ANALOGICAL NETWORKS FOR MEMORY-MODULATED 3D PARSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent breakthroughs in the applications of deep neural networks in visual perception, one setting that presents a persistent challenge is that of "few-shot learning." Works in the area of few shot visual learning mostly address the task of coarse image classification. Fine-grain visual parsing is necessary for scene understanding and action recognition. Thus far, a separate neural model is trained to parse each semantic category, which hinders knowledge sharing across objects, let alone few shot visual parsing. We present Analogical Networks, a model that casts fine-grained visual parsing into analogical inference: instead of mapping input scenes to part labels, which is hard to adapt in a few-shot manner to novel inputs, our model retrieves related scenes from memory and their corresponding part structures, and predicts **analogous** part structures in the input scene, via an end-to-end learnable modulation mechanism. By conditioning on more than one memory, compositions of structures are predicted, that mix and match parts from different visual experiences. We show Analogical Networks excel at few-shot learning, where instances of novel object categories are successfully parsed simply by expanding the model's memory, without any weight updates. Analogical Networks outperform existing state-of-the-art detection transformer models at part segmentation, as well as paradigms of meta-learning and few-shot learning. We show part correspondences emerge across memory and input scenes by simply training for a label-free segmentation objective, as a byproduct of the analogical inductive bias.

> Ask not what it is, ask what it is like.
>
> *Moshe Bar*

1 INTRODUCTION

The dominant paradigm in existing deep visual learning is to train high capacity networks that map input visual observations to task-specific output labels. Despite their success across a plethora of tasks, these models struggle to perform well in *few-shot* settings where only a small set of examples are available for learning. Meta-learning approaches provide one promising solution to this setting by enabling efficient task-specific adaptation of generic models, but this specialization comes at the cost of poor performance on the original tasks as well as the need to adapt separate models for each novel task considered. In this work, we introduce an alternate approach that unifies inference in both, few-shot and many-shot settings. Instead of pursuing prediction of task-specific outputs, we formulate an approach for analogy-driven prediction, where the desired prediction can be dynamically specified at inference given labelled examples from memory.

In his seminal work (Bar, 2007), Moshe Bar argued for the importance of analogies and associations in human reasoning—highlighting how associations of novel inputs to analogous representations in memory can drive perceptual inference. In this work, we operationalize these insights and propose Analogical Networks, an analogical learning framework for 3D visual parsing. Our approach encodes domain knowledge explicitly in a collection of structured parsed scene memories as well as implicitly, in model parameters. Given an input 3D shape, the model retrieves relevant memories on-the-fly and uses them to modulate inference and segment object parts in the input point cloud. The modulation mechanism operates as top-down feedback for the model to focus on entities present in the retrieved

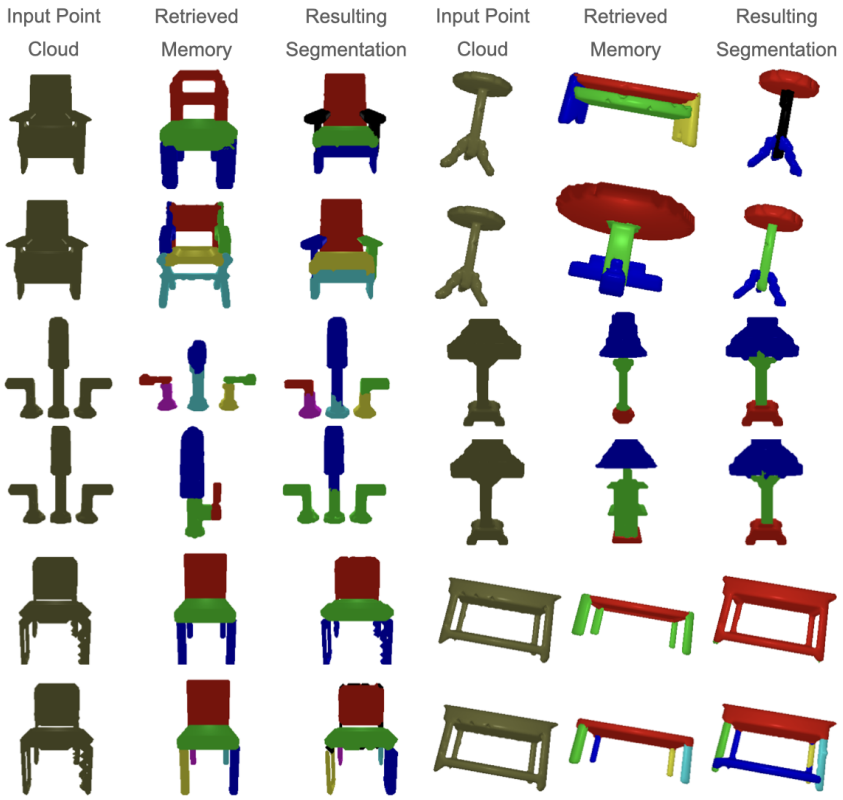

Figure 1: **Cross-object part correspondences emerge in Analogical Networks for 3D object segmentation**, without part association or semantic part labelling supervision. One model is trained across all object categories and segmentation granularities of the PartNet benchmark (Mo et al., 2019). Retrieved memories (2nd and 5th columns) modulate segmentation of the input 3D point cloud (1st and 4th columns, respectively). Contextualized memory part embeddings decode analogous parts in the input scene. We indicate corresponding parts between the memory and the input scene with the same color. Parts shown in black in columns 3, 6 are decoded from scene-agnostic queries and thus they are not in correspondence to any parts of the memory scene. *Rows 5, 6:* Conditioning the same input point cloud on memories with finer or coarser segmentations results in segmentation of analogous granularity.

set of memories, and infer the *analogous* parts in the input scene. During modulation, the input scene and the retrieved memories are jointly encoded and contextualized via cross-attention operations. Each memory is represented as a spatial graph of point and part feature embeddings. The memory part features are then used to decode *corresponding* parts in the input 3D scene. Instead of mapping the input point cloud to semantic class labels, the model reasons analogically, and maps the input to modifications and compositions of past visual experiences, without any explicit semantic label prediction step. Memory to input scene part associations emerge in our model without association supervision, simply as a byproduct of the analogical inductive biases of the modulation process, as shown in Figure 1.

In Analogical Networks, the encoder and modulator parameters are trained end-to-end for the visual parsing task. However, the retrieval process is not end-to-end differentiable. We devise a novel pre-training scheme where we train the parameters of the encoder and modulator sub-modules to parse an augmented version of a 3D scene given the original, un-augmented scene as the modulating memory, bypassing the retrieval process. We show this pre-training helps performance of our model, particularly in the few shot setting.

We test our model on the 3D object segmentation PartNet benchmark (Mo et al., 2019). We compare against state-of-the-art (SOTA) 3D object segmentors, as well as meta-learning and few-shot learning (Snell et al., 2017) baselines adapted for the task of 3D object parsing. Analogical Networks

outperform the baselines in the standard many shot train-test split, and particularly shine over the baselines in the few-shot setting: simply by expanding the memory repository with encodings of a few exemplars, and even without any weight updates, Analogical Networks segment novel instances much better than the baselines. We extensively ablate design choices of our model to quantify contributions of the modulation architecture, the pre-training scheme, and the memory retrieval.

In summary, our contributions are as follows:

1. **A mechanism for in-context memory-modulated learning.** We introduce a model that given a set of relevant examples predicts analogical part structures in the present input scene. The memory context is a natural way to communicate the task to the model, such as the segmentation granularity or object category. Analogical Networks cast every task as an analogical correspondence problem to an appropriate set of memories and their labels. This allows us to train **one model across all tasks**, where each inference is modulated with a different set of relevant memories, of appropriate category and segmentation granularity. In contrast, existing neural network models that are trained to map input to output (Carion et al., 2020; Wu et al., 2019; Yi et al., 2018b; Deng et al., 2019; Yang et al., 2019) break the general task of point cloud segmentation into sub-tasks, one per object category (e.g., chair, table, keyboard, etc.) and part label granularity (levels 1, 2, 3 in the PartNet benchmark (Mo et al., 2019)) and a different neural network is trained in each sub-task.
2. **Few-shot transfer via analogical prediction.** In parametric learning by gradient descent, any change to neural parameters impacts all aspects of agent's future behavior, leading to catastrophic forgetting (Kaushik et al., 2021). Analogical Networks learn from a few examples through memory expansion, without necessarily updating the weights of the networks, which resolves the interference problems of parametric models (Greff et al., 2020).
3. **Emergent correspondences without semantic supervision.** Correspondences emerge across exemplars through the proposed modulation mechanism that uses memory part features as expert queries during part detection, in place of learnable scene-agnostic queries used in SOTA transformer visual detectors (Carion et al., 2020).

Our code will be publicly available upon acceptance.

2 Related work

Frames, Schemas, and neural-symbolic models Analogical Networks represent knowledge explicitly, in terms of structured visual memories, where each one is a graph of part entity neural embeddings. A structured visual memory can be considered the neural equivalent of the earlier concept of a FRAME introduced in (Minsky, 1985), "a graph of nodes and their relations for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party", to quote Minsky (1985). FRAME nodes would operate as "slots" to be filled with specific entities, or symbols, in the visual scene. Symbol detection would be carried out by a separate state estimation process such as, general-purpose object detectors (Zhou et al., 2022), employed also by recent neuro-symbolic models (Yi et al., 2018a; Mao et al., 2019; Yi et al., 2019). In the wild detection of domain symbols (e.g., chair handles, table legs, faucet tips, fridge doors) typically fails, which is precisely the reason why these earlier symbolic systems of knowledge, largely disconnected from the sensory input, have not been widely adopted. Analogical Networks take a step towards addressing these shortcomings by including symbol detection as part of the FRAME inference itself, through a top-down modulation that uses the context represented in the FRAME memory graph, to jointly search for multiple entities and localize them in context of one another.

Few-shot Prediction: Meta-learning and Learning associations A key goal for our approach is to enable accurate inference in few-shot settings. Previous approaches (Finn et al., 2017; Rusu et al., 2018; Snell et al., 2017; Wang et al., 2018; Bar et al., 2022; Mangla et al., 2020; Wang et al., 2020; Nguyen & Todorovic, 2019; Liu et al., 2020; Tian et al., 2020; Yang et al., 2020) that target similar settings can be broadly categorized as relying on either meta-learning or learning better associations. Meta-learning approaches tackle few-shot prediction by learning a generic model that can be efficiently adapted to a new task of interest from a few labelled examples. While broadly

applicable, these methods result in catastrophic forgetting of the original task due to adaptation and thus require training a new model for each task of interest. Moreover, the goal of learning generic and rapidly adaptable models can lead to suboptimal performance over the base tasks with abundant data. An alternate approach for the few-shot setting is to learn better associations. For example, the category of a new example maybe inferred by transferring the label(s) from the one (or few) closest samples (Snell et al., 2017; Sung et al., 2018). While this approach obviates the need for adapting models and can allow prediction in few-shot and many-shot settings, the current approaches are only applicable to global prediction, e.g., image labels. Our work can be viewed as extending such association-based methods to allow predicting fine-grained and generic visual structures using our proposed modulation-based prediction mechanism.

Memory-augmented neural networks Memory-augmented neural networks (Santoro et al., 2016) are equipped with explicit external key-value memories, which they can attend on and retrieve key-value pairs relevant to the input. Memory-augmentation of parametric models permit **fast learning** with few examples, where the data are saved in the explicit memory immediately after their acquisition, while learning via parameter update is slow and requires multiple gradient iteration on de-correlated examples. Another benefit of explicit memories over implicit memories of model parameters is that computation only scales sub-linearly if approximate nearest neighbors schemes are used, while it scales quadratically with the number of model parameters. External memories have recently been used in scaling up language models (Borgeaud et al., 2021; Khandelwal et al., 2019) to alleviate from the limited context window of parametric transformers (Wu et al., 2022), to store knowledge in the form of entity mentions (de Jong et al., 2021), knowledge graphs (Das et al., 2022), and question answer pairs (Chen et al., 2022). Memory attention layers in these models are used to influence the computation of transformer layers, and have proven critical for factual question answering, but also for sentence completion over their parametric counterparts.

3 Analogical Networks for 3D Object Parsing

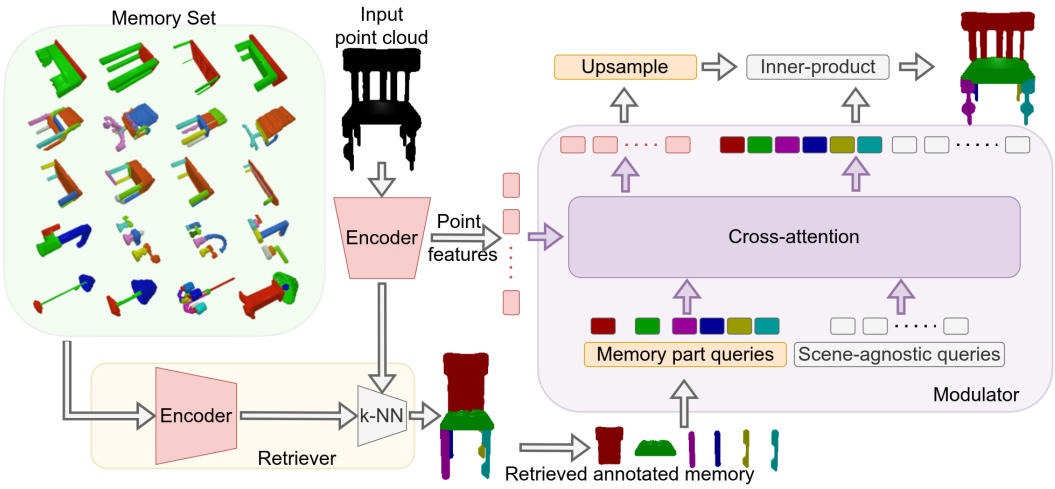

Figure 2: **Architecture for Analogical Networks.** Analogical Networks are comprised of an encoder, retriever and modulator sub-modules. Labelled memories and the (unlabelled) input point cloud are first separately encoded into feature embeddings and the top-k most similar memories to the present input are retrieved. Here for clarity of presentation we show the case for $k = 1$. Each retrieved memory part embedding initializes a query that is akin to a slot to be "filled" with the analogous part entity in the present scene. These queries are appended to a set of learnable scene-agnostic queries. The modulator contextualizes the queries with the input point cloud through iterative self and cross-attention operations that also update the point features of the input. When a memory part query decodes a part in the input point cloud, we say the two parts are put into analogical correspondence by the model. We color them with the same color to visually indicate this correspondence.

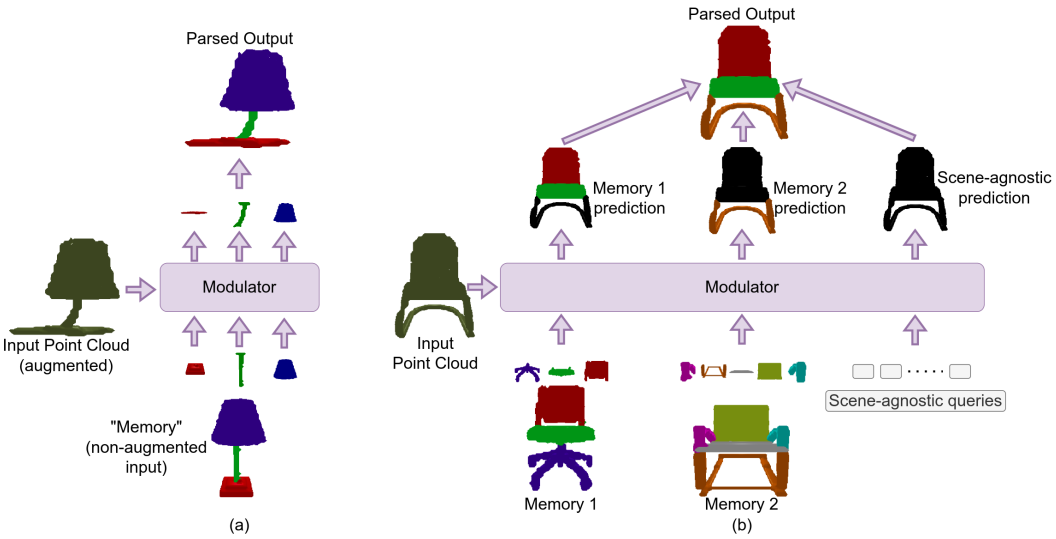

Figure 3: *Left:* **Within-instance correspondence pre-training.** *Right:* **Multi-memory Analogical Networks .**

The architecture of Analogical Networks for 3D part segmentation is illustrated in Figure 2. It is comprised of an encoder, a retriever and a modulator sub-network, and a collection of labelled object 3D point clouds of different categories and segmentation granularities as the memory set. The same object with finer or coarser segmentation labelling is represented in different memories.

In a nutshell, Analogical Networks work as follows: First, the encoder encodes the input point cloud to a set of point features, and each memory point cloud to point and part features. The retriever takes as input an 1D encoding of the input 3D point cloud and retrieves similar labelled objects from memory. The retrieved memory parts initialize queries that are tasked to decode the parts of the input scene. The modulator uses multiple iterations of self and cross-attention operations to jointly contextualize the encodings of the input point cloud and the retrieved memory part queries. Queries that are found confident decode the "analogous" parts in the present 3D point cloud and a Hungarian loss is computed between predicted and ground-truth parts, similar to detection transformers (Carion et al., 2020). When a memory part query is used to decode an analogous part in the input point cloud, we say the two parts are put in correspondence by the model. The colors in columns 2, 3 and 5, 6 of Figure 2 denote part correspondence between the retrieved memory parts and the input point cloud.

Encoder The encoder is a PointNet++ (Qi et al., 2017) and encodes the point cloud scene S and each memory scene M to a set of 3D point features $F^S, F^M \in \mathbb{R}^{N \times C}$ (where N is the number of sub-sampled points and C is the feature dimension). 3D positional encodings are then added to the 3D point features. Each labelled part p of a memory scene M is then encoded into a 1D feature vector f_p^M by average pooling of its point features. For retrieval, each scene (memory or the input) is summarized in an 1D vector by global average pooling.

Retriever The retriever retrieves top-k memories for the input scene S by computing an inner product between input scene query and memory key features.

Modulator The modulator sub-network resembles the decoder of a detection transformer (Carion et al., 2020) with two key differences: First, instead of only using a set of scene agnostic learnable query vectors (64 in our case), it also considers the retrieved memory part features f_p^M (where $M = 1..k$) as queries, i.e. candidates for decoding parts in the input point cloud scene. Second, during the iterative self and cross-attention layers, both the queries and the point features of the input scene are updated, while in DETR (Carion et al., 2020) only the queries are updated. We consider 6 layers of self and cross-attention. More details and pseudocode are provided in the supplementary.

Lastly, we upsample the point features to the original resolution (from N to N_p) using convolutional layers (Qi et al., 2017) and compute an inner product between each query and point feature to compute

the segmentation mask for each query. The output of the modulator is a set of N_q segmentation mask proposals and corresponding confidence scores, where N_q is the total number of queries. At training time, these proposals are matched to ground-truth instance binary masks using the Hungarian matching algorithm (Carion et al., 2020). For the proposals that are matched to a ground-truth instance, we compute the segmentation loss, which is a per-point binary cross-entropy loss (Vu et al., 2022; Cheng et al., 2021). We also supervise the confidence score of each query, similar to (Carion et al., 2020). The target labels are 1 for the proposals matched with a ground-truth part and 0 for non-matched. We found it beneficial to apply these losses after every cross-attention layer in the modulator. At test time, we multiply the per point mask occupancy probability with tiled confidence scores to get a $N_p \times N_q$ tensor (N_p is the number of points); then each point is assigned to the most confident query by computing inner product and taking per-point argmax over the queries.

Within-instance correspondence pre-training In Analogical Networks, the retrieval process is not end-to-end differentiable with respect to the downstream scene parsing task. This is the case because retrieved memories are contextualized with the input scene through dense cross-attention operations, instead of bi-encoder fusion (Miech et al., 2021); the former, though a stronger performer than bi-encoder models, it is not differentiable with respect to the memory feature vectors used by the retriever, as also explained in (Chen et al., 2022).

We devise a novel pre-training scheme where we pre-train the encoder and modulator parameters independent of the retriever, by augmenting (rotating and deforming (Kim et al., 2021)) each training 3D point cloud, and using the original point cloud as the modulating memory, as shown in the left side of Figure 3. In this special case, we actually have ground-truth information regarding part association between the original and the augmented point cloud. We thus supervise each memory part query to decode the corresponding deformed part in the input cloud and do not use Hungarian matching. We do not have such part association ground-truth during regular (cross-instance) training.

Upon within-instance pre-training, we train our model using as modulating memory set sampled from the retriever's top K memories. During regular training (i.e. cross-instance training), we further co-train with our within instance training data. We found that in this second training stage using the part association supervision in within-instance examples does not further contribute to performance, so we supervise with our standard losses and Hungarian matching.

4 Experiments

We test Analogical Networks on PartNet (Mo et al., 2019), an established benchmark for 3D object segmentation. PartNet contains multiple 3D object instances from 24 object categories, each segmented into three different segmentation granularities. We consider two experimental paradigms:

- **Many-shot:** In this setup, we split the exemplars of each of (base) categories into train and test sets. The model is tested on segmenting instances of the same categories in the test set. Our base categories are Chair, Display, StorageFurniture, Bottle, Clock, Door, Earphone, Faucet, Knife, Lamp, TrashCan, Vase.
- **Few-shot:** In this setup, K examples of novel categories are shown to the model, and the model is tasked to segment new examples of these categories. This tests the ability of the model for few shot adaptation. We consider $K = 1$ and $K = 5$ setups. Our novel categories are Table, Bed, Dishwasher, Refrigerator.

Our experiments aim to answer the following questions:

1. How do Analogical Networks compare to parametric SOTA neural networks and meta-learning networks for 3D object parsing in the many-shot and few-shot paradigms?
2. How well Analogical Networks learn part-based associations across exemplars without part association or part semantic label supervision?
3. How much weight updating over memory updating help few shot adaptation in Analogical Networks?
4. How much within-instance correspondence pretraining helps Analogical Networks?

Method	Fine-tuned?	Novel Categories		Base Categories
		1-shot ARI (↑)	5-shot ARI (↑)	Many shot ARI (↑)
3D-DETR	✗	29.2 ± 0.82	29.2 ± 0.82	70.5
	✓	34.3 ± 2.17	48.6 ± 2.46	-
Analogical Networks w/o memory part queries	✗	48.3 ± 2.07	48.9 ± 1.23	71.3
	✓	51.2 ± 2.64	56.6 ± 2.23	-
Analogical Networks single memory w/o pretrain	✗	36.8 ± 0.92	36.3 ± 1.60	71.8
	✓	46.1 ± 3.61	51.3 ± 1.32	-
Analogical Networks single memory (ours)	✗	53.0 ± 1.26	54.2 ± 1.36	**71.9**
	✓	53.8 ± 1.49	56.3 ± 1.59	-
Analogical Networks multi-memory (ours)	✗	$\mathbf{55.8 \pm 1.35}$	56.4 ± 1.30	70.6
	✓	-	$\mathbf{58.4 \pm 1.52}$	-

Table 1: **Results on few-shot and many-shot 3D object segmentation on PartNet (Mo et al., 2019).** Our few shot experiments use four held-out (novel) categories which are Table, Bed, Dishwasher, Refrigerator. We report mean and standard deviation for few-shot ARI performance over 10 tasks (each task consists of a different set of K-shot support set). **Without any finetuning, Analogical Networks outperform the baseline by 26% in the few shot setup.** Though weight finetuning helps both models, it brings a performance boost of 18% for the baseline and only 2% for our model. This means Analogical Networks can generalize and adapt few shot to new categories by expanding their memory, without any weight interference. Even upon finetuning, Analogical Networks outperform the 3D-DETR baseline by 10 % ARI.

5. How the proposed query-centric modulation compares against more conventional memory-augmented neural networks (Wu et al., 2022) where memories are attended on but not used directly as decoding slot queries?

6. How much the performance of Analogical Networks varies with different retrieval mechanisms?

Evaluation metrics We use Adjusted Random Index (ARI) as our evaluation metric for segmentation accuracy (Rand, 1971), which is a category label-agnostic clustering score (1 being the best and -1 being the worst). ARI calculates the similarity between two point clusters while being invariant to the ordering of the cluster centers. We report $100\times$ ARI as our metric. For this, we use the publicly available implementation of (Kabra et al., 2019).

Baselines We compare our model to the following models in the literature: *3D-DETR* is a 3D detection transformer neural network with the same backbone as our model, and similar segmentation prediction head and loss, but without any memory retrieval, akin to the 3D equivalent of a state-of-the-art 2D image segmentor (Carion et al., 2020; Athar et al., 2021). Note that point features are updated during the decoder layers, same as our model, which we found to help performance. *Prototypical Networks* is an adaptation of the episodic prototypical networks for image classification of (Snell et al., 2017) to the task of 3D object part segmentation. Specifically, given a set of N labelled point clouds, we form average feature vectors for each semantic labelled part and use them as queries to segment points into corresponding part masks. Contrary to 3D-DETR , Analogical Networks attend on external memories. Contrary to Prototypical Networks , Analogical Networks consider feature contextualization between memory prototypes and the input scene.

We further consider the following ablated versions of our model, each one aims to evaluate a specific aspect of its design: *Analogical Networks w/o memory part queries*, which is a model similar to the parametric 3D-DETR baseline augmented with attention to retrieved memories (Wu et al., 2022). Instead of using the part memory features as queries, this baseline updates the scene-agnostic query vectors by iteratively attending to the input feature cloud and the memory part features. Different from Analogical Networks, correspondence cannot emerge between a memory and the input scene, since the model does not encode such correspondence explicitly, but only implicitly, in the attention operations. *Analogical Networks single memory*, which is Analogical Networks with a single modulating memory. *Analogical Networks w/o pretrain*, which is our model without the self-supervised within-instance correspondence objective for training of the modulator and encoder weights.

4.1 MANY AND FEW SHOT 3D OBJECT PART SEGMENTATION

The PartNet benchmark supplies three levels of segmentation per object instance where level 3 is the most fine-grained. We train and test our model and baselines on all three levels of segmentation. We use a learnable level embedding concatenated to the input for our baseline 3D-DETR , as is usually the case in multi-task models (Jang et al., 2022). In the many shot setting, all examples in the training set become part of our model's memory. In the few shot setting, Analogical Networks adapt in two ways i) by expanding the memory of the model with the novel K-shot support examples, and ii) by further adapting the weights via finetuning to parse the K examples using both within-instance pre-training and cross-instance training. In this case, the memory set is only the novel labelled support set instances.

We show quantitative results in Table 1. Our conclusions are as follows: **(i) Analogical Networks dramatically outperform parametric DETR segmentors in few-shot learning**. While in the base category test set the two models have similar performance (70.5 versus 70.6), when adapting few-shot to the novel categories, Analogical Networks dramatically outperforms 3D-DETR (48.6 versus 58.4 in 5-shot ARI). **(ii) Analogical networks can adapt few-shot to novel categories without weight updates, simply by memory expansion**. Indeed, the 5-shot performance of our multi-memory model is very close before and after finetuning in the novel categories. **(iii) Multi-memory Analogical Networks outperform single memory Analogical Networks in few-shot adaptation**. **(iv) Within-instance correspondence pre-training helps Analogical Networks to adapt few-shot to the novel categories. (v) Attending on retrieved memory part features and contextualizing them with the input, as done in Analogical Networks w/o memory part queries, helps over parametric alone segmentors** in both many and few-shot settings. Our interpretation for this interesting result is that attending on relevant memories permits a model to operate in context and overcome challenges caused by imbalance of examples across categories. Rare instances are not forgotten as they are explicitly saved in the external memory, and upon retrieval and modulation with the input, can steer inference. In parametric alone networks, rare examples are in competition with examples in the mode of the training distribution for their representation in the parameters. We show per-class ARI scores in the supplementary. Still, this memory-augmented model exhibits much worse performance than our complete model without weight finetuning (49% versus 56.4 % ARI). Please, see our supplementary file for performance of our model under varying retrieval schemes.

4.2 EVALUATION OF EMERGENT PART CORRESPONDENCES

In this section, we evaluate the quality of cross-exemplar part correspondences that emerge in Analogical Networks . Our model is trained solely for segmentation of 3D objects, and no semantic part labels are used during training. The only cross-part association (correspondence) supervision is during pre-training, where the parts in a 3D point cloud are used as queries to decode (upon modulation) the corresponding part of the augmented 3D point cloud. Yet, through the use of memory part queries shown in Figure 2, our model can "label" input scenes with the labels of the corresponding memory parts. If the parts of visual memories are semantically labelled, those semantic labels will propagate to the input scene. Parts decoded from scene agnostic queries are not put in correspondence with any memory in our model. We found 80% of 3D points to be on parts decoded by memory part queries on average.

Here we evaluate the quality of label propagation through emergent correspondences by evaluating part instance segmentation and semantic segmentation in the few shot setting. We compare against *3D-DETR semantic* which is precisely our 3D-DETR baseline additionally supervised for classifying detected 3D parts into semantic part labels of the category considered in each case. Our baseline *Prototypical Networks* , similar to Analogical Networks, propagates semantic labels of the prototypical part features. We evaluate Prototypical Networks only for semantic segmentation since it cannot easily produce instance segments: if multiple part instances belong to the same label, e.g., chair leg, Prototypical Networks assume they belong to the same semantic prototypical part. We further train a version of Analogical Networks where only the scene agnostic queries are explicitly supervised for semantic part labels; we call it Analogical Networks semantic .

Evaluation metric We evaluate the part instance segmentation using mean Panoptic Quality (PQ) as our metric (Kirillov et al., 2019). We evaluate the 3D point semantic segmentation using mean intersection over union (mIoU) as our metric.

We show quantitative semantic and instance segmentation results in Table 2. Analogical Networks show very high semantic and instance segmentation accuracy via label propagation through memory part queries, **without any supervision.** We report two sets of numbers for our model. The first penalizes points that belong on parts decoded by scene agnostic queries and assumes they have wrong labels, the second only evaluates segmentation scores on the 3D points on parts decoded by part memory queries.

Method	Novel Category 1-shot		Novel Category 5-shot		Base Category Many-shot	
	mIoU	PQ	mIoU	PQ	mIoU	PQ
3D-DETR semantic	0.41	0.33	0.47	0.34	0.49	0.45
Prototypical Networks	0.37	-	0.39	-	0.54	-
Analogical Networks (ours)	$0.38/0.41^*$	$0.37/0.40^*$	$0.41/0.45^*$	$0.42/0.46^*$	$0.60/0.63^*$	$0.49/0.54^*$
Analogical Networks semantic (ours)	0.41	0.43	0.48	0.47	0.64	0.51

Table 2: **Part Semantic and Instance Segmentation performance on few-shot and many-shot 3D object segmentation on PartNet dataset (Mo et al., 2019).** Few-shot performance is calculated by averaging over the 10 different $K-$shot tasks (each task consists of a different support set, we observed std of 0.03 for all the reported values). * performance is calculated on the subset of input 3D points that were classified by memory part queries (and not by the scene agnostic queries since in the latter case semantic labels cannot be propagated.)

4.3 Limitations - Future directions

Below we present a set of future directions that are necessary for Analogical Networks to scale beyond segmentation of single object scenes. **(i)** The retriever in Analogical Networks operates currently over whole object memories and it is not end-to-end differentiable with respect to the downstream task. Sub-object part-centric memory representations would permit fine-grained retrieval of visual memory scenes. We further plan to explore alternative supervision for the retriever module inspired by works in the language domain (Izacard et al., 2022; Izacard & Grave, 2021). **(ii)** Scaling Analogical Networks to segmentation of complete, multi-object 3D scenes in realistic home environments requires scaling up the size of memory collection. It would further necessitate bootstrapping fine-grained object part annotations, missing from 3D scene datasets (Dai et al., 2017), by transferring knowledge of object part compositions from PartNet. Such semi-supervised fine-grained scene parsing is an exciting avenue of future work. **(iii)** So far we have assumed the input to Analogical Networks to be a complete 3D point cloud. However, this is hardly ever the case in reality. Humans and machines need to make sense of single view, incomplete and noisy observations. Extending Analogical Networks with generative heads that not only detect analogous parts in the input, but also inpaint or complete missing parts, is a direct avenue for future work.

5 Conclusion

We presented Analogical Networks, a model for associative 3D visual parsing that encodes domain knowledge explicitly, in a set of past labelled visual scenes, and implicitly, in model parameters. Analogical Networks forgo the established paradigm of mapping visual scenes to semantic labels and instead puts forward an analogical paradigm of mapping scenes to past scenes, their parts labels and their compositions. Given an input visual scene, the model retrieves a set of labelled memory point clouds, and contextualized those with the input through self and cross attention layers. The contextualized memory part features decode *corresponding* parts in the input point cloud. The encoder and modulation parameters of the model are trained end-to-end for the segmentation task, and the retriever parameters are trained with supervised contrastive feature learning. By casting visual parsing as analogical correspondence, Analogical Networks can be trained across and share knowledge across multiple tasks, where each task is communicated via the features and labels of appropriate conditioning memories. One shot, few shot or many shot learning are treated in a uniformly, by conditioning to the appropriate set of memories, whether taken from a single, few or many memory exemplars, and inferring analogous parses. We showed Analogical Networks outperform SOTA baselines in both many and few shot parsing setting. We further showed part correspondences emerge across exemplars without supervision, as a by-product of the analogical inductive bias.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the empirical results, we include a pseudo-code of the main model components and training pipeline in the Appendix. Upon publication, we will make our code publicly available on GitHub.

REFERENCES

Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Hodor: High-level object descriptors for object re-segmentation in video learned from static images, 2021. URL https://arxiv.org/abs/2112.09131.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. *NeurIPS*, 2022.

Moshe Bar. The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280–289, 2007. doi: 10.1016/j.tics.2007.05.005.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426, 2021. URL https://arxiv.org/abs/2112.04426.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proc. ECCV*, 2020.

Wenhu Chen, Pat Verga, Michiel de Jong, John Wieting, and William Cohen. Augmenting pre-trained language models with qa-memory for open-domain question answering, 2022. URL https://arxiv.org/abs/2204.04581.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. URL http://arxiv.org/abs/1702.04405. cite arxiv:1702.04405.

Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Robin Jia, Manzil Zaheer, Hannaneh Hajishirzi, and Andrew McCallum. Knowledge base question answering by case-based reasoning over subgraphs, 2022. URL https://arxiv.org/abs/2202.10610.

Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William Cohen. Mention memory: incorporating textual knowledge into transformers through entity mention attention. *CoRR*, abs/2110.06176, 2021. URL https://arxiv.org/abs/2110.06176.

Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey E. Hinton, and Andrea Tagliasacchi. Cvxnets: Learnable convex decomposition. *CoRR*, abs/1909.05736, 2019. URL http://arxiv.org/abs/1909.05736.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Gautier Izacard and Edouard Grave. Distilling Knowledge from Reader to Retriever for Question Answering. In *ICLR 2021 - 9th International Conference on Learning Representations*, Vienna, Austria, May 2021. URL https://hal.archives-ouvertes.fr/hal-03463398.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models, 2022. URL https://arxiv.org/abs/2208.03299.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: zero-shot task generalization with robotic imitation learning. *CoRR*, abs/2202.02005, 2022. URL https://arxiv.org/abs/2202.02005.

Rishabh Kabra, Chris Burgess, Loic Matthey, Raphael Lopez Kaufman, Klaus Greff, Malcolm Reynolds, and Alexander Lerchner. Multi-object datasets. https://github.com/deepmind/multi-object-datasets/, 2019.

Prakhar Kaushik, Alex Gain, Adam Kortylewski, and Alan L. Yuille. Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *CoRR*, abs/2102.11343, 2021. URL https://arxiv.org/abs/2102.11343.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models, 2019. URL https://arxiv.org/abs/1911.00172.

Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J Kim. Point cloud augmentation with weighted local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 548–557, 2021.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pp. 142–158. Springer, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *CoRR*, abs/1904.12584, 2019. URL http://arxiv.org/abs/1904.12584.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. *CoRR*, abs/2103.16553, 2021. URL https://arxiv.org/abs/2103.16553.

M. Minsky. A framework for representing knowledge. In R. J. Brachman and H. J. Levesque (eds.), *Readings in Knowledge Representation*, pp. 245–262. Kaufmann, Los Altos, CA, 1985. URL http://web.media.mit.edu/~minsky/papers/Frames/frames.html.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 622–631, 2019.

Charles Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proc. NIPS*, 2017.

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

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1842–1850. JMLR.org, 2016.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Thang Vu, Kookhoi Kim, Tung Minh Luu, Xuan Thanh Nguyen, and Chang-Dong Yoo. Softgroup for 3d instance segmentation on point clouds. *ArXiv*, abs/2203.01509, 2022.

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *CoRR*, abs/1803.08035, 2018. URL http://arxiv.org/abs/1803.08035.

Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. PQ-NET: A generative part seq2seq network for 3d shapes. *CoRR*, abs/1911.10949, 2019. URL http://arxiv.org/abs/1911.10949.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers, 2022. URL https://arxiv.org/abs/2203.08913.

Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *CoRR*, abs/1906.01140, 2019. URL http://arxiv.org/abs/1906.01140.

Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pp. 763–778. Springer, 2020.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. *CoRR*, abs/1810.02338, 2018a. URL http://arxiv.org/abs/1810.02338.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. *CoRR*, abs/1910.01442, 2019. URL http://arxiv.org/abs/1910.01442.

Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. GSPN: generative shape proposal network for 3d instance segmentation in point cloud. *CoRR*, abs/1812.03320, 2018b. URL http://arxiv.org/abs/1812.03320.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision, 2022. URL https://arxiv.org/abs/2201.02605.

6 APPENDIX

Our Appendix material is organized as follows: In Section 6.1, we provide implementation details such as pseudo code and hyper-parameters for training Analogical Networks. Section 6.2 reports the ARI scores for each category involved in base training phase. In section 6.3 we ablate Analogical Networks performance under varying memory retrieval schemes in 5-shot setting. We provide qualitative visual object parsing results for single and multi memory variants of Analogical Networks in section 6.4.

6.1 TRAINING PSUEDO CODE AND IMPLEMENTATION DETAILS

For both stages of training (i.e. within-instance correspondence pre-training and cross-instance training), we use AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of $1\mathrm{e}{-4}$, l_2 weight decay of $1\mathrm{e}{-2}$ and batch size of 8. We train it for 100 epochs with a learning rate schedule decay of 0.5 at every 30th epoch. For few-shot finetuing/evaluation, we use AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of $1\mathrm{e}{-5}$, l_2 weight decay of $1\mathrm{e}{-2}$ and batch size of 4. We finetune it for 30 epochs and we report the performance across 10 different task runs (each task has a different set of K-shot support samples). We describe our Analogical Networks training details in the following pseudo code:

Algorithm 1: Pseudo code for training our Analogical Networks

```
# pc_encoder: point cloud encoder, S: input point cloud, M: retrieved memory point
     cloud, Np: numbers of points in S or M, N: sub-sampled points, C: number of
    feature channels, P: number of parts in M
# Xp_Hungarian : Hungarian matched label assignment of points in S to queries
# yp_Hungarian : Hungarian matched label assignment of queries to GT parts in S
# part_encoder : Computes the part features using mean pooling
# pos_encode : Adds positional encoding, upsampler : Upsamples point cloud
# Objectness_loss: Binary cross entropy loss to decide which queries (scene-
    agnostic+learnable) would be responsible for decoding parts
# Segmentation_Loss: Cross entropy loss to assign each point to the hungarian
    matched query.

for S,M in dataloader: # load a batch with B samples
   # S : B x Np x 3 and M: B x Np x 3
   # Compute point features
   F^S = pc_encoder(S) # B x N x C
   F^M = pc_encoder(M) # B x N x C

   # Initialize memory part queries
   f^M = part_encoder(F^M) # B x P x C

   # Add positional embedding
   x = pos_encode(F^S)
   y = pos_encode(f^M)

   Loss = 0
   # Do multiple layers of modulation Self-Attn and Cross-Attn
   for layer in num_layers:
      x = Self-Attn(x) # B x N x C
      y = Self-Attn(y) # B x P x C
      x = Cross-Attn(x, y) # B x N x C
      y = Cross-Attn(y, x) # B x P x C

      X = upsampler(x) # B x Np x C
      point_query_score = matmul(normalize(X), normalize(y.T)) # B x N x P

      Loss += Segmentation_Loss(armax(point_query_similarity, -1), Xp_Hungarian) +
           Objectness_loss(y, yp_Hungarian)


   # optimizer step
   loss.backward()
   optimizer.step()
```

Methods	Base Categories (ARI ↑)											
	Chair	Display	Storage Furniture	Bottle	Clock	Door	Ear phone	Faucet	Knife	Lamp	Trash Can	Vase
3D-DETR	76.9	87.4	67.7	79.9	46.8	53.7	75.7	74.1	83.1	79.2	53.6	67.7
Analogical Networks w/o memory part queries	78.9	89.2	71.6	72.3	45.9	53.4	75.9	74.8	84.7	76.8	61.4	70.4
Analogical Networks single memory w/o pretrain	77.8	88.7	71.4	72.6	49.7	51.1	74.8	72.7	84.5	72.7	72.4	72.2
Analogical Networks single memory	79.1	89.1	73.5	69.4	50.7	52.7	74.9	75.9	84.2	78.3	64.4	69.5
Analogical Networks multi-memory	78.2	87.9	74.0	71.2	44.6	56.6	72.5	75.8	84.3	77.3	59.7	64.5

Table 3: Category specific ARI scores for base categories of PartNet dataset (Mo et al., 2019).

6.2 ARI Performance for Base Categories

In this section, we provide category specific ARI scores for different approaches as mentioned in table 3 that belongs to the base category set.

6.3 Performance under varying retrieval schemes.

In few-shot setting, we evaluate the performance of Analogical Networks under varying memory retrieval schemes in Table 4. We compare against a hypothetical *oracle* retriever that can fetch the most helpful (in terms of resulting ARI) memory for each input. Our conclusions are as follows: **(i) Analogical Networks with an oracle memory perform better than Analogical Networks using memories retrieved by the retriever**. This suggests that better training of our retriever or exploring its co-training with the rest of our model would have significant impact in improving its performance. **(ii) Considering any of the 5 shot exemplars randomly does slightly worse than using memories retrieved by our retriever.** Note that while in the few-shot setting we operate in each object category separately for both our model and baselines, in the base category case, our retriever needs to learn to retrieve memories of the correct category, which is critical for the good performance of our model.

Method	Fine-tuned?	Modulating Memory	Novel Category: 5-shot ARI (↑)
Analogical Networks single memory (ours)	✗	Random	52.2 ± 0.78
		Retriever	53.8 ± 1.49
		Oracle	61.9 ± 1.22
Analogical Networks single memory (ours)	✓	Random	53.5 ± 1.28
		Retriever	55.8 ± 1.91
		Oracle	62.3 ± 0.66

Table 4: **Ablations on ARI segmentation performance under varying retrieval schemes for 5-shot on 4 novel categories**.

6.4 Qualitative Parsing results for single-memory and multi-memory Analogical Networks

In this section, we show our model's qualitative results for object parsing. In the figures below (4, 5, 6 and 7), we visualize the figures with the following 5 columns: (a) Unlabelled input point cloud (b) Memory used for modulating the object parsing (c) Object parsing generated using only the memory part queries (not the scene-agnostic queries). In this column, black region of areas denote it would be eventually decoded by scene-agnostic query. (d) Final predicted segmentation parsing uses both memory initialized query and scene-agnostic query. (e) Input point cloud's ground truth segmentation at the segmentation granularity level of the memory. We qualitatively show the emergence of part correspondence between retrieved memory (column 2) and the input point cloud parsed using memory query (column 3). Parts having the same color in column 2 and 3 demonstrates correspondence. Furthermore, we show parsing results for Analogical Networks single memory w/o pretrain. We observe that all of memory part query are inactive in the parsing stage. This demonstrate the utility of within-instance pretraining as without this pretraining part correspondence doesn't emerge as shown in Figure 8.

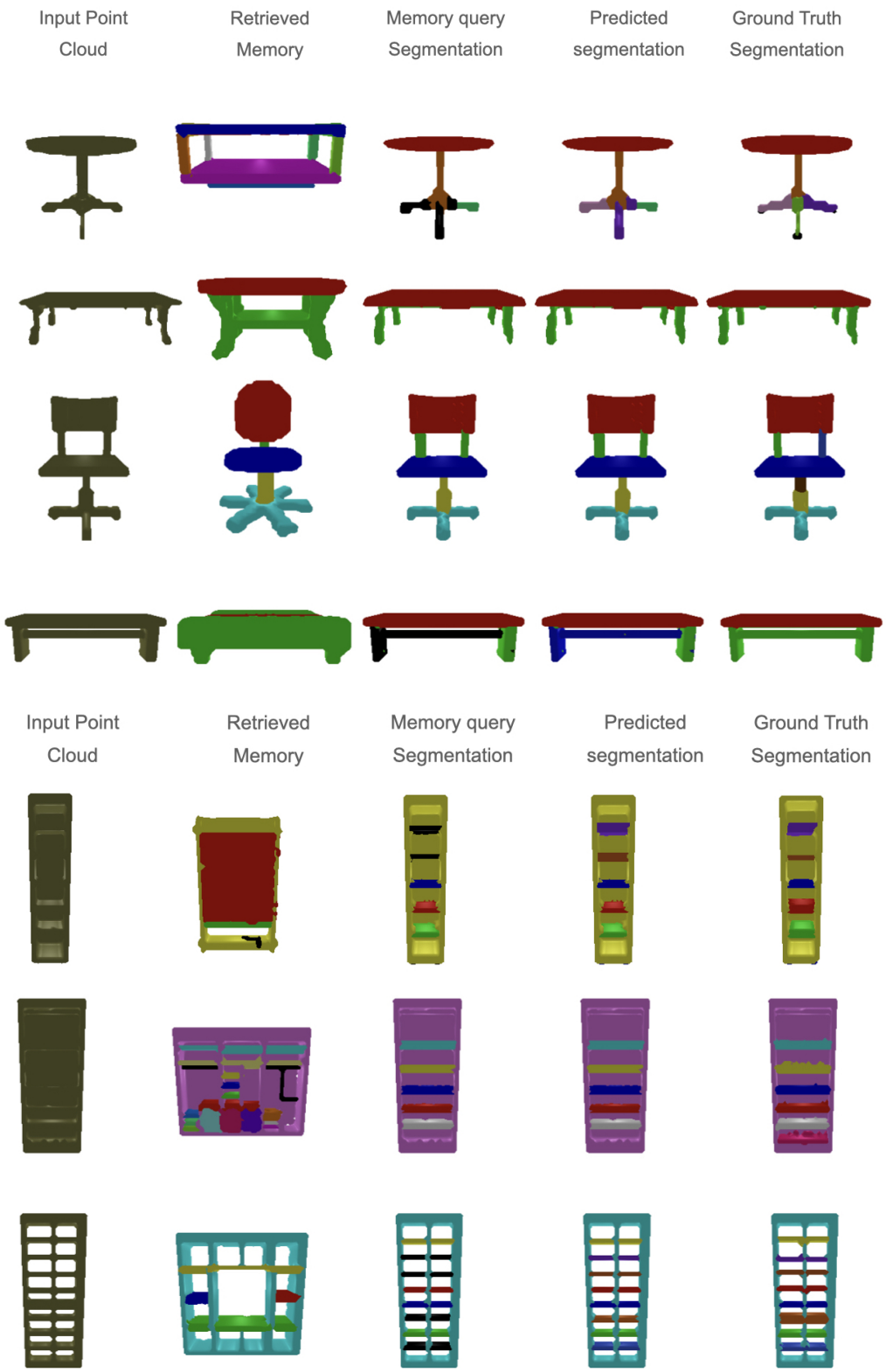

Figure 4: We show some qualitative object parsing results that are predicted by Analogical Networks .

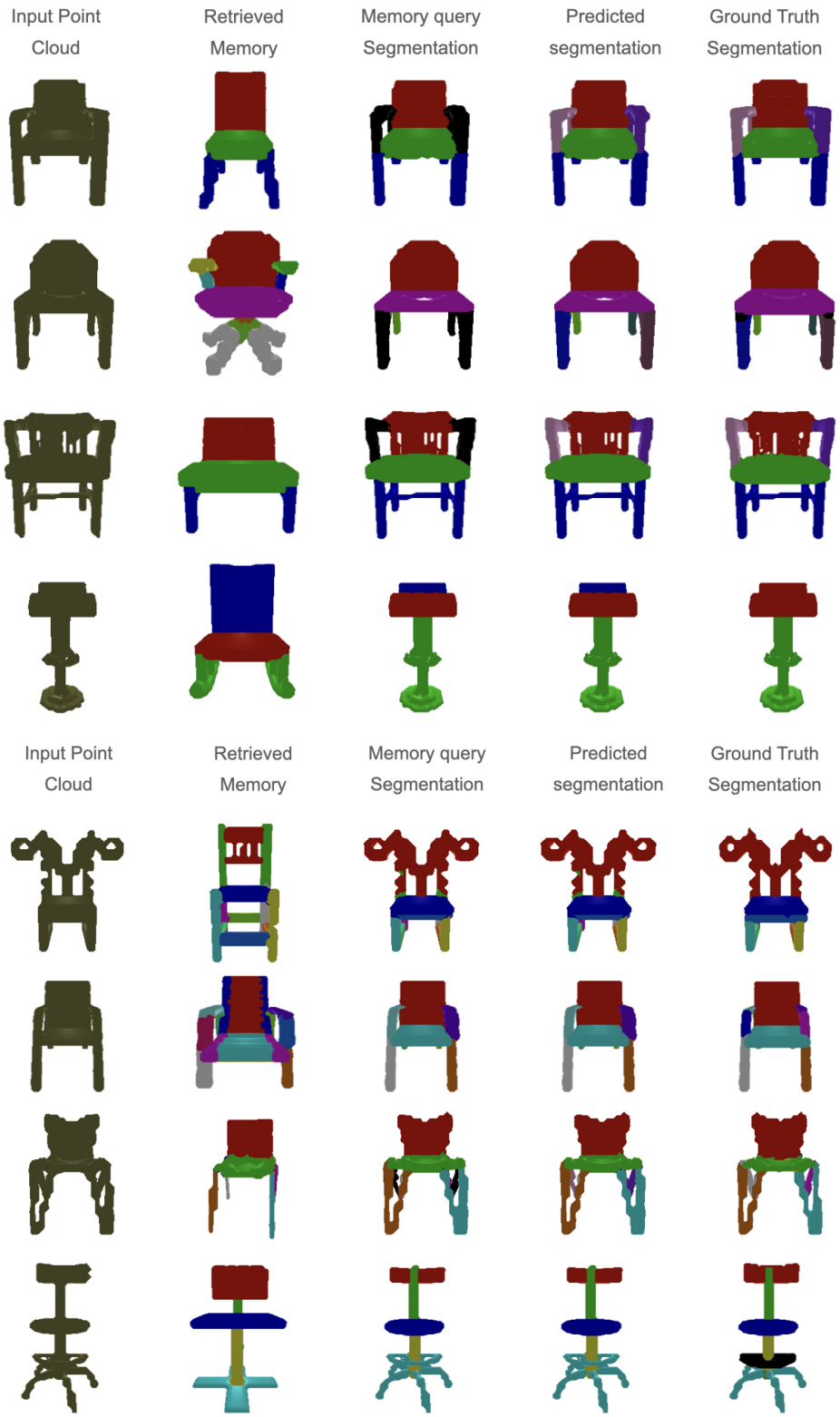

Figure 5: We show some qualitative object parsing results that are predicted by Analogical Networks .

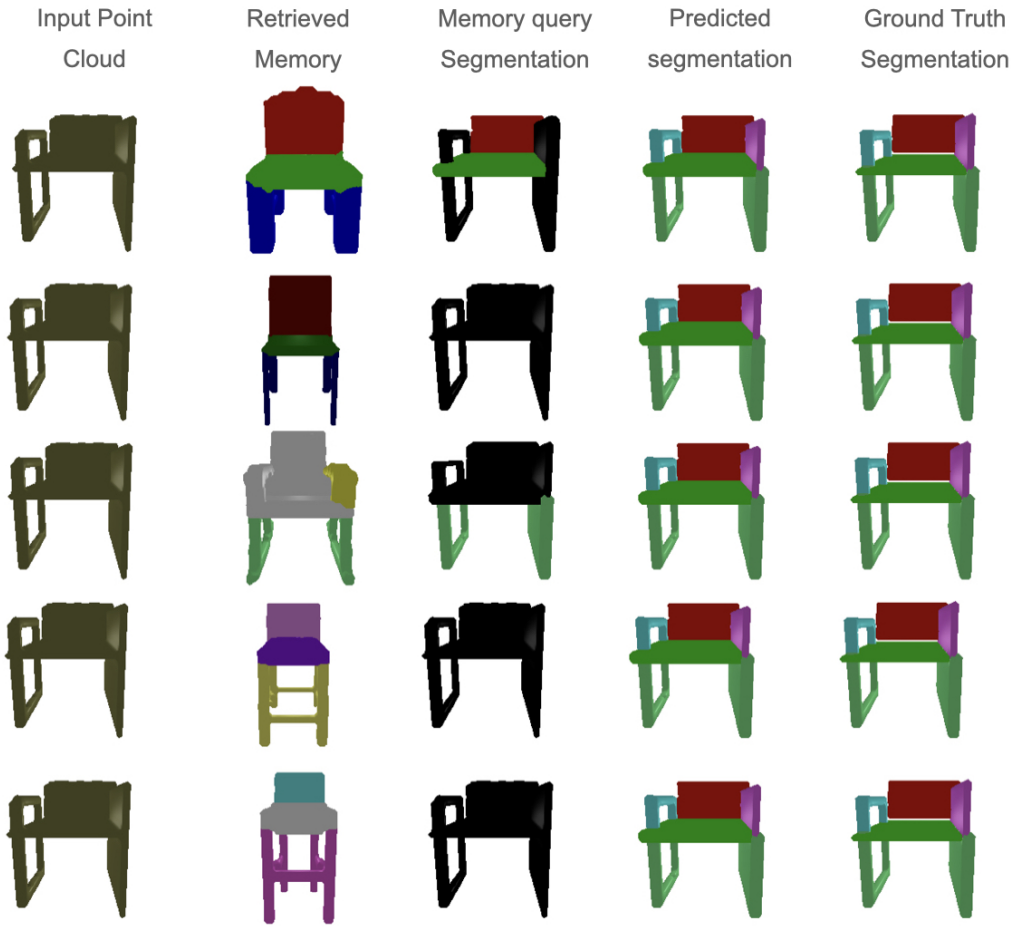

Figure 6: We show the parsing of input point cloud using our multi-memory Analogical Networks . Our model takes as input 5 different memories simultaneously and then parses the object.

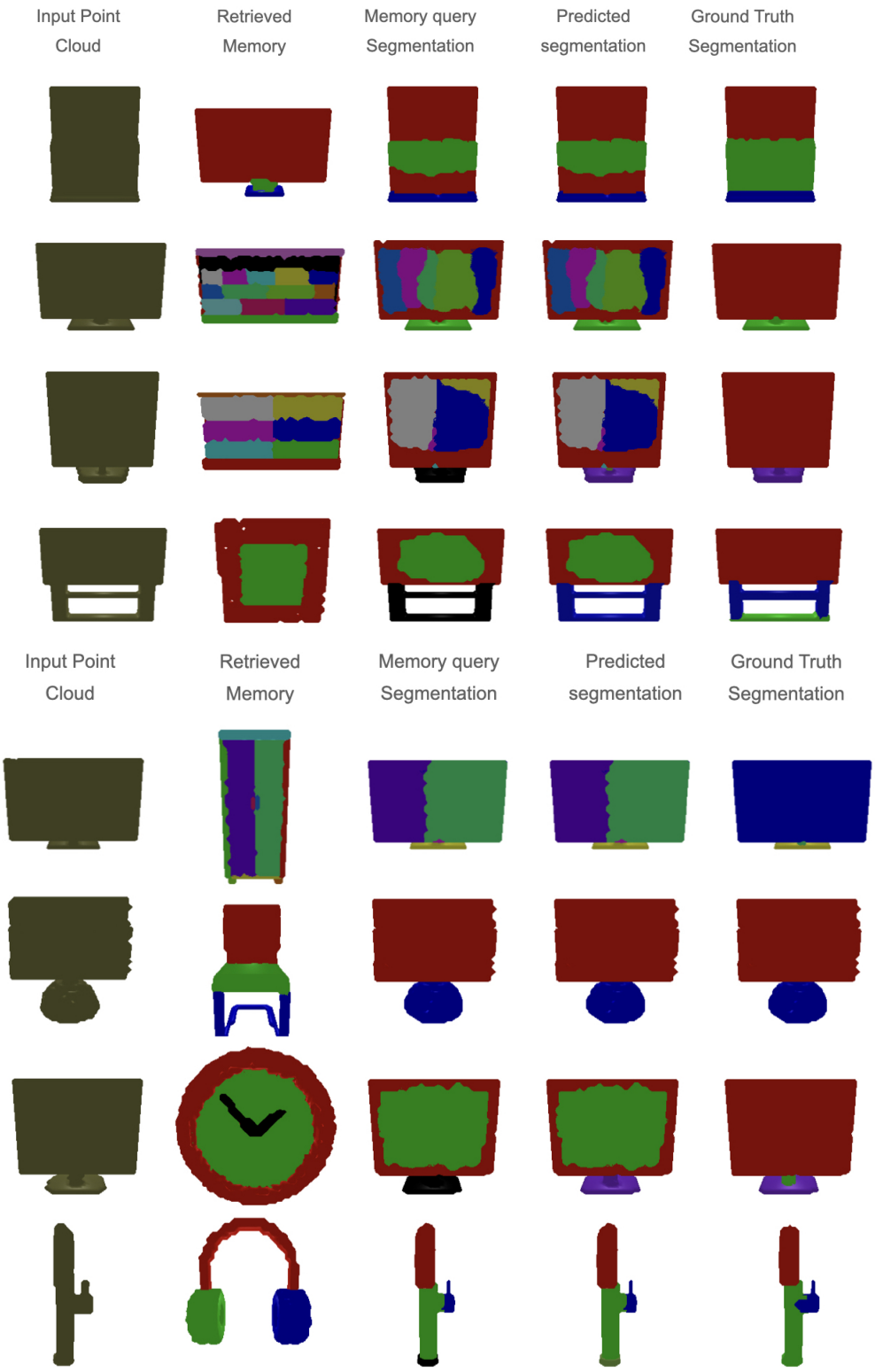

Figure 7: To qualitatively evaluate the effect of modulation in parsing, we modulate the input point cloud with a different category object and show its corresponding object parsing that are predicted by Analogical Networks .

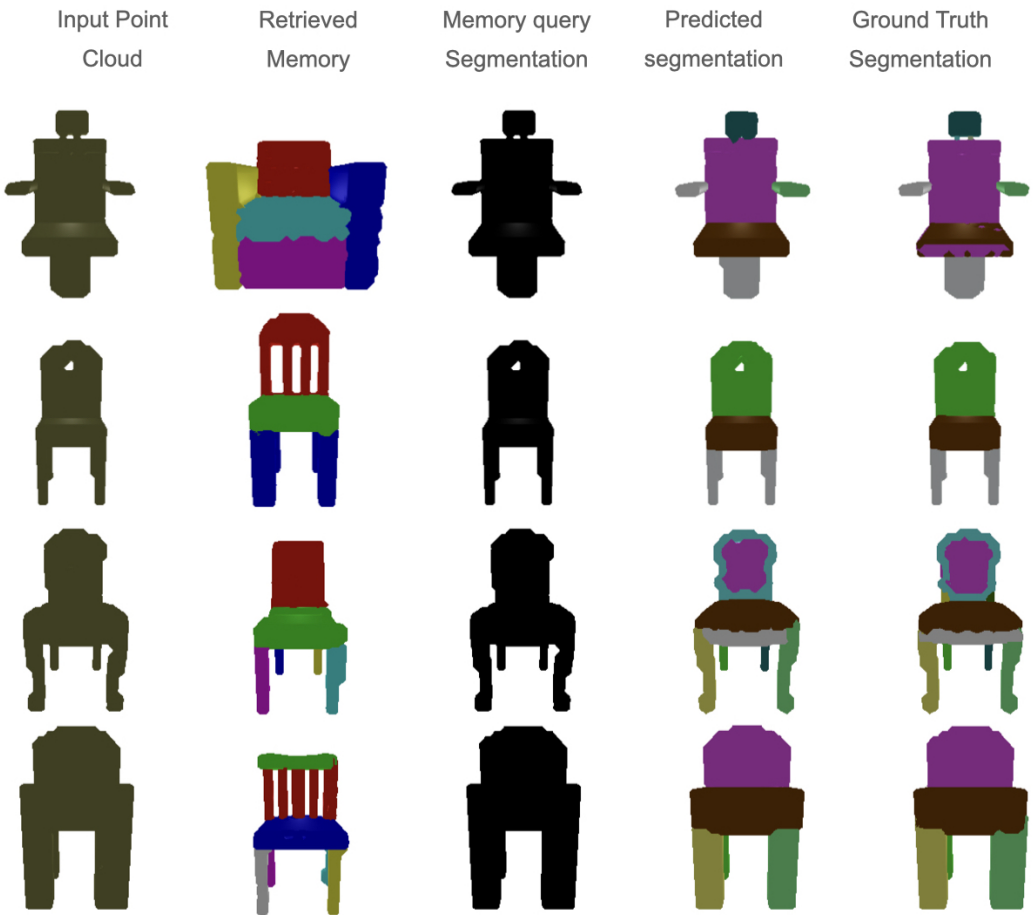

Figure 8: We show the parsing of input point cloud using Analogical Networks single memory w/o pretrain. Since most regions are black in column 3, it denotes that memory part query don't decode anything and everything is being decoded by scene agnostic queries. This highlights the role of within-instance pretraining for emergence of part correspondence.