

Domain Generalization for Vision Perception Models by Camera-Lidar Contrastive Learning

(Master's Thesis)

Nils Golombiewski

advised by Dr. Kira Maag and Prof. Dr. Hanno Gottschalk
with support from Christoph Hümmer and Manuel Schwonberg

Technical University Berlin

September 5, 2024

Domain Generalization

Train a model on a *source domain* \mathcal{D}_S and evaluate its performance on a *target domain* $\mathcal{D}_T \neq \mathcal{D}_S$. The model never learns on the target domain.

Domain Generalization

Train a model on a *source domain* \mathcal{D}_S and evaluate its performance on a *target domain* $\mathcal{D}_T \neq \mathcal{D}_S$. The model never learns on the target domain.

In order to evaluate our model, we need to specify a task and a performance measure. (In our case: Semantic Segmentation, see below.)

Distribution Shift Examples



Cityscapes



ACDC

Why care about Domain Generalization?

(Specifically for autonomous driving.)

Why care about Domain Generalization?

(Specifically for autonomous driving.)

- Target domain is too large \rightarrow train on subdomain (e.g. specific localities)

Why care about Domain Generalization?

(Specifically for autonomous driving.)

- Target domain is too large \rightarrow train on subdomain (e.g. specific localities)
- Events too rare (e.g. rare weather phenomena or traffic situations)

Why care about Domain Generalization?

(Specifically for autonomous driving.)

- Target domain is too large \rightarrow train on subdomain (e.g. specific localities)
- Events too rare (e.g. rare weather phenomena or traffic situations)
- Data distribution shifts over time

Semantic Segmentation

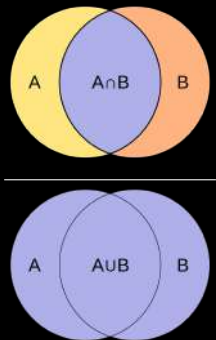
Computer vision task of assigning a semantic class label to each pixel of a 2D input image. No distinction between different instances of the same class.

Semantic Segmentation

Computer vision task of assigning a semantic class label to each pixel of a 2D input image. No distinction between different instances of the same class.

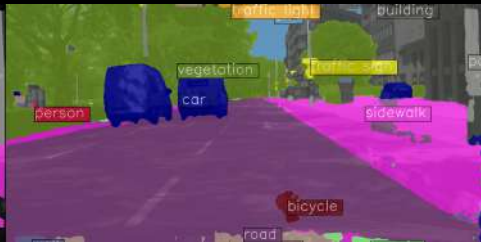


(Mean) Intersection over Union


$$\text{IoU}(A, B) = \frac{\text{Area of } A \cap B}{\text{Area of } A \cup B}$$

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^K \text{IoU}_k$$

Example: Ground Truth vs. Model Prediction



CLIP

Radford et al.: Learning transferable visual models from natural language supervision, 2021. (CLIP)

CLIP

Radford et al.: Learning transferable visual models from natural language supervision, 2021. (CLIP)

- Learn image representations from 400 million (image, text) pairs

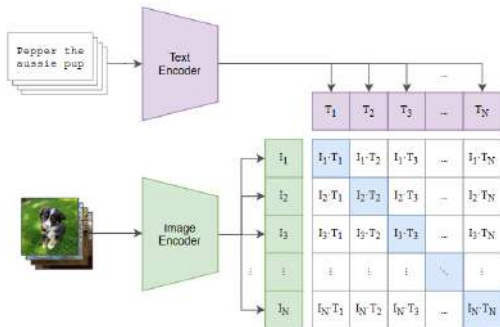
CLIP

Radford et al.: Learning transferable visual models from natural language supervision, 2021. (CLIP)

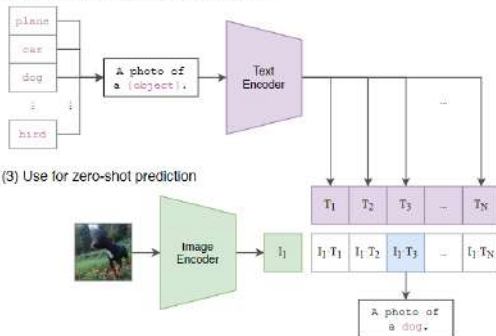
- Learn image representations from 400 million (image, text) pairs
- Reference learned visual concepts via natural language for zero-shot transfer to downstream tasks
- Evaluate performance on over 30 computer vision datasets and various tasks

CLIP Method

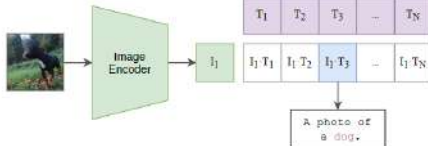
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Strengths of CLIP

- Generalizes well to most tasks, e.g. matches original ResNet-50 on ImageNet zero-shot without seeing any of its 1.28 million training examples.
- Self-supervised: Learns representations from raw, unannotated data.
- Exhibits superior domain generalization to SOTA supervised models.

Multimodal Contrastive Pretraining

Hümmer et al. 2023 (VLTSeg) shows that CLIP image encoder can improve domain generalization for semantic segmentation.

Multimodal Contrastive Pretraining

Hümmer et al. 2023 (VLTSeg) shows that CLIP image encoder can improve domain generalization for semantic segmentation.

Hypothesis

Multimodal pretraining can improve domain generalization, provided the co-modality is robust under distribution shift.

Multimodal Contrastive Pretraining

Hümmer et al. 2023 (VLTSeg) shows that CLIP image encoder can improve domain generalization for semantic segmentation.

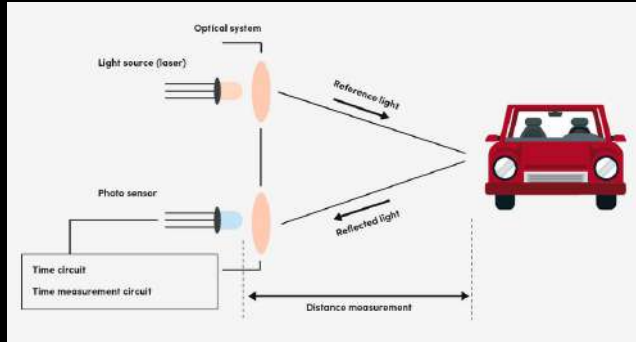
Hypothesis

Multimodal pretraining can improve domain generalization, provided the co-modality is robust under distribution shift.

We test this hypothesis by adapting CLIP for camera-lidar contrastive pretraining.

Lidar (Light Detection and Ranging)

Lidar devices emit laser pulses and measure the time for reflected light to return: Distance measurement via constant speed of light.



Lidar

As a co-modality we use lidar (*light detection and ranging*) because it is

- (partially) invariant to distribution shifts,
- a common sensor in autonomous vehicles.

Lidar

As a co-modality we use lidar (*light detection and ranging*) because it is

- (partially) invariant to distribution shifts,
- a common sensor in autonomous vehicles.

Each camera image in our training dataset (A2D2) comes with a corresponding *point cloud* (set of 3D points), obtained by a lidar device.

Mapping Lidar Point Cloud onto Camera Image



Approach

- 1 **Multimodal pretraining** an image encoder for alignment of pairs of camera images and lidar point clouds by optimizing a contrastive loss equivalent to CLIP.

Approach

- ① **Multimodal pretraining** an image encoder for alignment of pairs of camera images and lidar point clouds by optimizing a contrastive loss equivalent to CLIP.
- ② **Finetuning** the image encoder for the task of semantic segmentation.

Approach

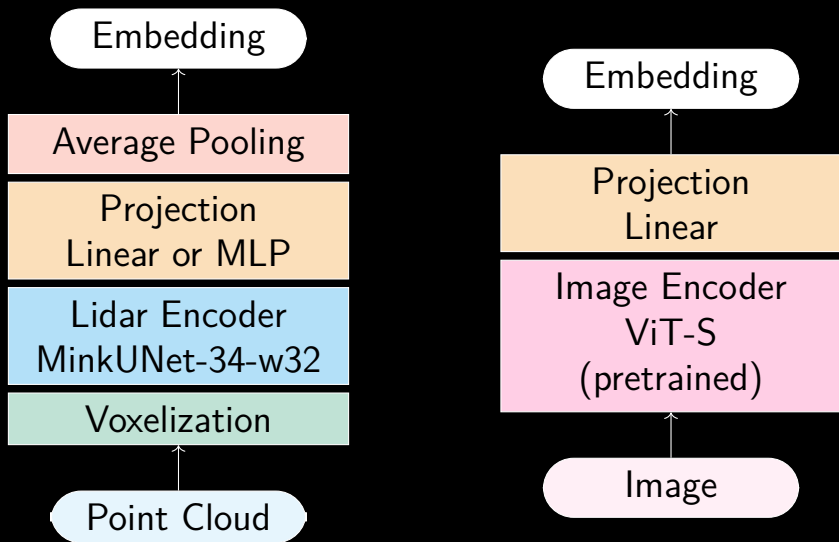
- ① **Multimodal pretraining** an image encoder for alignment of pairs of camera images and lidar point clouds by optimizing a contrastive loss equivalent to CLIP.
- ② **Finetuning** the image encoder for the task of semantic segmentation.
- ③ **Evaluation** of semantic segmentation performance on an unseen target domain to measure domain generalization.

Approach

- ① **Multimodal pretraining** an image encoder for alignment of pairs of camera images and lidar point clouds by optimizing a contrastive loss equivalent to CLIP.
- ② **Finetuning** the image encoder for the task of semantic segmentation.
- ③ **Evaluation** of semantic segmentation performance on an unseen target domain to measure domain generalization.

We compare this model to an image encoder that was only pretrained (supervised) for classification on ImageNet.

Camera-Lidar Contrastive Pretraining (CLCP)



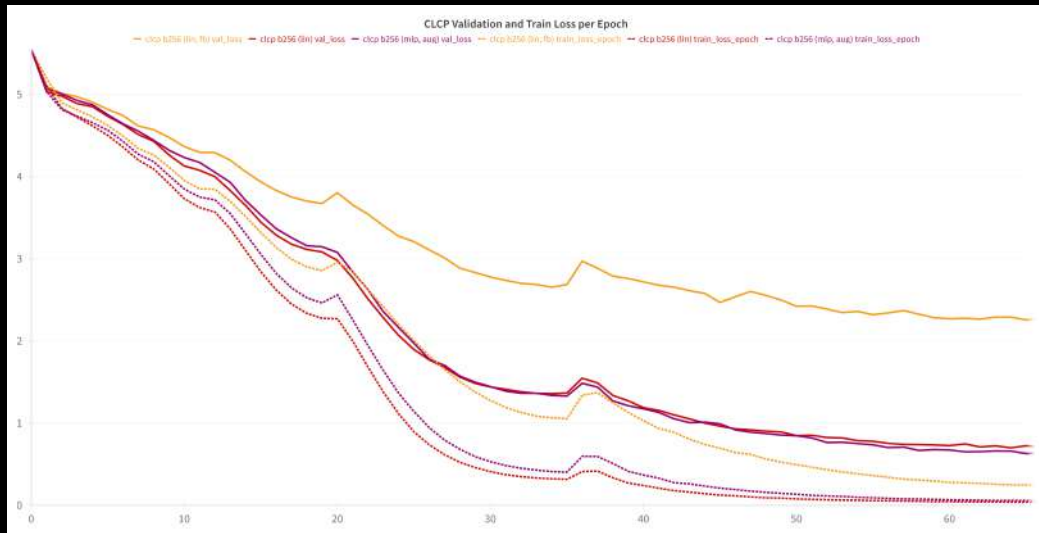
Loss

For a batch of size N , we compute N^2 similarity scores between image and lidar embeddings z_i^I and z_j^L .

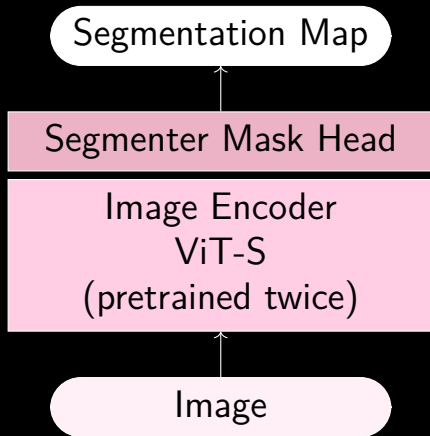
$$s_{i,j} = \frac{z_i^I \cdot z_j^L}{\|z_i^I\| \|z_j^L\|} \quad (\text{similarity scores})$$

$$\text{Loss} = \frac{1}{2N} \sum_{i=1}^N \left(-\log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^N \exp(s_{i,j}/\tau)} - \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^N \exp(s_{j,i}/\tau)} \right)$$

Some Pretraining Loss Curves



Finetuning for SemSeg on Cityscapes



Results

Experiment	mIoU (Cityscapes)	mIoU (ACDC)
timm-ViT-S	68.01	37.93
CLCP Ep19 (Lin, fb)	62.30	31.06
CLCP Ep35 (Lin, fb)	59.21	27.21
CLCP Ep19 (Lin)	63.52	33.48
CLCP Ep35 (Lin)	60.58	31.07
CLCP Ep67 (Lin)	58.11	29.28
CLCP Ep19 (MLP, aug)	62.35	30.57
CLCP Ep35 (MLP, aug)	59.32	28.94
CLCP Ep67 (MLP, aug)	57.11	28.02

Conclusions

Our experiments don't support the hypothesis.

Conclusions

Our experiments don't support the hypothesis.

Possible reasons:

- 1 Camera-lidar alignment is not useful (enough) with respect to segmentation task (as opposed to natural language supervision).

Conclusions

Our experiments don't support the hypothesis.

Possible reasons:

- ① Camera-lidar alignment is not useful (enough) with respect to segmentation task (as opposed to natural language supervision).
- ② Amount and diversity of pretraining dataset is not sufficient to increase domain generalization.

Discussion

From Fang et al: Data Determines Distributional Robustness in Contrastive Language-Image Pre-training (CLIP), 2022:

“Our main result is that CLIP’s robustness is determined almost exclusively by the training distribution. Language supervision at training time does not make the resulting models more robust than standard supervised learning when the images in the training set are the same.”

Outlook

Is vision-only multimodal learning a dead-end?

Outlook

Is vision-only multimodal learning a dead-end?

- Self-supervised representation learning still allows pretraining on much larger scale.

Outlook

Is vision-only multimodal learning a dead-end?

- Self-supervised representation learning still allows pretraining on much larger scale.
- Further investigation should reveal whether camera-lidar can improve over camera-only pretraining.

Outlook

Is vision-only multimodal learning a dead-end?

- Self-supervised representation learning still allows pretraining on much larger scale.
- Further investigation should reveal whether camera-lidar can improve over camera-only pretraining.
- Same holds for vision-only vs. vision-language pretraining.

Outlook

Is vision-only multimodal learning a dead-end?

- Self-supervised representation learning still allows pretraining on much larger scale.
- Further investigation should reveal whether camera-lidar can improve over camera-only pretraining.
- Same holds for vision-only vs. vision-language pretraining.