Chameleon: A Data-Efficient Generalist for Dense Visual Prediction in the Wild

Donggyun Kim[1], Seongwoong Cho[1], Semin Kim[1], Chong Luo[2], Seunghoon Hong[1]

[1]School of Computing, KAIST [2]Microsoft Research Asia

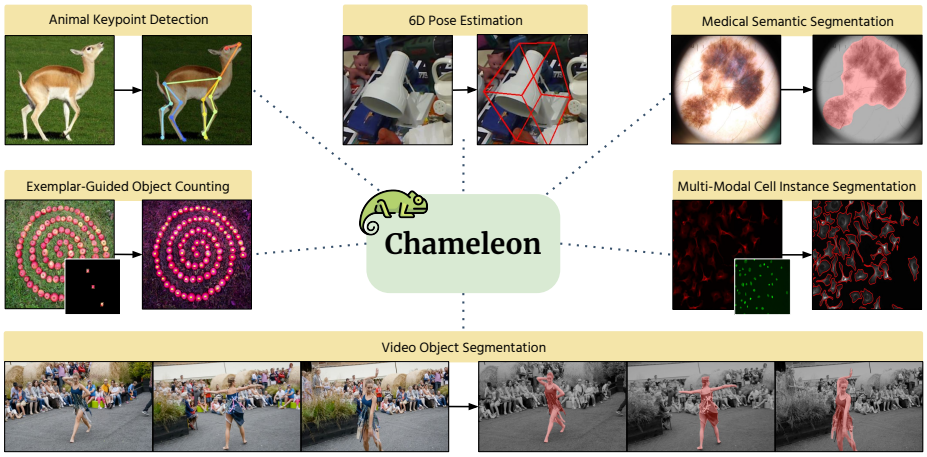


Fig. 1: Chameleon is a data-efficient generalist that can adapt to various **unseen** dense visual prediction tasks in the wild with arbitrary output structures using a handful of examples (dozens). It can also learn to utilize multi-modal inputs and user-interactions.

Abstract. Despite the success in large language models, constructing a data-efficient generalist for dense visual prediction presents a distinct challenge due to the variation in label structures across different tasks. In this study, we explore a universal model that can flexibly adapt to unseen dense label structures with a few examples, enabling it to serve as a data-efficient vision generalist in diverse real-world scenarios. To this end, we base our method on a powerful meta-learning framework and explore several axes to improve its performance and versatility for real-world problems, such as flexible adaptation mechanisms and scalability. We evaluate our model across a spectrum of unseen real-world scenarios where low-shot learning is desirable, including video, 3D, medical, biological, and user-interactive tasks. Equipped with a generic architecture and an effective adaptation mechanism, our model flexibly adapts to all of these tasks with at most 50 labeled images, showcasing a significant advancement over existing data-efficient generalist approaches. Codes are available at https://github.com/GitGyun/chameleon.

Keywords: Vision Generalist · Low-shot Learning · Dense Prediction

1 Introduction

Generalist models have gained significant attention across various fields [1, 7, 22, 27, 46] with their data efficiency in learning new tasks. In contrast to specialist models designed specifically to achieve certain tasks, generalist models aim to address a broad range of tasks, including those unseen during training. Moreover, generalist models have even begun competing with specialist models while using much less supervision attributed to incorporating two key ingredients: (1) a universal learning framework and (2) large-scale pre-training. For instance, large language models [7, 37, 39] have exhibited exceptional generalization abilities, benefiting from the universal nature of natural language and unsupervised pre-training on extensive corpora. Similarly, in the fields of algorithmic learning and reinforcement learning, large-scale training through universal interfaces—graph neural networks [22, 34, 47] and transformers [46, 49, 51], respectively—have demonstrated decent generalization performance.

However, building a data-efficient generalist for dense visual prediction tasks, which involve high-dimensional outputs with vastly diverse structure and semantics [3], remains less explored. Most of the prior efforts for general dense visual prediction [11, 25, 33] mainly focus on unifying a range of *pre-defined* tasks into a single model, rather than generalizing to *unseen* tasks. Conversely, in-context learning approaches [61, 62] attempt to solve various tasks with few demonstrations by framing the dense prediction as an image-to-image translation problem. Yet, these methods often struggle to generalize to out-of-distribution tasks that have distinct output structures and semantics unseen during training, which limits their applicability to various real-world problems. Figure 2 highlights the necessity of a more flexible adaptation mechanism in building data-efficient vision generalists for arbitrary dense visual prediction.

In this work, we aim to explore the potential of a powerful and flexible data-efficient generalist for diverse real-world dense prediction tasks. To this end, we build our method based on the framework of Visual Token Matching (VTM) [24], which directly addresses out-of-distribution tasks, with three key improvements. First, we design an encoding mechanism to incorporate varying numbers and types of input modalities, which expands the scope of adaptable tasks and addresses drifts in data modality or multi-input scenarios. Second, we enhance the task-specific adaptation mechanism by introducing a task-adaptive feature re-weighting module in the hierarchical architecture. Lastly, we enlarge and diversify the meta-training data to make the model acquire more general prior knowledge, as well as scale up the modal capacity and resolution. We meta-train the model on a large-scale dataset constructed by combining six existing datasets from diverse domains, which consists of 17 different dense visual prediction tasks.

We evaluate our method, termed Chameleon, in six downstream benchmarks composed of unique and unseen structured outputs, including tasks with video, 3D, medical and biological data, and user-interactive tasks. Our results show that existing in-context learning approaches, even if they are empowered by prompt tuning at test time, have limited generalization capability to out-of-distribution tasks, while our method successfully adapts to each scenario using

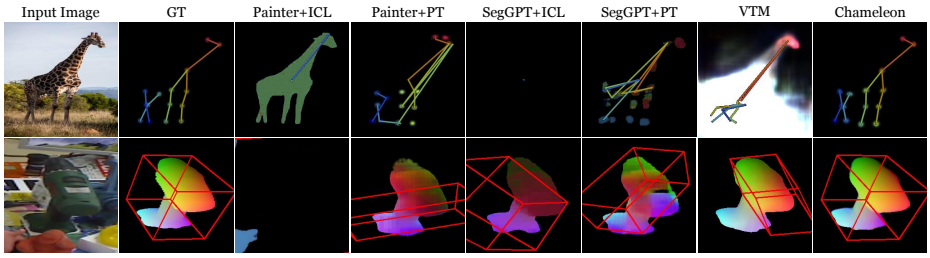


Fig. 2: Existing generalist models struggles to learn out-of-distribution tasks of unseen label semantics (6D pose) or structure (animal keypoint) during training. ICL and PT denote in-context learning and prompt tuning is used for adaptation, respectively.

at most 50 labeled examples per task, significantly outperforming the generalist baselines. Our extensive analyses also suggest that effective encoding mechanism with flexible adaptation and meta-training on a rich dataset are the key factors of successful generalization to out-of-distribution tasks.

2 Related Work

Generalist Models. Recently, generalist models have emerged as an effective approach to tackle a variety of tasks seamlessly within a single framework. In computer vision, generalist models for dense visual prediction have mainly focused on multi-task learning and prompting approaches. Multi-task learning approaches [11, 17, 25, 33, 64] train a unified architecture to solve diverse tasks, but they require numerous labeled data for each task and lack generalization ability to unseen tasks. In-context learning approaches [61, 62] consider unseen tasks but they either address in-distribution tasks whose label structures or semantics are seen during training or focus on segmentation tasks.

Few-shot Learning. Few-shot learning also targets a wide range of tasks within a single framework, but its main focus is on learning from a few labeled examples. In computer vision, most attention is paid to a specific set of tasks with dedicated architectures, such as image classification [5, 31, 52, 58], object detection [15, 18, 60], and semantic segmentation [21, 36, 50], which are not suitable for out-of-distribution generalization. Visual Token Matching [24] proposes a universal few-shot learning problem for dense visual prediction, whose main focus is out-of-distribution generalization to arbitrary tasks with only a few labels. However, it has only been demonstrated in a constrained setting where both the meta-training and testing are from the same narrow domains (*i.e.*, indoor scene), leaving its potential as a generalist in various real-world applications in question.

3 Approach

Chameleon is a data-efficient generalist based on the Visual Token Matching [24] framework, improving its design and scalability to address low-shot learning

problems in broader and more challenging real-world applications. In this section, we first present our problem setting and overall framework, then describe our improved encoder designs for handling variable multi-modal inputs (Section 3.1) and enhancing the adaptation mechanism (Section 3.2).

Problem Setting. Chameleon is designed as a versatile model capable of learning arbitrary dense prediction tasks with minimal labeled data. Formally, given a (multi-modal) query image $X^q \in \mathbb{R}^{3I_\mathcal{T} \times H_\mathcal{T} \times W_\mathcal{T}}$, our goal is to produce the per-pixel label $Y^q \in \mathbb{R}^{O_\mathcal{T} \times H_\mathcal{T} \times W_\mathcal{T}}$ of an arbitrary task $\mathcal{T}$ adaptively based on the small number of labeled examples $\mathcal{S}_\mathcal{T}$ (*i.e.* support set) by:

$$Y^q = \mathcal{F}(X^q; \mathcal{S}_\mathcal{T}), \quad \mathcal{S}_\mathcal{T} = \{(X^i, Y^i)\}_{i \leq N}. \tag{1}$$

Importantly, Chameleon does not presuppose specific priors on dense prediction tasks, allowing its application to various *unseen* tasks with the unique number of inputs $I_\mathcal{T}$ and output channels $O_\mathcal{T}$ as well as their semantics and spatial resolutions $(H_\mathcal{T}, W_\mathcal{T})$. These include a wide range of real-world problems whose inputs and outputs are defined over pixels, such as segmentation, stereo depth estimation, dense pose estimation, and exemplar-guided object counting.

Overall Framework. To support versatility, Chameleon employs the universal token matching framework [24] that formulates dense prediction as a token-level matching problem between query and support images as follows:

$$g\left(\mathbf{y}_k^q\right) = \sum_{i \leq N} \sum_{j \leq M} \sigma\left(f_\mathcal{T}\left(\mathbf{x}_k^q\right), f_\mathcal{T}\left(\mathbf{x}_j^i\right)\right) \cdot g\left(\mathbf{y}_j^i\right), \quad \forall\ k \leq M, \tag{2}$$

where $f_\mathcal{T}(\mathbf{x}_k)$ and $g(\mathbf{y}_k)$ denote the k-th token embeddings obtained by an image X and a label Y, respectively, σ is a similarity function, and M is the number of tokens per image. In this framework, the prediction for the k-th query token is produced by interpolating the support label embeddings based on its similarity to the support image embeddings. To incorporate various similarities for dense prediction in a single framework, a small amount of task-specific parameters $\theta_\mathcal{T}$ are introduced in the image encoder to adapt the image token embeddings $f_\mathcal{T}(\mathbf{x}) = f(\mathbf{x}; \theta, \theta_\mathcal{T})$ while sharing the other parameters across all tasks. After the matching in Eq. (2) is performed, the predicted query token embeddings are decoded into the query label by a label decoder $h \approx g^{-1}$.

The training protocol consists of two stages: episodic meta-training and few-shot fine-tuning. During episodic training, the whole model is trained with various dense prediction tasks sampled from a meta-training dataset to learn a general concept of matching. At this stage, Chameleon maintains and tunes separate sets of task-specific parameters $\theta_\mathcal{T}$ of the image encoder for each training task $\mathcal{T}_\text{train}$. After meta-training, Chameleon adapts to an unseen target task $\mathcal{T}_\text{test}$ by fine-tuning the task-specific parameters $\theta_\mathcal{T}$ with a small support set $\mathcal{S}_{\mathcal{T}_\text{test}}$. To further adapt the model to unseen output structures, we also fine-tune a part of the label decoder h (*e.g.,* a linear head) while fixing the rest.

The key design of Chameleon lies in how to produce the image token embeddings $f_\mathcal{T}(\mathbf{x})$. Since the matching (Eq. (2)) requires the number of the tokens in the image and the label to be consistent, we need to design a flexible encoding

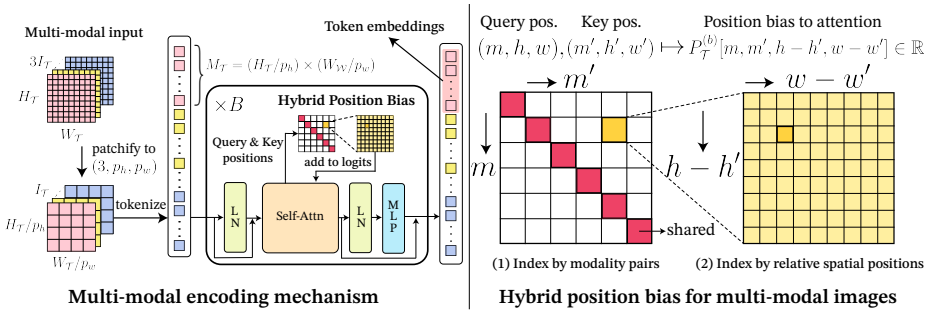


Fig. 3: Encoding mechanism of the image encoder to handle multiple input images.

mechanism that handles arbitrary input space with a varying number of modalities $I_{\mathcal{T}}$. At the same time, the encoding mechanism should reflect the unique correlation between the input modalities, which varies significantly per task. Another crucial component is the adaptation mechanism of the image encoder $f_{\mathcal{T}}(\mathbf{x}) = f(\mathbf{x}; \theta, \theta_{\mathcal{T}})$ and the choice of task-specific parameters $\theta_{\mathcal{T}}$. It should be flexible enough to adapt in order to predict vastly diverse semantics and structures of labels that are unseen during training, while not overfitting to the small support set. In the following sections, we explain how we design each component.

3.1 Encoder for Variable Input Images

To effectively handle tasks with varying numbers and types of input modalities, we design an encoding mechanism based on a transformer [57] as illustrated in Figure 3. First, we patchify a multi-modal input $X \in \mathbb{R}^{3I_{\mathcal{T}} \times H_{\mathcal{T}} \times W_{\mathcal{T}}}$ with a fixed patch size $(3, p_h, p_w)$, which results in $I_{\mathcal{T}} \times M_{\mathcal{T}}$ tokens where $M_{\mathcal{T}} = (H_{\mathcal{T}}/p_h) \times (W_{\mathcal{T}}/p_w)$ denotes the number of tokens per modality. Then we encode all the tokens at once by a transformer encoder, which contextualizes the token embeddings across modalities. Importantly, we should also encode the positional information about tokens, such that the encoder can incorporate the varying relationship between input modalities as well as the spatial prior adaptively per task. Besides the example-level contextualization, such information allows our model to learn and adapt global correlation across the input modalities per task.

To model the positional relationships between the multi-modal tokens, we design a learnable positional embedding that extends the relative position bias [4, 43]. In each b-th attention layer, the position bias between a query token at position (m, h, w) and a key token at position (m', h', w') is computed by indexing a learnable embedding $P_{\mathcal{T}}^{(b)}$ as follows:

$$P_{\mathcal{T}}^{(b)}[m, m', h - h', w - w'] \in \mathbb{R}. \tag{3}$$

The first two indices (m, m') distinguish each modality pair, such that different types of *inter-modal* interaction between tokens can be modeled. Then the remaining indices $(h - h', w - w')$ distinguish the relative spatial positions, which

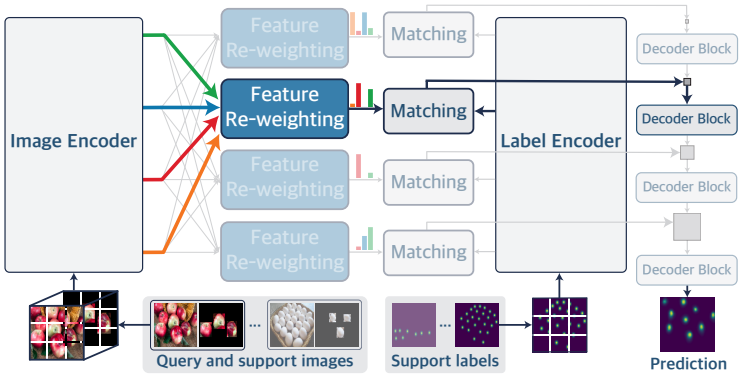


Fig. 4: Task-adaptive feature re-weighting mechanism with a hierarchical architecture. The figure highlights the matching module at the third level of the hierarchy ($l = 3$).

effectively encodes the translation-equivariance along the spatial axes. Note that we assign different embeddings $P_{\mathcal{T}}^{(b)}$ for each task as a part of task-specific parameters $\theta_{\mathcal{T}}$. This ensures the encoder not only handles different numbers of positions but also adapts to contextualize distinct relationships between modalities of each task separately. Having that the information from other modalities is contextualized to each modality, we use the first $M_{\mathcal{T}}$ tokens as image token embeddings for the matching (Eq. (2)).

3.2 Feature Modulation of the Image Encoder

To adapt to tasks with unseen semantics and structures of labels, Chameleon modulates the image encoder in two ways. First, the bias parameters $\mathbf{b}_{\mathcal{T}}$ of each image encoder layer are tuned separately for each task $\mathcal{T}$. This has been proven to efficiently modulate the features in a transformer encoder [24, 66]. Second, we introduce a feature re-weighting mechanism to associate different levels of image and label features. While it is known that using multi-level image features is beneficial for dense prediction in general [10, 29, 44], it is not straightforward in our matching formulation (Eq. (2)) to associate both features at different levels. To enable our model to capture an arbitrary correspondence between image and label features, we design a hierarchical architecture that adaptively relates different levels of image and label features depending on each task.

To this end, we extract the image and label features at L levels of their encoders and perform matching at each label feature level using all levels of the image features, as illustrated in Figure 4. To control the contribution of the image feature levels on each matching module task-specifically, we introduce a learnable matrix $\Lambda_{\mathcal{T}} \in \mathbb{R}^{L \times L}$ for each task $\mathcal{T}$ that re-weights the multi-level image features $\hat{F}_{\mathcal{T}} = [\hat{f}_{\mathcal{T}}^{(1)}(\mathbf{x}), \cdots, \hat{f}_{\mathcal{T}}^{(L)}(\mathbf{x})] \in \mathbb{R}^{L \times d}$ via matrix multiplication:

$$F_{\mathcal{T}} = [f_{\mathcal{T}}^{(1)}(\mathbf{x}), \cdots, f_{\mathcal{T}}^{(L)}(\mathbf{x})] = \Lambda_{\mathcal{T}} \hat{F}_{\mathcal{T}}, \tag{4}$$

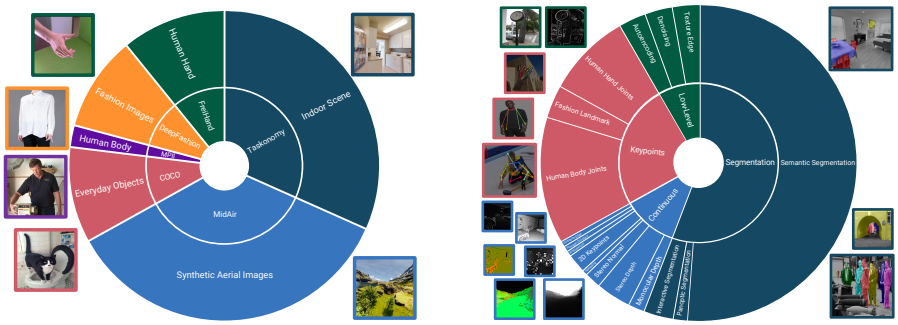


Fig. 5: Summary of our meta-training dataset. Left: image domains (outer circle) and source datasets (inner circle). Sizes correspond to the dataset size. Right: task categories (inner circle) and specific tasks (outer circle). Sizes correspond to the sampling ratio.

where each row of $\Lambda_{\mathcal{T}}$ is normalized to sum to 1 such that the total contribution from the image feature levels remains constant. Then each re-weighted feature $f_{\mathcal{T}}^{(l)}(\mathbf{x})$ is passed to l-th matching module:

$$g^{(l)}\left(\mathbf{y}_k^q\right) = \sum_{i \le N} \sum_{j \le M} \sigma^{(l)}\left(f_{\mathcal{T}}^{(l)}\left(\mathbf{x}_k^q\right), f_{\mathcal{T}}^{(l)}\left(\mathbf{x}_j^i\right)\right) \cdot g^{(l)}\left(\mathbf{y}_j^i\right), \quad 1 \le l \le L. \tag{5}$$

After performing the matching at L levels, we convert the outputs into a feature pyramid whose resolution increases as the level decreases, which are progressively decoded by a convolutional decoder [44].

In this way, our model can adapt to various tasks having different optimal correspondence between the image and label features (see Figure 7 for the learned feature weights in downstream tasks), as well as adapting the image features themselves via bias tuning. Since the task-specific parameters introduced in the image encoder $\theta_{\mathcal{T}} = (P_{\mathcal{T}}, \mathbf{b}_{\mathcal{T}}, \Lambda_{\mathcal{T}})$ occupy a small portion of the whole parameters, Chameleon is robust to over-fitting during fine-tuning.

4 Scaling up the Data and the Model

We investigate strategies to enhance the generalization of Chameleon over various unseen dense prediction tasks by collecting a large-scale meta-training dataset (Section 4.1) and scaling up the model capacity and resolutions (Section 4.2).

4.1 Meta-Training Data with Diverse Tasks and Domains

To achieve robust generalization across real-world scenarios, meta-training on diverse domains and tasks constitutes a crucial element of Chameleon. To this end, we curated a large-scale meta-training dataset comprising around 1.2 million images drawn from six prominent datasets: Taskonomy [68], COCO [8, 30], MidAir [16], MPII [2], DeepFashion [32], and FreiHand [70]. As summarized in

Figure 5, the dataset covers a wide range of domains (indoor to outdoor) and tasks (continuous to categorical) considered in mainstream vision benchmarks, which makes Chameleon generally applicable to many real-world scenarios.

Our meta-training dataset consists of dense labels from 14 different dense prediction tasks, which can be roughly categorized into continuous signal prediction, semantic segmentation, and keypoint detection. We also augment the dataset with three unsupervised tasks, namely autoencoding, denoising, and edge detection (see Figure 5 for the sampling ratio of each task). To include tasks with multi-modal input, we use stereo images offered by the MidAir dataset. In addition, we simulate an interactive segmentation task using instance segmentation labels in the COCO dataset by composing a pair of images as input, where the first element is an RGB image and the second image includes marked positions of several pixels sampled within the target object instances to be segmented.

4.2 Scaling up the Model

To boost the performance of Chameleon in the wild, we scale up the model capacity from a base implementation of VTM [24]. Since the image encoder plays a central role in the matching, we scale it up to pre-trained BEiTv2-Large [40]. To match the correspondence between the image and label encoders, we also scale up the label encoder to ViT-Large [14] and increase the dimension and number of heads in the matching module accordingly. Finally, the number of convolution channels in the label decoder has increased from 96 to 256.

Since the performance of dense prediction is generally sensitive to the image resolution, Chameleon adapts to the resolution $(H_{\mathcal{T}}, W_{\mathcal{T}})$ defined for each target task $\mathcal{T}$. This can be done by performing spatial interpolation of the positional embeddings of the transformer encoders, both for images and labels. To avoid the heavy meta-training at high resolution, we meta-train Chameleon with $(224, 224)$ resolution and then fine-tune it with the adapted resolution, which efficiently boosts up the downstream performance.

5 Experiments

This section presents the evaluation results of Chameleon on six benchmark datasets and internal analysis. More results and detailed descriptions of implementation and experiments are in the Appendix.

Generalist Baselines. We compare our model with three data-efficient generalist approaches: Painter [61], SegGPT [62], and VTM [24]. Painter and SegGPT can be used in unseen tasks with or without test-time adaptation through In-Context Learning (ICL) or Prompt Tuning (PT), respectively. Therefore, we evaluate Painter and SegGPT under both settings, where we apply SegGPT+ICL in only segmentation tasks since the model cannot handle the continuous label. For a fair comparison, we use the same support set for fine-tuning (VTM, Painter+PT, SegGPT+PT) and prompting (Painter+ICL, SegGPT+ICL). As all of these baselines do not support multiple input images, we apply them on tasks having a single input image.

Specialist Baselines. To provide a reference, we also report the performance of two specialist models for each task trained with full supervision. Since our goal is *not* beating the state-of-the-arts in individual benchmarks but demonstrating the generality, we avoid specialists that incorporate heavy task-specific post-processing or extra supervision as they are orthogonal to the model.

5.1 Downstream Tasks

To evaluate the generality of our method in real-world few-shot settings, we select six downstream tasks covering diverse output semantics and structures as well as input domains and modalities that are **unseen** in the meta-training.

Animal Keypoint Detection. To test whether our model can flexibly adapt to unseen output structures, we select animal keypoint detection. The objective is to predict the joint locations of animals, which can be converted to a multi-channel dense heatmap. Note that the output structure, *i.e.,* definition of keypoints and their spatial relationships, are unseen during meta-training. We evaluate our model on the AP-10K [65] dataset, where we select eight species with distinctive features (antelope, cat, elephant, giraffe, hippo, horse, mouse, and pig) and report the mean average precision (AP) [65] over them. For simplicity of post-processing, we exclude images with multiple instances.

6D Pose Estimation. To test whether our model can also adapt to unseen output semantics, we select 6D pose estimation. The objective is to predict the 6D extrinsic camera matrix that represents the rotation and translation of a target object. We formulate it as a dense prediction by predicting dense correspondence between each image pixel and 3D vertex of the provided CAD model, from which the 6D pose is obtained by Perspective-n-Point algorithm [26]. Indeed, the labels have distinct semantics and structure from those of meta-training tasks. We evaluate our model on the LineMOD [20] dataset and report the ADD score [42] measuring the distance of vertices in 3D space.

Exemplar-Guided Object Counting. To test whether our model can exploit a user interaction as an extra image modality, we select exemplar-guided object counting. The objective is to count all objects in an image specified by three bounding box exemplars, which are represented by two images: RGB image and an exemplar guide that highlights the bounding box areas. In this task, the model must use the exemplar guide to figure out target objects to be counted. We formulate the task to predict the heatmap of object centers, from which the number of objects is obtained by counting the modes. We employ the FSC-147 [45] dataset and report mean absolute error (MAE) following the literature [9, 13].

Cell Instance Segmentation. Cell instance segmentation also has multi-modal input images, with distinct domains from the natural images. The objective of this task is to segment all cell instances within a bi-model image (one for cytoplasm and another for nuclei). Following [54], we formulate the task as flow estimation, where the model predicts vertical and horizontal gradients of each cell towards its center along with foreground segmentation. As in 6D pose estimation, this output representation has distinct semantics and structures from the meta-

training tasks. We evaluate our model on the Cellpose [54] dataset and report average precision with threshold IoU=0.5 (AP_{50}).

Skin Lesion Segmentation. We select skin lesion segmentation as an *in-distribution* but *out-of-domain* task, where the objective is to segment the skin lesion in dermatoscopic images. We employ ISIC 2018 [35] dataset and report average F1 score of 5-fold cross-validation, following the literature [19, 56].

Video Object Segmentation. Finally, to further explore the potential of our model in the wild, we select video object segmentation. The objective is to track target objects over an entire video, which are specified in the first frame. We formulate this task as 1-shot image segmentation by treating the first frame as support and the remaining as queries, where we augment the 1-shot support with random cropping. Note that, unlike common specialists in this literature, we neither exploit any temporal correlation nor train our model on video data. We employ the DAVIS 2017 [41] dataset and report the $\mathcal{J}\&\mathcal{F}$ score [12, 59].

5.2 Main Results

Table 1 summarizes the performance of our model and baselines on the six downstream tasks. In general, our model significantly outperforms the generalist baselines in all tasks, which shows the effectiveness of our approach in low-shot learning of diverse dense visual prediction in real-world applications. We discuss the results of each task in the following paragraphs.

Animal Keypoint Detection. In this task, our model should understand not only the appearance of distinctive animal body parts but also their spatial priors to resolve ambiguities in prediction. Since these are largely different across species and any objects in meta-training data, the task requires rapid adaptation to unseen domains and output structures. As shown in Figure 8, Chameleon successfully predicts keypoints of eight species with varying appearance and body configurations. Interestingly, our model seems to leverage the spatial prior to localizing the missing parts (occlusions in Antelope and Cat) and distinguish left and right. We also observe that the generalist baselines struggle to learn this task despite the test-time adaptation (see Figure 2), showing the effectiveness of our model in adapting to unseen output structures.

6D Pose Estimation. In this task, our model has to predict 6D pose of an object, which is different from any meta-training tasks in both knowledge to solve it and the output structure. Without leveraging a dedicated architecture for 3D understanding, Chameleon successfully adapts to the task, even outperforming some of the specialized baselines. To further analyze if our model really understands the task, we visualize the attention score in the matching (Eq. (2)) in Figure A.15. It shows that the similarity of the query image patch with the support images is highly correlated with 3D positions, which is desirable for the task. We also note that learned weights in the feature re-weighting (Figure 7) tend to be inversely correlated with the feature levels. It indicates that the model leverages the high-level semantics to capture fine details in labels, which is reasonable in 3D understanding. These observations indicate that our model can adapt to novel 3D understanding tasks with unique output structure.

Table 1: Comparison with specialists for each task and generalists based on in-context learning and parameter-efficient fine-tuning. Generalists use 1-shot support for DAVIS 2017, 20-shot for AP-10K and ISIC 2018, and 50-shot for the others.

	animal kp. AP-10K	6D pose LineMOD	skin les. seg. ISIC 2018	video obj. seg. DAVIS 2017	obj. count. Cellpose	cell inst. seg. FSC-147
	AP ↑	ADD ↑	F1 ↑	$\mathcal{J}\&\mathcal{F}$ ↑	MAE ↓	AP_{50} ↑
Specialists, fully-supervised						
SimpleBaseline [63]	64.9	-	-	-	-	-
HRNet [55]	69.8	-	-	-	-	-
DPOD [67]	-	83.0	-	-	-	-
CDPN [28]	-	89.9	-	-	-	-
FTN [19]	-	-	89.7	-	-	-
UNeXt [56]	-	-	89.8	-	-	-
XMem [12]	-	-	-	87.7	-	-
ISVOS [59]	-	-	-	88.2	-	-
CounTR [9]	-	-	-	-	12.0	-
LOCA [13]	-	-	-	-	10.8	-
Stardist [48]	-	-	-	-	-	67.0
Cellpose [54]	-	-	-	-	-	70.4
Generalists, low-shot						
Painter [61] + ICL	0	0	36.3	34.6	-	-
Painter [61] + PT	0.4	15.7	86.8	53.9	-	-
SegGPT [62] + ICL	-	-	60.2	75.6	-	-
SegGPT [62] + PT	2.0	23.1	88.1	67.0	-	-
VTM [24]	9.1	59.3	84.1	54.0	-	-
Chameleon (ours)	67.2	85.2	88.5	77.5	12.0	70.3

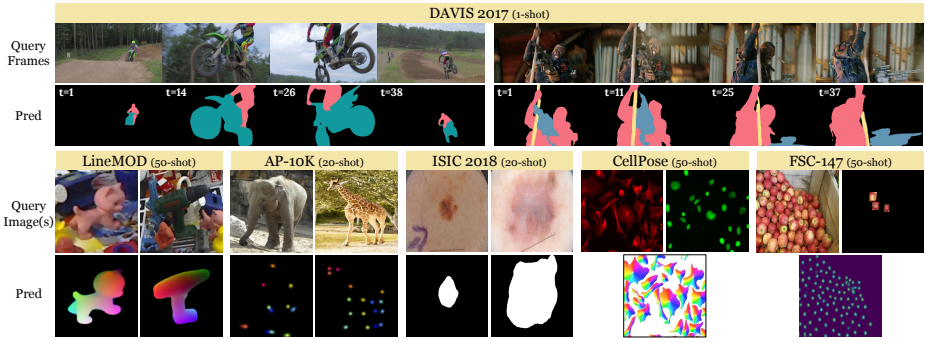


Fig. 6: Qualitative results of Chameleon in six downstream benchmarks. We color-coded outputs from different channels. t denotes the frame number.

Medical Semantic Segmentation. In this task, the model has to adapt to a huge domain shift from natural images in meta-training data to medical images. As shown in Table 1 and Figure A.16, our model successfully adapts even with such domain gaps, while in-context learning methods struggle. Not surprisingly, with prompt tuning, Painter and SegGPT become competitive with our

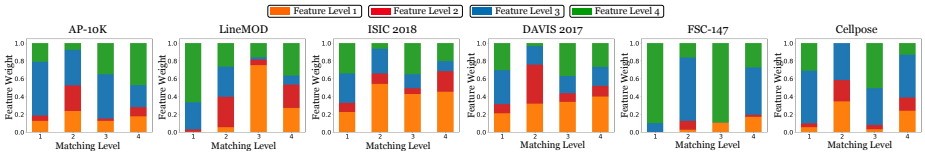


Fig. 7: Learned feature weights in downstream tasks.

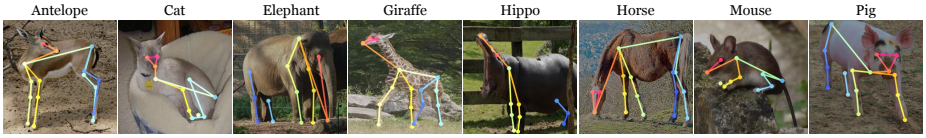


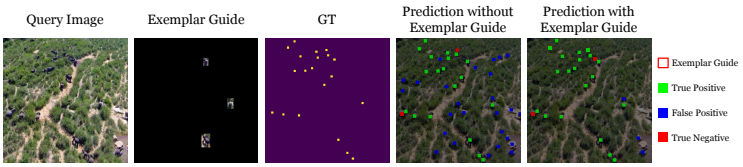
Fig. 8: Keypoint prediction of Chameleon on eight animal species.

model, as they can address out-of-domain tasks with seen label semantics and structures. Still, Chameleon outperforms all the generalist baselines, showing its effectiveness.

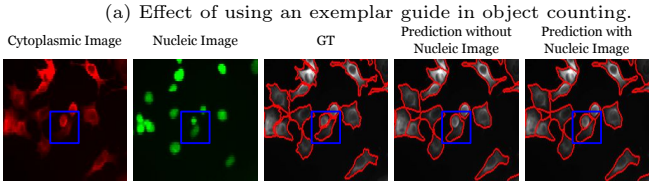
Video Object Segmentation. Although the meta-training tasks include segmentation, generalizing our model to video object segmentation is challenging since it is learned on images and unaware of relating temporally distant objects. Surprisingly, by matching each frame independently with the label of the first frame, Chameleon successfully tracks objects under significant appearance variations (Figure 6), achieving comparable performance to the specialist baselines that heavily rely on temporal correlation. As shown in Figure A.20, most failure cases of our method are due to ambiguous distractors, which additional frame labels can resolve. Indeed, our method can naturally incorporate such additional labels while it is not straightforward in baselines due to the causal inference, and Chameleon begins to surpass them with four frame labels (Figure 10).

Exemplar-Guided Object Counting. The ability to process multiple input images also allows Chameleon to be applied to user-interactive tasks, such as exemplar-guided object counting. In this task, our model exploits the exemplar guide given as the second image to identify objects to count. As shown in Figure 9 (a), counting objects without such guidance inevitably includes many false positives, whereas our method successfully excludes them using the guidance. Together with cell instance segmentation tasks, it shows that our method can adapt to multi-modal inputs with vastly different semantics effectively with the encoding mechanism introduced in Section 3.1.

Cell Instance Segmentation. This task involves out-of-domain images and labels, but more interestingly, solving this task requires understanding of bi-modal images of cytoplasm and nuclei. It requires our model to take these two images and learn to leverage their exclusive cues for cell instance segmentation by adapting the multi-modal position bias. As shown in Figure 9 (b), Chameleon successfully utilizes such information, by distinguishing two instances entangled in cytoplasmic image using information in nucleic image.



(a) Effect of using an exemplar guide in object counting.



(b) Effect of using a nucleic image in cell instance segmentation.

Fig. 9: Effect of using multi-modal input. (a) In object counting, Chameleon excludes false positives (bushes) by using the exemplar guide. (b) In cell instance segmentation, Chameleon separates two cells in the blue box by exploiting the nucleic image.

5.3 Ablation Study

Component-wise Analysis. We conduct an ablation study to analyze the effect of each component introduced in Chameleon. As our model is based on the VTM framework, we ablate our improvements from VTM one by one. As shown in Table 2, all components contribute to improving the downstream performance (scaling up the model and meta-training data), as well as broadening the scope to multi-modal applications (encoding mechanism for multi-modal inputs). Notably, we observe that feature re-weighting improves the performance considerably especially when the structure and semantics of the labels are largely different from meta-training tasks, such as 6D pose estimation, object counting, and cell instance segmentation. As shown in Figure 7, the learned weights vary substantially across tasks, showing its effectiveness in adapting matching modules to the out-of-distribution tasks.

Ablation Study on Meta-Training Data. To further analyze the effect of meta-training, we conduct an ablation study in Table 3 by gradually increasing the scale and diversity of meta-training data. The performance of Chameleon tends to consistently improve as we diversify domains and tasks in the meta-training dataset. Interestingly, such improvements are often from adding tasks *less or barely correlated* with the downstream tasks. For instance, adding synthetic drone images with continuous labels (MidAir) improves the animal pose estimation by a large margin, or adding keypoint detection tasks (KP-4) improves 6D pose estimation. It shows that our method can effectively leverage the indirect correlations of meta-training and downstream tasks through universal matching, which is critical in generalization to out-of-distribution tasks.

Effect of Support Size. To study the effect of support set size, we plot the performance of Chameleon with three different shots in Figure 10. We observe that performance consistently increases as support size increases, and beats specialist

Table 2: Ablation study on the contributions of each component.

Model Variant	AP-10K	LineMOD	ISIC 2018	DAVIS 2017	FSC-147	Cellpose
	AP ↑	ADD ↑	F1 ↑	$\mathcal{J}\&\mathcal{F}$ ↑	MAE ↓	AP_{50} ↑
VTM	9.1	59.3	84.1	54.0	-	-
+ Enlarged Backbone	25.3	73.9	85.4	66.8	-	-
+ Large and Diverse Data	65.0	73.6	86.5	73.6	-	-
+ Variable-Input Encoder	63.6	73.3	**88.9**	76.0	17.9	67.2
+ Feature Weighting **(Ours)**	**67.2**	**85.2**	88.5	**77.5**	**12.3**	**70.3**

Table 3: Ablation study on meta-training dataset. COCO (seg.) refers to using only segmentation labels in COCO dataset, and KP-4 refers to using four keypoint detection datasets (COCO, MPII, Deepfashion, and Freihand).

Taskonomy	MidAir	COCO (seg.)	KP-4	AP-10K	LineMOD	ISIC 2018	DAVIS 2017	FSC-147	Cellpose
				AP ↑	ADD ↑	F1 ↑	$\mathcal{J}\&\mathcal{F}$ ↑	MAE ↓	AP_{50} ↑
✓	✗	✗	✗	19.3	84.0	84.6	66.4	-	-
✓	✓	✗	✗	42.4	80.0	87.3	69.3	17.1	66.9
✓	✓	✓	✗	65.1	83.3	88.0	74.6	14.9	69.2
✓	✓	✓	✓	**67.2**	**85.2**	**88.5**	**77.5**	**12.3**	**70.3**

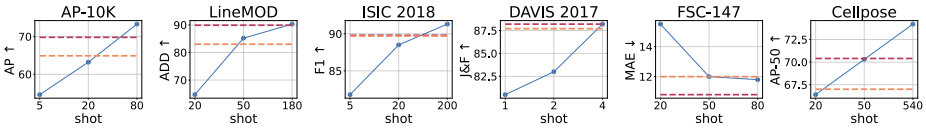


Fig. 10: Downstream performance of Chameleon (blue line) by varying the support set size. Dotted lines correspond to the performance of specialist models of each task.

baselines in all benchmarks with only dozens of labels at most. This demonstrates the potential of Chameleon in various dense visual tasks in the wild, whose available supervision ranges between a couple of examples to dozens.

6 Conclusion

We proposed Chameleon, a data-efficient generalist for arbitrary unseen dense visual prediction. Based on a token-level matching framework, we introduced a flexible encoding mechanism for multiple input images and a powerful task-specific adaptation mechanism for hierarchical architecture. We have also collected a meta-training dataset by curating six datasets containing diverse dense visual tasks from various domains. Through extensive experiments, we showed that Chameleon can learn various unseen tasks with distinct label structures and semantics from training with at most dozens of labels.

Acknowledgements. This work was supported in part by the National Research Foundation of Korea (RS-2024-00351212), IITP grant (RS-2022-II220926, RS-2022-II220959, and RS-2021-II212068) funded by the Korean government (MSIT), and Naver Cooperation.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 2
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014) 7, 2
3. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:2307.13721 (2023) 2
4. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=p-BhZSz59o4 5
5. Bateni, P., Goyal, R., Masrani, V., Wood, F., Sigal, L.: Improved few-shot visual classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14493–14502 (2020) 3
6. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008) 1
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2
8. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 7, 2
9. Chang, L., Yujie, Z., Andrew, Z., Weidi, X.: Countr: Transformer-based generalised visual counting. In: British Machine Vision Conference (BMVC) (2022) 9, 11, 13
10. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 6
11. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 909–919 (2023) 2, 3
12. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 640–658. Springer (2022) 10, 11
13. Djukic, N., Lukezic, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18872–18881 (2023) 9, 11, 13
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.:

An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 8, 16
15. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4013–4022 (2020) 3
16. Fonder, M., Droogenbroeck, M.V.: Mid-air: A multi-modal dataset for extremely low altitude drone flights. In: Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (June 2019) 7, 2
17. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Hu, H., Chen, D., et al.: Instructdiffusion: A generalist modeling interface for vision tasks. arXiv preprint arXiv:2309.03895 (2023) 3
18. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5321–5330 (2022) 3
19. He, X., Tan, E.L., Bi, H., Zhang, X., Zhao, S., Lei, B.: Fully transformer network for skin lesion analysis. Medical Image Analysis **77**, 102357 (2022) 10, 11, 9
20. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: 2011 international conference on computer vision. pp. 858–865. IEEE (2011) 9
21. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2022) 3
22. Ibarz, B., Kurin, V., Papamakarios, G., Nikiforou, K., Bennani, M., Csordás, R., Dudzik, A.J., Bošnjak, M., Vitvitskyi, A., Rubanova, Y., et al.: A generalist neural algorithmic learner. In: Learning on Graphs Conference. pp. 2–1. PMLR (2022) 2
23. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an image edge detection filter using the sobel operator. JSSC (1988) 1
24. Kim, D., Kim, J., Cho, S., Luo, C., Hong, S.: Universal few-shot learning of dense prediction tasks with visual token matching. In: The Eleventh International Conference on Learning Representations (2023) 2, 3, 4, 6, 8, 11, 1, 17
25. Kolesnikov, A., Susano Pinto, A., Beyer, L., Zhai, X., Harmsen, J., Houlsby, N.: Uvim: A unified modeling approach for vision with learned guiding codes. Advances in Neural Information Processing Systems **35**, 26295–26308 (2022) 2, 3
26. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o(n) solution to the pnp problem. International journal of computer vision **81**, 155–166 (2009) 9, 6
27. Li, H., Zhu, J., Jiang, X., Zhu, X., Li, H., Yuan, C., Wang, X., Qiao, Y., Wang, X., Wang, W., et al.: Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2691–2700 (2023) 2
28. Li, Z., Wang, G., Ji, X.: Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7678–7687 (2019) 11, 6
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 6
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 7, 1

31. Liu, L., Hamilton, W.L., Long, G., Jiang, J., Larochelle, H.: A universal representation transformer layer for few-shot image classification. In: International Conference on Learning Representations (2020) 3
32. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 7, 2
33. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=E01k9048soZ 2, 3
34. Mahdavi, S., Swersky, K., Kipf, T., Hashemi, M., Thrampoulidis, C., Liao, R.: Towards better out-of-distribution generalization of neural algorithmic reasoning tasks. Transactions on Machine Learning Research (2022) 2
35. Milton, M.A.A.: Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802 (2019) 10
36. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6941–6952 (2021) 3
37. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2** (2023) 2
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 16
39. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022) 2
40. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022) 8, 16
41. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017) 10
42. Rad, M., Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: Proceedings of the IEEE international conference on computer vision. pp. 3828–3836 (2017) 9
43. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020) 5
44. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021) 6, 7
45. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3394–3403 (2021) 9
46. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-maron, G., Giménez, M., Sulsky, Y., Kay, J., Springenberg, J.T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., de Freitas, N.: A generalist agent. Transactions on Machine Learning Research (2022), https://openreview.net/forum?id=1ikK0kHjvj, featured Certification 2

47. Rodionov, G., Prokhorenkova, L.: Neural algorithmic reasoning without intermediate supervision. Advances in Neural Information Processing Systems **36** (2024) 2
48. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 265–273. Springer (2018) 11
49. Schubert, I., Zhang, J., Bruce, J., Bechtle, S., Parisotto, E., Riedmiller, M., Springenberg, J.T., Byravan, A., Hasenclever, L., Heess, N.: A generalist dynamics model for control. arXiv preprint arXiv:2305.10912 (2023) 2
50. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: BMVC (2017) 3
51. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Conference on Robot Learning. pp. 785–799. PMLR (2023) 2
52. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017) 3
53. Steder, B., Rusu, R.B., Konolige, K., Burgard, W.: Narf: 3d range image features for object recognition. In: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). vol. 44, p. 2. Citeseer (2010) 1
54. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature methods **18**(1), 100–106 (2021) 9, 10, 11, 14
55. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 11
56. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 23–33. Springer (2022) 10, 11, 9
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 5
58. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016) 3
59. Wang, J., Chen, D., Wu, Z., Luo, C., Tang, C., Dai, X., Zhao, Y., Xie, Y., Yuan, L., Jiang, Y.G.: Look before you match: Instance understanding matters in video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2268–2278 (2023) 10, 11
60. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: International Conference on Machine Learning. pp. 9919–9928. PMLR (2020) 3
61. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. arXiv preprint arXiv:2212.02499 (2022) 2, 3, 8, 11
62. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1130–1140 (2023) 2, 3, 8, 11
63. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018) 11

64. Ye, H., Xu, D.: Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In: The Eleventh International Conference on Learning Representations (2022) 3
65. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 (2021) 9
66. Zaken, E.B., Goldberg, Y., Ravfogel, S.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1–9 (2022) 6
67. Zakharov, S., Shugurov, I., Ilic, S.: Dpod: 6d pose object detector and refiner. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1941–1950 (2019) 11, 6
68. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3712–3722 (2018) 7, 1
69. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 17
70. Zimmermann, C., Ceylan, D., Yang, J., Russel, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: IEEE International Conference on Computer Vision (ICCV) (2019), "https://lmb.informatik.uni-freiburg.de/projects/freihand/" 7, 2