmark [Tian *et al.*, 2023] to evaluate the performance of TDGNet in the visual planning task. In VP^2, a given agent uses the predicted frames from video prediction models to complete various control tasks. All elements are provided except the video prediction models, thus comparing them under a standardized benchmark. We use the RoboDesk environment in VP^2 for evaluation, which includes 7 kinds of control tasks. The agent's success rate to complete the tasks is used as the evaluation metric.

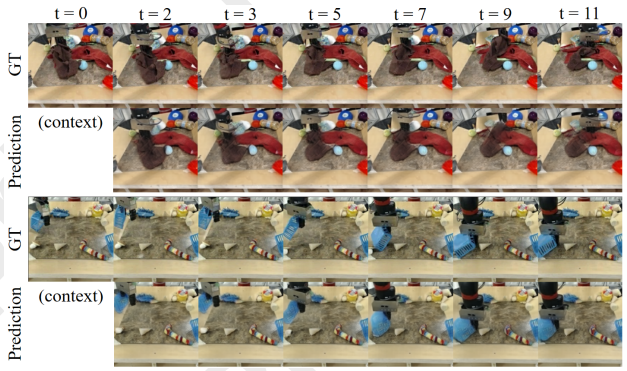


Figure 4: Video Prediction on RoboNet.

Result. In Figure 5, we present a comparison between TDGNet and a set of baseline models [Wu *et al.*, 2024; Babaeizadeh *et al.*, 2021; Villegas *et al.*, 2019; Voleti *et al.*, 2022; Minderer *et al.*, 2020; Gupta *et al.*, 2022]. 'Simulator' is the success rate of the agent when it directly use the simulator as the dynamics, representing an upper bound of model performance. TDGNet attains performance that is comparable or superior to existing SOTA models across the majority of tasks. In addition, we compute the average success rate, normalized by the simulator's performance, demonstrating that TDGNet outperforms existing SOTA models.

4.3 Ablative Experiments

We conduct ablative experiments on the components of TDGNet during training and inference, including the L1 and LPIPS loss used for reconstruction, as well as the top-down guidance (TDG) and conflict detection (CD) performed by the top-down pathway during training and inference. According to the result in Table 5, the LPIPS loss improves the performance of baseline models trained only with L1 loss, particularly for MOVi-C, with improvements of 5.7, 10.4, and 11.8, respectively, in terms of ARI-FG, mBO, and mIoU. The top-down pathway further provides considerable performance gains. TDG and CD bring about improvements of 3.9, 2.4, and 2.9 on CLEVRTex, as well as 9.9, 10.3, and 10.7 on MOVi-C. The top-down pathway offers greater improvements on more complex multi-object scenes (i.e., MOVi-C), indicating that the top-down pathway in TDGNet tends to help the model overcome more challenging samples.

4.4 Analysis

Low-level Feature Optimization with TDG

We visualize the internal features of models in Figure 6(a) by taking the low-level features from TDGNet and the baseline model without TDG (i.e., BO-QSA [Jia *et al.*, 2023]), using

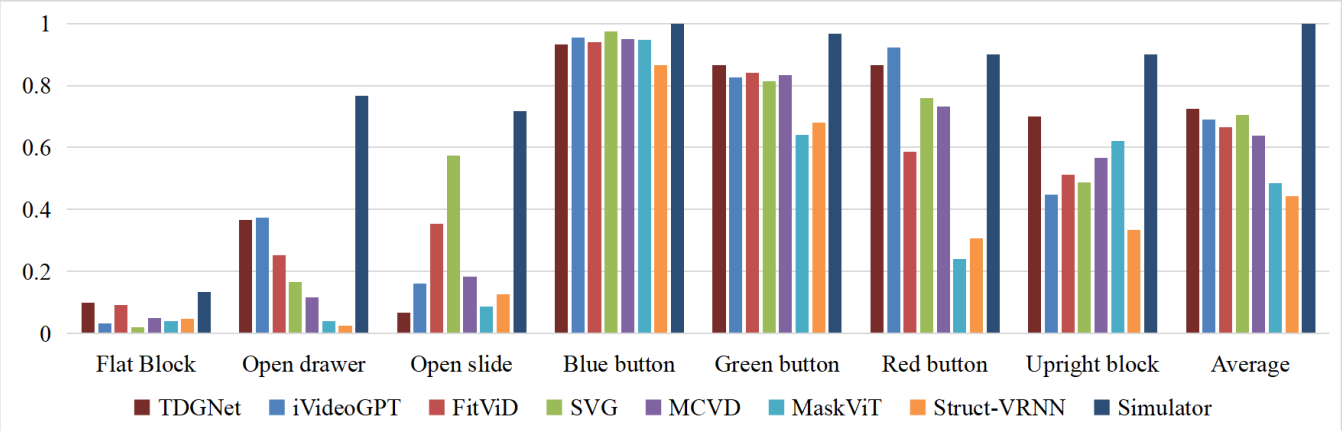


Figure 5: Visual planning on the VP^2 benchmark. On the right, we show the mean scores for each model averaged across all tasks, normalized by the performance of the simulator.

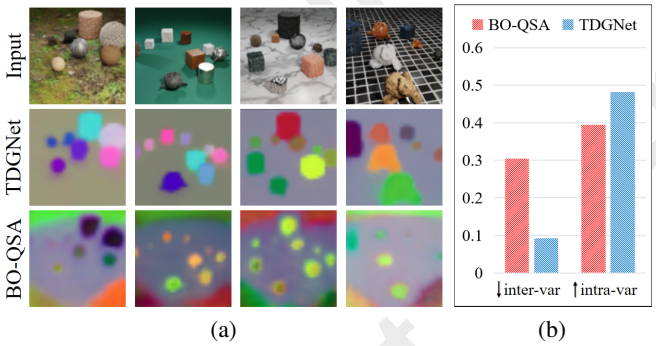


Figure 6: **(a)** Feature visualization on CLEVRTex. We use PCA to reduce dimensions to 3. **(b)** Comparison of object feature variance between TDGNet and BO-QSA. TDGNet has smaller intra- and larger inter-object feature variance.

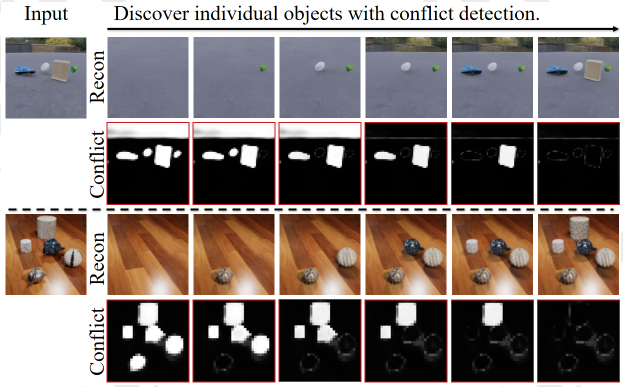


Figure 7: Refinement process of Conflict Detection. We remove all slots except for the background and gradually incorporate objects into the network through CD, and eventually discover all the objects.

PCA [Abdi and Williams, 2010] to reduce the feature dimension to 3 for visualization. Feature maps from BO-QSA are fuzzy. Some of the small objects are almost unseen in the feature map. By contrast, the features extracted by TDGNet are more conducive to identifying individual objects. The features of the same objects are highly similar, while definite boundaries separate adjacent objects. Further in Figure 6(b), we calculate the intra- and inter-object feature variances, which confirms that TDGNet achieves a higher inter-object feature variance and a much lower intra-object feature variance, making objects more distinguishable in the low-level features.

Iterative Refinement with CD

Compared with extracting objects directly through attention competition, the process of detecting and resolving conflicts provides a more explainable way of object discovery. We visualize the process of CD in Figure 7. The model repeatedly computes conflicts between low-level features and slots and includes features with large conflicts into the slot set. To clearly illustrate this process, we manually corrupt the bottom-up pathway, assuming that no object is detected and only the background is given at first. In this extreme case, CD repeatedly calculates the conflicts and includes objects. After multiple iterations, all conflicts are solved, representing that all objects are included.

5 Conclusion

Observing that OCL models provide suboptimal object-centric representations such as missing objects or splitting objects into parts, we analyze existing models and propose that OCL models trained solely with reconstruction loss cannot learn distinguishable low-level features. To address this issue, we refer to human vision and propose TDGNet, which introduces a top-down pathway that guides low-level features with the high-level representations (i.e., slots). We verify that the top-down pathway makes objects more distinguishable in the low-level features. Our experiment results show that TDGNet outperforms existing SOTA models in the object discovery task while also demonstrating its potential for downstream tasks such as prediction and planning in the robotics field.

Acknowledgments

This work was supported by Chinese National Natural Science Foundation Projects U23B2054, 62276254, 62206280, 62376265, the Beijing Science and Technology Plan Project Z231100005923033, Beijing Natural Science Foundation L221013, L242092, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA0480103, the Youth Innovation Promotion Association CAS Y2021131 and InnoHK program.

References

[Abdi and Williams, 2010] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[Ahissar and Hochstein, 2004] Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–464, 2004.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[Babaeizadeh *et al.*, 2021] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction, 2021.

[Beck and Kastner, 2009] Diane M Beck and Sabine Kastner. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, 49(10):1154–1165, 2009.

[Biza *et al.*, 2023] Ondrej Biza, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin F. Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames, 2023.

[Burgess *et al.*, 2019] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[Chang *et al.*, 2022] Michael Chang, Thomas L. Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation, 2022.

[Dasari *et al.*, 2020] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning, 2020.

[Downs *et al.*, 2022] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022.

[Engelcke *et al.*, 2019] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations, 2019.

[Greff *et al.*, 2019] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pages 2424–2433. PMLR, 2019.

[Greff *et al.*, 2022] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti, Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator, 2022.

[Gupta *et al.*, 2022] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction, 2022.

[Hochstein and Ahissar, 2002] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

[Jia *et al.*, 2023] Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[Jiang *et al.*, 2023] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion, 2023.

[Kahneman *et al.*, 1992] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992.

[Kakogeorgiou *et al.*, 2024] Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. Spot: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22776–22786, June 2024.

[Karazija *et al.*, 2021] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation, 2021.

[Karimi-Rouzbahani *et al.*, 2017] Hamid Karimi-Rouzbahani, Nasour Bagheri, and Reza Ebrahimpour. Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. *Scientific reports*, 7(1):14402, 2017.

[Liang and Hu, 2015] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3367–3375, 2015.

[Liu *et al.*, 2024] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2(1):16, 2024.

[Locatello *et al.*, 2020] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention, 2020.

[Minderer *et al.*, 2020] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos, 2020.

[Ramanishka *et al.*, 2017] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7206–7215, 2017.

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

[Seitzer *et al.*, 2022] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

[Singh *et al.*, 2021] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. *arXiv preprint arXiv:2110.11405*, 2021.

[Song *et al.*, 2024] Yeon-Ji Song, Suhyung Choi, Jaein Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. Unsupervised dynamics prediction with object-centric kinematics. *arXiv preprint arXiv:2404.18423*, 2024.

[Tian *et al.*, 2023] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction, 2023.

[Villar-Corrales *et al.*, 2023] Angel Villar-Corrales, Ismail Wahdan, and Sven Behnke. Object-centric video prediction via decoupling of object dynamics and interactions. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 570–574. IEEE, 2023.

[Villegas *et al.*, 2019] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks, 2019.

[Voleti *et al.*, 2022] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation, 2022.

[Wen *et al.*, 2018] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network for object recognition. In *International conference on machine learning*, pages 5266–5275. PMLR, 2018.

[Wolfe, 2021] Jeremy M Wolfe. Is guided search 6.0 compatible with reverse hierarchy theory. *Journal of Vision*, 21(9):36–36, 2021.

[Wu *et al.*, 2023] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models, 2023.

[Wu *et al.*, 2024] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models, 2024.

[Yin *et al.*, 2022] Zhaoyuan Yin, Pichao Wang, Fan Wang, Xianzhe Xu, Hanling Zhang, Hao Li, and Rong Jin. Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 73–89. Springer, 2022.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.