Supplementary to
GroupViT: Semantic Segmentation Emerges from Text Supervision

Jiarui Xu[1*] Shalini De Mello[2] Sifei Liu[2] Wonmin Byeon[2]
Thomas Breuel[2] Jan Kautz[2] Xiaolong Wang[1]
[1]UC San Diego [2]NVIDIA

1. Implementation Details

1.1. Architecture

The architecture of GroupViT is based on ViT-S [7, 14] with 12 Transformer layers. Each layer consists of a multi-head self-attention block and an MLP block. The input to each block is normalized by layer normalization [1]. We connect the group tokens in the different grouping stages via MLP-Mixer layers [13]. Our text-encoder consists of 12 Transformer layers, each with a hidden dimension of 256. Following [10], the Transformer operates on a lower-cased byte pair encoding (BPE) representation of the text with a vocabulary of 49,152 words.

1.2. Fully-Supervised Transfer to Semantic Segmentation

To implement the baselines for fully-supervised transfer to semantic segmentation, we fine-tune the pre-trained ViT model jointly with a 1×1 convolutional layer appended to it for pixel-wise classification. We scale each input image by a randomly selected factor in the range of $[0.5, 2]$ and then crop random 224×224 patches from each image during training. We use the Adam [8] optimizer with a weight decay of 0.05 and a learning rate 0.001. We train all models for 4k iterations with a batch size of 16. During inference, we resize each input image to have a shorter side of size 448 pixels. We open-source our code at https://github.com/NVlabs/GroupViT.

2. Qualitative Results

PASCAL VOC 2012 We show additional qualitative results of GroupViT on the PASCAL VOC 2012 dataset, i.e. examples with a single object in Fig. 2; multiple objects from the same category in Fig. 3; and multiple objects from different categories in Fig. 4. Observe that GroupViT successfully groups and correctly classifies the objects in these various challenging scenarios.

*Jiarui Xu was an intern at NVIDIA during the project.

Figure 1. **Concepts Learnt by Group Tokens.**. We highlight the regions that group tokens attend to in different stages.

PASCAL Context We show more qualitative results of GroupViT on the PASCAL Context dataset in Fig. 5. The PASCAL Context dataset annotates not only *object* classes from PASCAL VOC 2012, e.g. `car` and `dog`, but also *stuff* classes related to the context, e.g. `sky` and `water`. Observe that GroupViT successfully segments *object* and *stuff* classes in the PASCAL Context dataset, e.g., `cat` and `window` in the second row, and `dog` and `water` in the sixth row.

3. Additional Experiments and Analysis

Concepts Learnt by Group Tokens We visualize what the group tokens learn in Fig. 1. We select some group tokens and highlight their attended regions across images from the PASCAL VOC 2012 dataset. We find that the different group tokens learn different semantic concepts. In the first stage, group tokens usually focus on mid-level concepts such as “eyes” (row 1) and “limbs”(row 2). Interestingly, the group token 36 attends to “hands” if people are in the image, while focusing on “feet” if animals like bird and dog are present. Group tokens in the second stage are more

Figure 2. **Qualitative Results of GroupViT on PASCAL VOC 2012.** The results in columns labeled "Stage 1/2" show grouping results prior to assigning labels, where the regions belonging to the same group are indicated by the same color. All these examples contain a single object from a category.

Figure 3. **Qualitative Results of GroupViT on PASCAL VOC 2012.** The results in columns labeled "Stage 1/2" show grouping results prior to assigning labels. The regions belonging to the same group are indicated by the same color. These examples contain multiple objects from the same category.

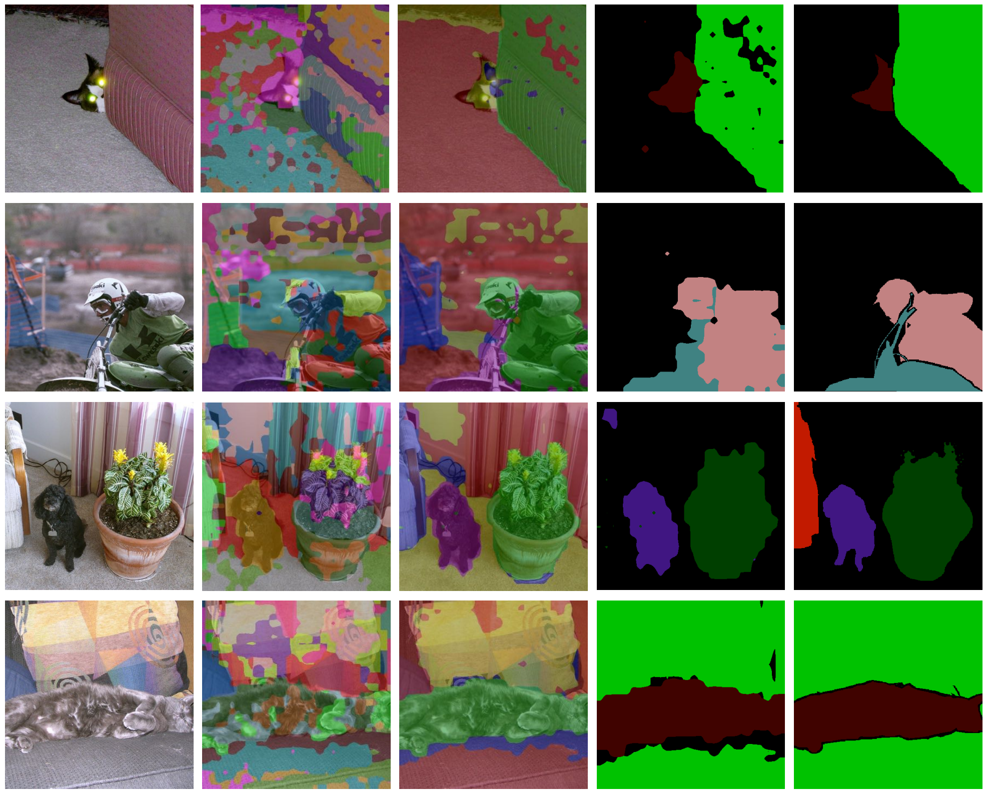

Figure 4. **Qualitative Results of GroupViT on PASCAL VOC 2012.** The results in columns labeled "Stage 1/2" show grouping results prior to assigning labels, where the regions belong to the same group are in the same color. These examples contain multiple objects from multiple different categories.

associated with high-level concepts, e.g., "grass", "body" and "face". Fig. 1 also shows that the learnt concepts in the first stage could be aggregated into higher level concepts in the second stage.

3.1. Image Classification

We compare the performances of the GroupViT and ViT architectures for the task of object classification on ImageNet. Following CLIP [10], here we train both architectures using supervision from text only via an image-text contrastive loss. In Table 1, we report both the zero-shot and the linear probing accuracy on the ImageNet [5] validation split. The zero-shot and linear probing evaluation follow the same setting as CLIP [10]. GroupViT's ImageNet classification performance is comparable to (if not better than) that of ViT, thus demonstrating that our proposed grouping mechanism enhances the baseline ViT architecture with the capability to perform semantic pixel grouping and zero-shot transfer to semantic segmentation, without affecting its object classification performance.

model	zero-shot Acc@1	linear Acc@1
ViT	42.4	69.2
GroupViT	42.9	69.8

Table 1. **ImageNet Accuracy**.

3.2. Mask Probing

We follow the procedure outlined in DINO [2] to evaluate the quality of the masks generated by GroupViT and by the baseline ViT model pre-trained using prior methods in a fully supervised [14], self-supervised [2, 4] or text-supervised [10] manner. For the ViT models, similar to DINO [2] for each final attention head, we compute its similarity to the `[CLS]` token and derive an attention mask for the pixels with the highest attention values. We then compute the Jaccard similarity of each head's attention mask to the ground truth mask and retain the attention mask with the highest similarity. As for GroupViT, it does not have a multi-head design in the Grouping Block. Thus, we directly select the group most similar to the ground truth, as measured by the Jaccard index for each image. As Table 2 shows, the mask probing result of GroupViT is significantly better than that of all variants of the baseline ViT architecture. Hence, compared to ViT, our GroupViT more effec-

Figure 5. **Qualitative Results of GroupViT on PASCAL Context.** Columns labeled “Stage 1/2” show grouping results prior to assigning labels, where the regions belonging to the same group are indicated by the same color. GroupViT can successfully segment *object* and *stuff* classes, e.g. `cat` and `window` in row 2, `dog` and `water` in row 6.

arch	model	dataset	supervision	Jaccard Similarity
ViT	Random	VOC	-	23.6
ViT	DeiT [14]	ImageNet	class	24.6
ViT	MoCo [4]	ImageNet	self	28.2
ViT	DINO [2]	ImageNet	self	45.9
ViT	DINO [2]	CC+YFCC	self	41.8
ViT	CLIP [10]	CC+YFCC	text	28.6
GroupViT	Ours	CC+YFCC	text	51.8

Table 2. **Comparison of mask probing performance** GroupViT outperforms all other variants of the baseline ViT architecture at effectively grouping image regions on semantic groups.

tively groups semantically-related visual inputs together.

3.3. Limitations

We find that the mIoU of GroupViT on PASCAL Context is significantly lower than that on PASCAL VOC 2012. This could be attributed to the presence of background classes in PASCAL Context, e.g., `ground`, `road` and `wall` that result in low IoU ($\sim$1.5) on zero-shot transferring GroupViT to semantic segmentation on PASCAL Context. Through visual inspection, we find that while the pixels belonging to these background classes are typically correctly grouped into a single group by GroupViT, the group as a whole may be miss-classified into the wrong class on being compared to the text embedding of the various class labels. We hypothesize that this, in turn, happens due to the low probability of the background classes being described in textual sentences used during training. We show examples of such failure case in Fig. 6. We further conduct an oracle experiment to verify this finding. In the oracle experiment, for each output group from GroupViT, we compute its IoU with all ground truth masks and assign to each group the class label that results in the the maximum IoU. This represents the upper bound of GroupViT's performance since here we leverage ground truth masks to predict each group's class label. We use our 2-stage GroupViT trained on the CC and YFCC datasets for this oracle experiment, which is the same model labeled "Ours" in Table 5 of the main paper. We report the oracle experiment's results on PASCAL Context in Table 3. The large gap between the performance of the original and oracle mIoU values on the PASCAL Context dataset, shows that while GroupViT's grouping results are reasonably good, there is room to further improve the groups' classification to segmentation class labels via image-text embedding similarity computation.

arch	mask mIoU	oracle mask mIoU
GroupViT	22.4	54.6

Table 3. **Original versus oracle results on PASCAL Context**.

3.4. COCO Dataset

We evaluate the performance of GroupViT on the COCO dataset [9], which contains 80 object classes. We combine the instance masks of the same category to get the semantic segmentation masks for each image. We report semantic segmentation mIoU on COCO in Table 4. It demonstrates that GroupViT is able to transfer to complex datasets with various number of classes.

arch	mask mIoU
GroupViT	24.3

Table 4. **Results on COCO Dataset**.

3.5. Training on RedCaps

To show that our approach is generalizable to other training datasets, besides CC [3, 11] and filtered YFCC [12], we also train GroupViT on the recently released RedCaps dataset [6], which contains 12 millions image-text pairs from Reddit, of similar size as filtered YFCC. We report mIoU for zero-shot transfer to various image segmentation benchmarks datasets in Table 5. Replacing YFCC with RedCaps yields similar accuracy on Pascal VOC, Pascal Context and COCO datasets. It demonstrates that GroupViT is able to learn grouping with properly filtered image text pairs.

arch	Training Dataset	PASCAL VOC	PASCAL Context	COCO
GroupViT	CC+YFCC	52.3	22.4	24.3
GroupViT	CC+RedCaps	50.8	23.7	27.5

Table 5. **Results trained with CC+RedCaps**.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 4, 6

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *CVPR*, 2021. 6

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 4, 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[6] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS*, 2021. 6

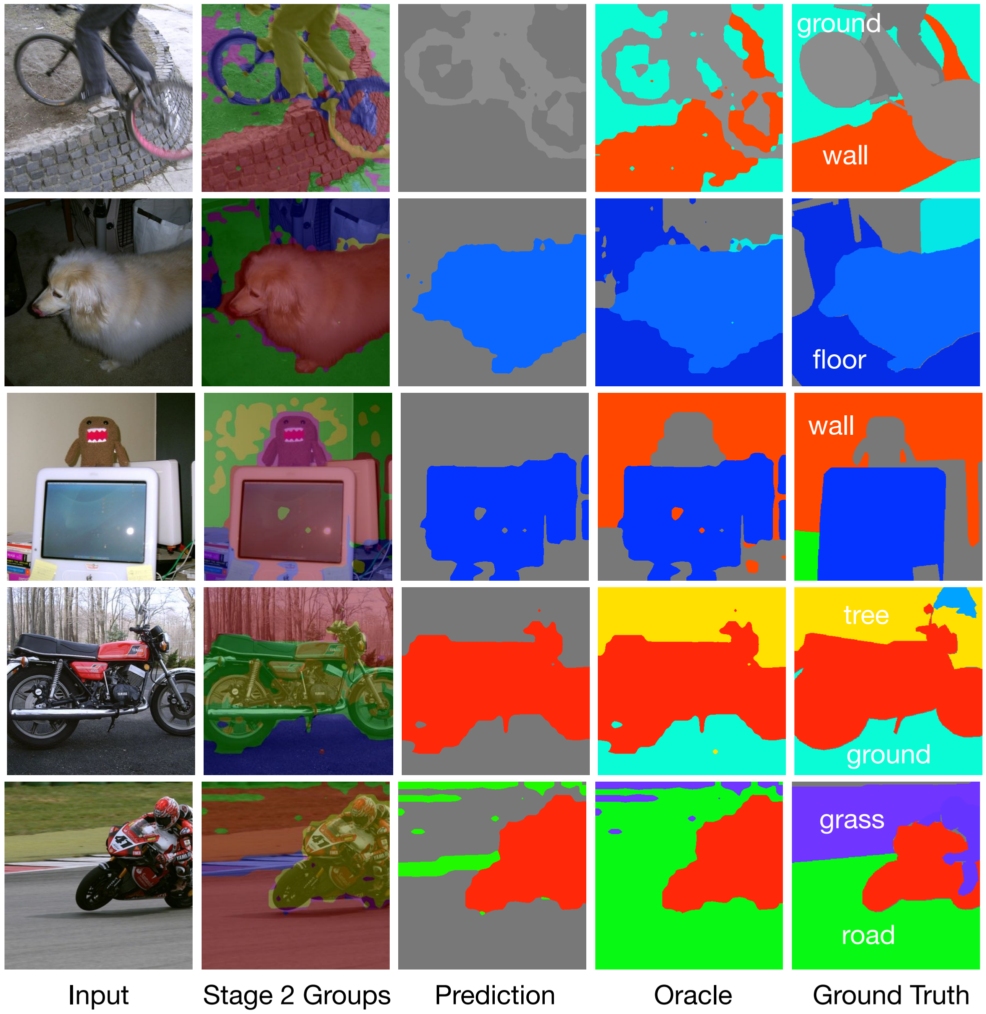

Figure 6. **Failure cases on PASCAL Context.** "Oracle" shows the results of assigning groups to segmentation classes based on their IoU with the ground truth masks. Although GroupViT successfully groups *stuff* classes, e.g. `ground`, `road` and `wall`, it is not able to classify them correctly using the similarity between the visual and text embedding.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 4, 6

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 6

[12] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6

[13] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, 2021. 1

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 4, 6