THU-ACV Final Project

1 Point distribution

At most 3 students per team, and you can solo the project if you like.

Proposal: 30% Presentation: 20% Final Report and Codebase: 50%

2 Project Structure

Your team need to prepare a proposal, a presentation and a final report on one of the provided topics.

2.1 Topics Decision

On March. 24 we will have an optional tutorial. You can come to discuss with TAs about the topics. You need to choose a topic and group up a team by March. 25.

2.2 Proposal

Based on your final project topics, we assigned a TA-in-charge for each of you. You can discuss with your TA-in-charge. BUT they will NOT help with debugging, thinking the main bulk of your idea, writing paper, etc. It is your own responsibility to conduct the WHOLE pipeline.

On Mar. 31 we will have an optional tutorial. You can come to discuss with your TA-in-charge about the proposal. Please send your final project proposal to your TA-in-charge before April. 8 .

The proposal should be 1-2 pages, including introduction to the task, a survey of related works, a brief description of your idea or method, and the tentative timeline.

Your TA will give you some feedback on the proposal.

The email address of TAs:

simachonghao@pjlab.org.cn

hanj19@mails.tsinghua.edu.cn

tianchangyao@sensetime.com

2.3 Presentation

On May. 12 we will have an optional tutorial. You can come to discuss with your TA-in-charge about the presentation. You are required to submit the draft presentation slides before May. 20. You may make some minor modification on the slides before the day of presentation. The project presentation will be held on the whole day of May. 28.

2.4 Final Report and Codebase

The template for your final report is on the website (4 pages excluding references). You should submit it with your codebase befor June. 11.

Paper Report of Final Project for Advanced Computer Vision.

3 Timeline

All deadline are due at 23:59, Beijing local time, on that day.

Mar. 18: Release Topics

Mar. 25: Deadline of topics and teams decision

Apr. 8: Deadline of proposal submission

May. 20: Deadline of draft presentation submission

May. 28: Presentation Day

June. 11: Final report and codebase.

4 Topics

You are encouraged to select one of the topics we provided below, and formulate it into an **achievable and solid** research topic. Topics in **Vision X** are more challenging, and you can try them if you are interested (they also need concrete problem formulations).

4.1 3D Perception

- Monocular Lane/Object Detection
- Lidar Segmentation
- Depth Estimation
- Sensor Fusion
- BEV Representation
- Connecting Industry to Academia in Autonomous Driving through Evaluation Metrics

In recent years, there has been tremendous progress on 3D vision for analysis and understanding of 3D data, such as 3D semantic segmentation, 3D object detection and tracking. These advances have however not yet translated to significant progress in several fundamental challenges for the domain of robotics. Active perception in static and dynamic environments, inference of spatial relations in 3D scenes, activity recognition, and behavior prediction in real-world settings are a few examples of challenging robotics problems. To successfully tackle these problems, we should leverage the inherent 3D nature of the physical world, and apply deep learning approaches that learn 3D representations that are robust to input perturbation and generalize to real-world variations with high sample efficiency (e.g., transformation invariance).

Suggested dataset: KITTI, OpenLane 300 sequence

Suggested baseline: SMOKE, 3D/Gen-LaneNet, MonoDepth, Cylinder3D

Suggested Papers:

1. Garnett N, Cohen R, Pe'er T, et al. 3d-lanenet: end-to-end 3d multiple lane detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2921-2930.
2. Liu Z, Wu Z, Tóth R. Smoke: Single-stage monocular 3d object detection via keypoint estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 996-997.
3. Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 270-279.
4. Zhou H, Zhu X, Song X, et al. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation[J]. arXiv preprint arXiv:2008.01550, 2020.

4.2 Adversarial Attack: Robustness in Deep Learning

- 2D image classification
- 2D object detection

Adversarial attacks are the phenomenon in which machine learning models can be tricked into making false predictions by slightly modifying the input. Most of the times, these modifications are imperceptible and/or insignificant to humans, ranging from colour change of one pixel to the extreme case of images looking like overly compressed JPEGs. But it turns out even state of the art models can produce highly confident yet bizarre results to such altered inputs.

This vulnerability gives rise to two serious questions to the current state of AI:

1. Are machine learning models actually understanding the abstract ideas and conceptual hierarchy of our world as we would like them to, or are they relying on statistical nuisances of the inputs to make predictions?
2. Can we safely and reliably deploy our models in the production environment without the risk of exploitation and unintended consequences?

Discovered by Szegedy, adversarial attacks have become a major avenue of research in machine learning. The main worrying attributes of adversarial attacks are:

1. **Imperceptibility**: Adversarial examples can be generated effectively by adding small amount of perturbations or even by just slightly modifying the values along limited number of dimensions of the input. These subtle modification makes them almost impossible to be detected by humans, but the models classify them incorrectly with high confidence challenging our understanding of how the model synthesise inputs, focus attention and learn semantics.
2. **Targeted Manipulation**: Attack samples can be generated in a way that manipulates the model to output the exact incorrect class as intended by the adversary. This opens up the possibility of severe manipulation of the system to one's gain instead of simply breaking it.
3. **Transferability**: Adversarial examples generated for one model can deceive networks with even different architectures trained on the same task. Even more surprisingly, these different models often agree with each other on the incorrect class. This property allows attackers to use a surrogate model (not necessarily the same architecture or even the same class of algorithm) as an approximation to generate attacks for the target model (also known as oracle).
4. **Lack of theoretical model**: There are currently no widely accepted theoretical models on why adversarial attacks work so effectively. Several hypothesis have been put forward such as linearity, invariance and non-robust features leading to several defence mechanisms, but none of them have acted as a panacea for coming up with robust models and resilient defences.

Suggested Dataset: CIFAR-10; Tiny ImageNet; COCO

Suggested Baseline: Cleverhans(including FGSM and a set of classical attack methods)

Suggested Papers:

1. Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
2. Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016: 372-387.
3. Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

4.3 Transfer, few-shot, semi- and self-supervised Learning

- Person Re-Identification
- Object detection

4.3.1 Unsupervised Domain Adaptive Person Re-identification

Person re-identification (re-ID) aims at retrieving the same persons' images from images captured by different cameras. However, even with large-scale datasets, for person images from a new camera system, the person re-ID models trained on existing datasets generally show evident performance drops because of the domain gaps. Unsupervised Domain Adaptation (UDA) is therefore proposed to adapt the model trained on the source image domain (dataset) with identity labels to the target image domain (dataset) with no identity annotations.

1. Noise pseudo labels. State-of-the-art UDA methods for person re-ID group unannotated images with clustering algorithms and train the network with clustering generated pseudo labels. Although the pseudo label generation and feature learning with pseudo labels are conducted alternatively to refine the pseudo labels to some extent, the training of the neural network is still substantially hindered by the inevitable label noise. The noise derives from the limited transferability of source-domain features, the unknown number of target-domain identities, and the imperfect results of the clustering algorithm. The refinery of noisy pseudo labels has crucial influences to the final performance.
2. Large domain gap. UDA re-ID is to transfer the knowledge from the labeled source domain to improve the model's discriminability on the unlabeled target domain. It is a challenging problem because the source and target domains can have two extreme distributions, and there can be no overlap between the two domains' label space.

You are free to study the former two questions of UDA pserson re-ID. Or you can study other relevent questions about unsupervised/semi-supervised learning.

Suggested Dataset: Market1501, MSMT17

Suggested Baselines: MMT, IDM

Suggested Papers:

1. Dai, Yongxing, et al. "Idm: An intermediate domain module for domain adaptive person re-id." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
2. Ge, Yixiao, Dapeng Chen, and Hongsheng Li. "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification." arXiv preprint arXiv:2001.01526 (2020).
3. Zheng, Kecheng, et al. "Group-aware label transfer for domain adaptive person re-identification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

4.3.2 Few-Shot Object Detection

Compared to the ability of human to quickly extract novel concepts from few examples, deep models are still far from satisfactory. Few-shot object detection (FSOD) is a much more challenging task than both few-shot classification and object detection. At present, most FSOD approaches prefer to follow the meta-learning or fine-tuning paradigm.

1. Meta-learning methods aims to acquire more task-level knowledge and generalize better to novel classes. You may propose new meta-learning methods which generalize well on novel classes with fast adaption.
2. Finetune-based methods are very simple and efficient. By adopting a two-stage fine-tuning scheme, this series is comparable to meta methods. Yet, due to most parameters are pre-trained on base domain and then frozen on novel set, they may fall down the severe shift in data distribution and underutilization of novel data.
3. Regardless of the meta-based or finetune-based method, Faster R-CNN has been widely used as the basic detector and achieved good performance. However, its original architecture is designed for conventional detection and lacks of tailored consideration for few-shot scenario.

You are free to delve deep into the former questions of few-shot object detection. Or you can study other relevent questions about few-shot learning.

Suggested Dataset: VOC07+12, MS COCO

Suggested Baselines: Meta RCNN FsDet

Suggested Papers:

1. Kang, Bingyi, et al. "Few-shot object detection via feature reweighting." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
2. Yan, Xiaopeng, et al. "Meta r-cnn: Towards general solver for instance-level low-shot learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
3. Wang, Xin, et al. "Frustratingly simple few-shot object detection." arXiv preprint arXiv:2003.06957 (2020).

4.4 General Vision

- Image caption
- Long-tailed learning
- Multi-task Learning
- Rethinking Backbone for downstream task
- Prompt Learning

4.4.1 Long-tailed learning.

Compared to standard datasets, real-world data often exhibits a long-tailed distribution. Take image classification as an example, the head classes usually possess much more sample points than the tail ones. Such distribution bias could lead to a dramatic decrease in model performance.

Previous methods tried to solve this problem mainly from three different perspectives, which are:

1. Class Re-balancing. Can we explicitly boost the model's attention to tail classes by re-sampling or adjusting its original output logits?
2. Information Augmentation. Maybe it's time for us to add more external information through transfer learning, data augmentation or knowledge distillation, to help model generalize better on tail ones.
3. Module Improvement. Should the model's architecture be redesigned or optimized to better capture the inter- and intra- class variance?
4. Or you can come up with more impressive ideas!

Suggested Datasets: Places-LT, CIFAR-100-LT, ImageNet-LT

Suggested Baselines: OLTR, CB-LWS

Suggested Papers:

1. A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in International Conference on Learning Representations, 2021.
2. B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, “Decoupling representation and classifier for long-tailed recognition,” in International Conference on Learning Representations, 2020.
3. Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, “Large-scale long-tailed recognition in an open world,” in Computer Vision and Pattern Recognition, 2019, pp. 2537–2546.
4. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

4.4.2 Image Captioning

As a classic and challenging task, image captioning requires models capable of both visual and linguistic understanding. Although the standard encoder-decoder structure based on CNN and RNN (or even transformers) has become today's mainstream paradigm, however, there still exists some other challenges remained to be further explored, such as:

1. Diverse Captioning. How to generate more diverse captions instead of simply following the ground truth formats? And how can we better measure such diversity quantitatively?
2. Novel Object Captioning. What if we ask the model to generate captions for novel objects which is not included in the training set? How to increase the model's generalization power?
3. Controllable Captioning. Can we make the model generate desired captions under our human's control, which means the generation process may need to be manually intervened?
4. Or we are looking forward to your own questions and answers!

Suggested Dataset: COCO Captions, Nocaps, Flicker30k

Suggested Baselines: SAT

Suggested Papers:

1. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015.
2. R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in ICCV, 2017.
3. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided Open Vocabulary Image Captioning with Constrained Beam Search," in EMNLP, 2017.
4. A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra, "Diverse Beam Search for Improved Description of Complex Scenes," in AAAI, 2018.
5. M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "MeshedMemory Transformer for Image Captioning," in CVPR, 2020.

4.5 Vision X

- Application in real-life
- Rethinking ImageNet, what's next?
- Medical Imaging
- Domain Adaptation
- Auto Labelling
- Differentiable computer vision, graphics and physics
- Vision QA system
- Beyond BackPropagation: novel ideas for training NN
- Efficient NERF
- Object-Oriented Learning: Perception, Representation, and Reasoning
- Vision-Language Models
- Multimodal Perception
- Video Understanding, Object Tracking
- Learning from Limited or Imperfect Data