

Supplementary - Visual-Semantic Decomposition and Partial Alignment for Document-based Zero-Shot Learning

Anonymous Authors

The supplementary material provides:

- Section 1.1: Ablation over different prompts.
- Section 1.2: Ablation on p in partial score.
- Section 2.1: Comparison with SOTA on different LLMs.
- Section 2.2: Comparison with generative methods.
- Section 2.3: Comparison with CLIP without prior on labels.
- Section 3: Training details.
- Section 4: Details of less-described categories.
- Section 5: Examples of category documents.

1 EXTRA ABLATIONS ON EMDEPART

1.1 Ablation over Different Prompts

To find a robust prompt, we consider the following three prompts for enriching less-described documents:

Direct Prompt: “What does a *{class name}* look like? Please describe within 500 words.”

Detailed Prompt: “Now you are a *{type}* expert. I will give you *{type}* name, and you need to give detailed information. I want you to define *{class name}*. Please describe within 500 words.”

Visual Prompt: “Now you are a *{type}* expert. I will give you *{type}* name, and you need to give detailed visual information about its shape, color, appearance, habitat, etc. I want you to define *{class name}*. Please describe within 500 words.”

We leverage the coarse-grained species as *{type}*, i.e., animal for AWA2, bird for CUB, and flower for FLO. Besides, *{class name}* denotes the name of the labeled class.

In Table 1, we see that all prompts achieve performance improvements, verifying the effectiveness of enriching less-described documents. The visual prompt achieves the best result, which enriches less-described classes with more diverse descriptions.

Table 1: Ablation over different prompts on AWA2 and FLO datasets. We test each prompt with three times. The best results overall are in bold.

Template Style	AWA2		FLO	
	T1	H	T1	H
Wiki	81.4	81.5	47.2	59.5
Direct Prompt	82.2 \pm 0.95	82.6 \pm 0.81	48.2 \pm 0.68	59.9 \pm 0.57
Detailed Prompt	84.1 \pm 0.53	83.7 \pm 0.47	51.2 \pm 0.45	64.7 \pm 0.87
Visual Prompt	86.1\pm0.16	84.8\pm0.29	53.3\pm0.41	67.3\pm0.82

1.2 Ablation on p in Partial Score

In Table 2, we ablate the value of p in the partial score function to determine the optimal value on three datasets. The number of view embeddings k are set to 4, 5, and 4 for AWA2, CUB, and FLO datasets, respectively.

Table 2: Ablation on p in partial score function. The best results overall are in bold.

Model	AWA2	CUB	FLO
	T1	T1	T1
w/o Partial Score	85.71	52.58	52.99
$p = 1$	85.32	52.64	53.31
$p = 2$	86.00	52.71	53.19
$p = 3$	86.13	52.84	52.91
$p = 4$	-	52.74	-

The results show that the optimal value of p in the partial score function varies with the dataset. Specifically, for the AWA2 and CUB datasets, a higher value ($p = 3$) leads to better performance. In contrast, for the FLO dataset, a lower value ($p = 1$) is more effective. It is noteworthy that lower values, such as $p = 1$ for AWA2 and CUB, result in the performance decrease. This is due to insufficient engagement with semantic information from view embeddings.

2 EXTRA EXPERIMENTS

2.1 Comparison with SOTA on Different LLMs

In Table 3, we compare the SOTA method in document-based ZSL on different LLMs (GPT3 [1], PaLM [3], ChatGPT [6]). We see that our EmDepart with Wiki outperforms I2MVFormer with Wiki+ChatGPT across all metrics. It verifies that modeling partial association is helpful for knowledge transfer and significantly improves performance. With the larger scale of LLMs, the EmDepart achieves better performance with more diverse visual descriptions. The model with ChatGPT achieves the best result.

Table 3: Comparison with SOTA methods on different LLMs. The best results within a method are underline. The best results overall are in bold.

Model	Auxiliary Information	AWA2		FLO	
		T1	H	T1	H
I2MVFormer [9]	Wiki	73.6	73.8	41.3	51.2
	Wiki+GPT3	74.2	74.2	44.2	54.5
	Wiki+PaLM	79.6	76.6	46.2	57.1
	Wiki+ChatGPT	<u>79.9</u>	<u>80.5</u>	<u>47.7</u>	<u>59.1</u>
EmDepart (Ours)	Wiki	81.4	81.5	47.2	59.5
	Wiki+GPT3	82.3	82.2	53.2	65.5
	Wiki+PaLM	84.1	82.5	50.2	64.3
	Wiki+ChatGPT	<u>86.1</u>	<u>84.8</u>	<u>53.3</u>	<u>67.3</u>

Table 4: Comparison with SOTA methods in document-based ZSL on three benchmark datasets. We evaluate methods on documents sourced from Wiki. The best results overall are in bold.

Type	Model	Auxiliary Information	Zero-Shot Learning			Generalized Zero-Shot Learning								
			AWA2	CUB	FLO	AWA2			CUB			FLO		
			T1	T1	T1	U	S	H	U	S	H	U	S	H
Generative	GAZSL [17]	Wiki	83.1	42.9	34.2	56.8	94.7	71.0	15.9	50.4	24.1	28.8	90.1	43.7
	f-VAEGAN-D2 [16]	Wiki	85.1	41.9	36.9	73.2	81.7	77.2	33.4	57.3	42.2	30.0	97.3	45.8
Discriminative	I2DFormer [10]	Wiki	76.4	45.4	40.0	66.8	76.8	71.5	35.3	57.6	43.8	35.8	91.9	51.5
	I2MVFormer [9]	Wiki	73.6	42.1	41.3	66.6	82.9	73.8	32.4	63.1	42.8	34.9	96.1	51.2
	EmDepart (Ours)	Wiki	81.4	50.2	47.2	76.0	87.8	81.5	42.6	56.3	48.5	42.7	97.6	59.5

2.2 Comparison with Generative Methods

In Table 4, we compare our EmDepart with SOTA generative methods, which generate samples for unseen classes and convert the ZSL to a supervised task.

On the AWA2 dataset, generative methods achieve superior performance. This is because AWA2 is a coarse-grained dataset, where collected documents provide richer discriminative information than traditional attributes. By generating samples for unseen classes, these methods achieve better results. However, generative methods fail on CUB and FLO datasets compared to discriminative methods. This is due to non-visual noisy descriptions (such as sound, diet, and organ) in documents, Which lead to hard knowledge transfer and generate low-quality samples for unseen classes.

In contrast, discriminative methods enhance the fine-grained alignment between image patches and text words, implicitly filtering out irrelevant information. Our EmDepart generates embeddings from multiple semantic views, accurately modeling the semantic alignment according to the matching information and achieving better performance.

2.3 Comparison with CLIP without Prior on Class Labels

Recently, some work [7, 8, 11, 13, 14] shows that visual descriptions are helpful for improving the performance of vision-language models like CLIP [12] on image classification tasks. However, these methods heavily rely on prior information on class names. We consider the following prompts to evaluate the generalization ability of models under the situation with and without the class name prior.

Prompt with document: “A photo of a {class name}. {document}.”

Prompt with document and without class name prior: “A photo of a {type}. {document}”

The {type} denotes the species of categories, such as animal, bird, and flower. Besides, the {class name} denotes labels in the dataset. We enrich each class with documents sourced from the encyclopedia.

In Table 5, we show the results of CLIP and our EmDepart in this situation. We see that the performance of CLIP increases with the help of documents similar to [7, 8, 11, 13, 14]. However, the performance decreases a lot when the model is without prior information on the class name. In this situation, our EmDepart outperforms CLIP and demonstrates superior generation ability.

Table 5: Comparison with CLIP in different settings. The best and worst results are in bold and red, respectively. We evaluate mean per-class accuracy for CLIP.

Model	AWA2	CUB	FLO
CLIP [12]	92.0	54.1	66.8
CLIP w/ document	92.2	57.6	71.0
EmDepart	84.8	51.9	67.3
CLIP w/ document and w/o class name	47.2	14.1	18.2
EmDepart w/o class name	68.5	43.2	56.4

3 TRAINING DETAILS

3.1 Calibrated Stacking

We apply calibrated stacking (CS) [2] to trade-off calibration degrees in the GZSL settings. This is helpful for reducing the bias towards seen classes. We modify Eq. 13 in the main paper:

$$\hat{y} = \arg \max_{d' \in \mathcal{D}^s \cup \mathcal{D}^u} (S_P(x, d') - \gamma \mathbb{I}_{\mathcal{D}^s}(\hat{y})). \quad (1)$$

Here, $\mathbb{I}_{\mathcal{D}^s}$ represents an indicator function, which is 1 when $d' \in \mathcal{D}^s$ and 0 otherwise. A calibrated factor γ is applied to trade off the calibration degree on seen classes.

3.2 Grid Search

Hyperparameters are optimized by grid search in the validation split. We set the range for $\lambda_{local} \in [0, 0.01, 0.1, 0.25, 0.5, 0.75, 1]$, $\lambda_{var} \in [0, 0.01, 0.1, 0.5, 1, 3, 5]$, $\lambda_{div} \in [0, 0.01, 0.1, 0.5, 1, 3, 5]$, and $k \in [0, 1, 2, 3, 4, 5, 6, 7, 8]$. Once the hyperparameters are confirmed, we merge the validation with the training split to obtain the performance on the test split. The effect of hyperparameters in EmDepart is shown in Figure 7 in the main paper.

3.3 Additional Training Details

We implement our framework with Pytorch and train on an Nvidia GeForce RTX 3090 GPU. Similar to [9, 10], we use the ViT-B/16 [5] pre-trained on ImageNet 1K [4] as the visual backbone, which respects the GUB split [15]. The detailed hyperparameters are shown in Table 6 for AWA2, Table 7 for CUB and Table 8 for FLO.

Table 6: Hyperparameters settings for AWA2 dataset.

Config	AWA2
Regular Training Setting	
optimizer	Adam
base learning rate	1.0e-4
dropout	0.35
batch size	64
learning rate schedule	cosine decay
warmup epochs	0
epochs	32
augmentation	RandomResizedCrop
Specific Settings in EmDepart	
λ_{local}	0.1
λ_{var}	1.0
λ_{div}	3.0
number of view embeddings k	4
p in partial score	3
τ in Eq. 9	32.0
ϵ in Eq. 4	1e-4
γ in Eq.4	0.10
dimension of semantic embedding r	256
layers of text encoder	2
layers of MLP in image perceiver	2
layers of visual SDM	2
layers of textual SDM	2

Table 7: Hyperparameters settings for CUB dataset.

Config	CUB
Regular Training Setting	
optimizer	Adam
base learning rate	8.0e-4
dropout	0.15
batch size	40
learning rate schedule	cosine decay
warmup epochs	2
epochs	32
augmentation	RandomResizedCrop
Specific Settings in EmDepart	
λ_{local}	0.5
λ_{var}	1.0
λ_{div}	3.0
number of view embeddings k	5
p in partial score	3
τ in Eq. 9	4.2
ϵ in Eq. 4	1e-4
γ in Eq.4	0.25
dimension of semantic embedding r	64
layers of text encoder	2
layers of MLP in image perceiver	2
layers of visual SDM	2
layers of textual SDM	2

Table 8: Hyperparameters settings for FLO dataset.

Config	FLO
Regular Training Setting	
optimizer	Adam
base learning rate	5.0e-4
dropout	0.12
batch size	48
learning rate schedule	cosine decay
warmup epochs	0
epochs	40
augmentation	RandomResizedCrop
Specific Settings in EmDepart	
λ_{local}	0.5
λ_{var}	1.0
λ_{div}	3.0
number of view embeddings k	4
p in partial score	1
τ in Eq. 9	4.0
ϵ in Eq. 4	1e-4
γ in Eq.4	0.75
dimension of semantic embedding r	128
layers of text encoder	2
layers of MLP in image perceiver	2
layers of visual SDM	2
layers of textual SDM	2

4 DETAILS OF LESS-DESCRIBED CATEGORIES

In our work, we leverage LLMs to supplement less-described category documents. To save computation costs, we select a set of categories instead of all categories for enriching documents. The less-described categories for each dataset are shown below.

Table 9: Details of less-described categories.

	AWA2	CUB	FLO
number of classes	50	200	102
number of less-described classes	21	74	59

AWA2: dalmatian, persian cat, german shepherd, blue whale, siamese cat, moose, gorilla, ox, fox, rabbit, chihuahua, collie, dolphin, grizzly bear, skunk, hippopotamus, spider monkey, wolf, weasel, zebra, buffalo.

CUB: Laysan Albatross, Parakeet Auklet, Yellow headed Blackbird, Bobolink, Lazuli Bunting, Gray Catbird, Yellow breasted Chat, Eastern Towhee, Chuck will Widow, Red faced Cormorant, Shiny Cowbird, Fish Crow, Mangrove Cuckoo, Least Flycatcher, Scissor tailed Flycatcher, Vermilion Flycatcher, American Goldfinch, Eared Grebe, Pied billed Grebe, Ivory Gull, Anna Hummingbird, Ruby throated Hummingbird, Long tailed Jaeger, Blue Jay, Florida Jay, Green Jay, Tropical Kingbird, Gray Kingbird, Pied Kingfisher, Hooded Merganser, Red breasted Merganser, Clark Nutcracker, White breasted Nuthatch, Orchard Oriole, Ovenbird, Horned Puffin, Common Raven, American Redstart, Baird Sparrow, Clay colored

Sparrow, Henslow Sparrow, Vesper Sparrow, Cape Glossy Starling, Summer Tanager, Elegant Tern, Black capped Vireo, Philadelphia Vireo, Bay breasted Warbler, Black throated Blue Warbler, Blue winged Warbler, Canada Warbler, Cerulean Warbler, Hooded Warbler, Kentucky Warbler, Magnolia Warbler, Prairie Warbler, Prothonotary Warbler, Louisiana Waterthrush, Red bellied Woodpecker, Red cockaded Woodpecker, Red headed Woodpecker, Bewick Wren, Rock Wren, Black footed Albatross, Least Auklet, Acadian Flycatcher, Yellow bellied Flycatcher, Pomarine Jaeger, Mockingbird, Black throated Sparrow, Cape May Warbler, Golden winged Warbler, Northern Waterthrush, Bohemian Waxwing.

FLO: hard-leaved pocket orchid, canterbury bells, sweet pea, monkshood, colt's foot, spear thistle, yellow iris, purple coneflower, peruvian lily, balloon flower, giant white arum lily, fritillary, grape hyacinth, prince of wales feathers, artichoke, sweet william, carnation, garden phlox, love in the mist, alpine sea holly, ruby-lipped cattleya, cape flower, great masterwort, sword lily, poinsettia, wallflower, marigold, oxeeye daisy, wild pansy, primula, sunflower, gaura, black-eyed susan, silverbush, californian poppy, osteospermum, bearded iris, windflower, thorn apple, morning glory, passion flower, lotus, toad lily, anthurium, frangipani, clematis, hibiscus, columbine, tree mallow, magnolia, canna lily, bee balm, ball moss, foxglove, bougainvillea, mexican petunia, blanket flower, trumpet creeper, blackberry lily.

5 EXAMPLES OF CATEGORY DOCUMENTS

We show two examples of category documents for three datasets, which are the auxiliary information in our EmDepart. We will release all the documents after the review process.

5.1 AWA2

Giraffe: The Giraffe is an animal with an enormously long neck which allows it to exploit the leaves and vegetation that are too high up for other animals to find. Despite their length, the neck of the Giraffe actually contains the same number of bones as numerous other hoofed mammals but they are simply longer in shape. The giraffe's elongated neck leads into a short body, with long and thin, straight legs and a long tail that is tipped with a black tuft that helps to keep flies away. The Giraffe tends to be white in color with brown or reddish markings that cover its body (with the exception of its white lower legs). The markings of each Giraffe are not only unique to that individual but they also vary greatly between the different Giraffe species in size, color, and the amount of white that surrounds them. All giraffes though have large eyes that along with their height give them excellent vision, and small horn-like ossicones on the top of their heads. Giraffes are animals that inhabit open woodlands and savannah where using their height they are able to see for great distances around them to watch out for approaching danger. A giraffe is a tall and slender mammal with a long neck and legs. Their coat is pale yellow with dark brown spots or irregular patches covering their whole body except for their underbelly. Their spots are unique to each individual and help them to blend into their habitat, making them difficult to spot by predators. Giraffes have small horn-like ossicones on top of their head, and their ears are small and tufted with hair. Their eyes are

large, dark, and have long eyelashes. Long neck: Giraffes are well-known for their long necks, which are an adaptation for reaching high branches and leaves in trees.

Horse: All horses have long necks that hold up their large, long heads. They have big eyes and ears, which are well-adapted for many environments. A mane of long hair grows down along their necks and their short tails are covered in coarse hairs, too. They come in a variety of colors because they have been bred so long for different traits. These animals are famously a hoofed mammal with one large toe at the end of each leg. Their hooves consist of horn material which comes in different colors. Black is the most common hoof color, but horses with white feet often have white hoofs. White hooves are actually more brittle than pigmented ones. Appaloosa horses have a beautiful mixture of multiple colors. These types of painted horses often have striped hoofs that include both pigmented and white hoof material. These animals are well-suited to all kinds of environments and climates. Domestic horses can live almost anywhere as long as they have shelter, food, and space to run. horses are generally known for their distinct physical characteristics such as a long face, large nostrils, muscular build, four legs, hooves, a mane and tail of hair, and varying coat colors and patterns. They also have big, expressive eyes, long necks, and pointed ears. horses are generally found in open fields, meadows, pastures, and sometimes in stables or barns if they are domesticated. Their natural habitat includes grassy plains, hills, and forests with access to water sources. Their surroundings are usually green and have varying degrees of vegetation cover.

5.2 CUB

Brown Creeper: Tiny, lanky songbirds with long, spine-tipped tails, slim bodies, and slender, decurved bills. Length: 4.7-5.5 in (12-14 cm). Weight: 0.2-0.3 oz (5-10 g). Wingspan: 6.7-7.9 in (17-20 cm).The bill is slender and decurved, perfect for probing into crevices in tree bark to find insects and spiders.Brown Creeper have streaked brown and buff upperparts, with a broad, buffy stripe over the eye. The underparts are white, usually hidden against the tree trunk.Their legs and feet are specialized for clinging to tree trunks, supporting their unique foraging behavior.The tail is long and spine-tipped, used for support as Brown Creeper hitch upward in a spiral around tree trunks.Their wings are well-suited for short flights between trees, necessary for their foraging style.Brown Creeper forage by hitching upward in a spiral around tree trunks and limbs, using their stiff tails for support, and fly weakly to the base of another tree to continue foraging.Brown Creeper are found in mature evergreen or mixed evergreen-deciduous forests for breeding. In winter, Brown Creeper can be found in a broader variety of forests, including deciduous woodlands.

Mockingbird: Medium-sized songbird, more slender than a thrush with a longer tail. Length: 8.3-10.2 in (21-26 cm). Weight: 1.6-2.0 oz (45-58 g). Wingspan: 12.2-13.8 in (31-35 cm).Long, thin bill with a hint of a downward curve.Overall gray-brown, paler on the breast and belly. Two white wingbars on each wing and a white patch in each wing. White outer tail feathers are also flashy in flight.Long legs that are well-adapted for running and hopping on the ground.Long tail that is gray-brown like the body, which

appears particularly long in flight and aids in balance and maneuverability. Short, rounded, and broad wings, which are efficient for quick takeoffs and agile flight. Mockingbird are known for their songs and mimicry. Mockingbird sit conspicuously on high vegetation, fences, eaves, or telephone wires, or run and hop along the ground. Mockingbird are territorial and will aggressively chase off intruders. Found in towns, suburbs, backyards, parks, forest edges, and open land at low elevations.

5.3 FLO

King Protea: The king protea, scientifically known as *Protea cynaroides*, is an extraordinary flowering plant native to the southwestern coastal regions of South Africa. As the largest and most iconic member of the Protea family, this majestic flower possesses an enchanting beauty that captures the essence of its royal title. The king protea boasts a distinctively regal appearance, with a large, spherical inflorescence that can reach up to 12 inches (30 centimeters) in diameter. This magnificent flower is characterized by its intricate structure, composed of multiple layers of petals surrounding a prominent central cone. The cone is adorned with an array of feathery, needle-like styles that extend outwards, adding a unique texture to the overall appearance. The petals of the king protea are large and sturdy, each measuring around 4 to 6 inches (10 to 15 centimeters) in length. They can vary in color, ranging from soft hues of creamy whites, blush pinks, and delicate mauves to vibrant shades of deep crimson, burgundy, and coral. The petals exhibit a velvety texture and often feature a slightly waxy coating, enhancing their appeal and adding a touch of lustrous shine. One of the distinguishing characteristics of the king protea is the presence of a prominent ring of long, stiff bracts that encircle the base of the flower. These bracts, often referred to as phylloid bracts, serve to protect and support the blooms, giving the inflorescence an impressive crown-like appearance. The bracts themselves can vary in color and are typically seen in shades of pale green, silvery grey, or even a reddish hue. In its natural habitat, the king protea thrives in a diverse range of environments, predominantly found in the fynbos biome of South Africa. It prefers well-drained soils and can be seen flourishing along sandy coastal areas, mountain slopes, and even in the slightly more arid landscapes of the region. This resilient flower has adapted to withstand harsh conditions, including periods of drought and occasional wildfires, showcasing its remarkable ability to survive and retain its majestic allure. The king protea is not merely a flower; it represents a symbol of strength, beauty, and resilience. Its captivating presence has made it an iconic emblem and a highly sought-after ornamental bloom. This regal flower is often used as a centerpiece or featured in floral arrangements, adding a sense of grandeur and elegance to any setting, whether it be a sophisticated event or a serene garden.

Sunflower: Sunflowers, scientifically known as *Helianthus annuus*, are iconic and dazzling flowers that invoke thoughts of warm, sunny days and vibrant landscapes. These majestic plants belong to the Asteraceae family and can reach impressive heights, often towering over other plants in gardens and fields. With their distinct appearance and widespread popularity, sunflowers have become a true symbol of joy, vitality, and positivity. The most striking feature of a sunflower is, undoubtedly, its enormous flower head. These

flower heads, also known as inflorescences, are an incredible sight to behold, with an impressive size measuring between 10 to 30 centimeters in diameter or even larger in some cultivated varieties. Sunflowers are aptly named due to their stunning resemblance to the sun, both in shape and color. When we examine a sunflower's blossoming head, we find a captivating arrangement of intricate details. A circular or semi-circular cluster of florets forms the center of the flower, aptly called the disk florets or the central disc. These disc florets are small, tubular-shaped, and densely packed together, creating a textured surface in stunning shades of dark brown, deep maroon, or even a rich purple-black hue. Upon closer inspection, the disc florets reveal intricate patterns and textures, often showcasing a striking contrast to the vibrant yellow petals surrounding them. The disk florets are embraced by a ring of larger, elongated florets called ray florets, which contribute to the iconic shape of a sunflower. These ray florets possess a petal-like appearance, featuring vivid yellow or sometimes orange tones. Standing upright and radiating from the center, the ray florets resemble the rays of sunshine, providing an ethereal aura to the flower. While most sunflowers bear yellow petals, cultivated varieties may display a delightful array of shades, including vibrant oranges, warm reds, and even pale creams. Sunflowers possess a well-defined reproductive structure, positioned at the center of the broad flower head. This structure consists of pistils and stamens, responsible for the plant's pollination and fertilization processes. Bees, butterflies, and other insects are commonly attracted to the sunflower's nectar and vibrant colors, aiding in the transfer of pollen from flower to flower. In terms of habitat, sunflowers are native to North America and are highly adaptable plants, capable of thriving in diverse environments. They prefer areas with abundant sunlight, often gracing the landscape of fields, meadows, and gardens. Sunflowers have a notable preference for fertile, well-draining soil, but they can also withstand periods of drought.

REFERENCES

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.
- [2] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In *Computer Vision - ECCV 2016. (Lecture Notes in Computer Science, Vol. 9906)*. Springer, 52–68.
- [3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *CoRR abs/2204.02311* (2022). arXiv:2204.02311
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society*

- Conference on Computer Vision and Pattern Recognition (CVPR 2009). IEEE Computer Society, 248–255.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net.
- [6] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *CoRR* abs/2303.10130 (2023). arXiv:2303.10130
- [7] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E. O'Connor. 2023. Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops*. IEEE, 262–271.
- [8] Sachit Menon and Carl Vondrick. 2023. Visual Classification via Description from Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.
- [9] M. Naeem, M. Ali Khan, Y. Xian, M. Afzal, D. Stricker, L. Van Gool, and F. Tombari. 2023. I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 15169–15179.
- [10] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. 2022. I2DFormer: Learning Image to Document Attention for Zero-Shot Image Classification. In *NeurIPS*.
- [11] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 15645–15655.
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021. (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.
- [13] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. 2023. ChatGPT-Powered Hierarchical Comparisons for Image Classification. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.
- [14] Karsten Roth, Jae-Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. 2023. Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 15700–15711.
- [15] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 9 (2019), 2251–2265.
- [16] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. Computer Vision Foundation / IEEE, 10275–10284.
- [17] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A Generative Adversarial Approach for Zero-Shot Learning From Noisy Texts. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. Computer Vision Foundation / IEEE Computer Society, 1004–1013.