List of Master Thesis projects proposed by TomTom, 2020:
For more information please contact mohsen.ghafoorian@tomtom.com.

A study for improving generation fidelity of VAEs:

Generative modeling is a task that ml-researchers have tried to solve for a very long time. Among the currently published techniques, GANs [1] and VAEs [2] are more interesting than the other two i.e. Autoregressive and Flow-based because of their slow inference and slow training times respectively. Both GANs and VAEs have their pros and cons. But, in terms of the sheer fidelity of the generated images, GANs happen to be quite ahead in the game.
VAEs, on the other hand, are way more useful because of the ability to do controlled latent-space exploration (mainly allowed because of the presence of an Encoder) and also the ability to numerically compute the likelihood of validation samples at runtime; which is something that GANs don't allow. So, it is quite logical and interesting to do a thorough study focused on improving the generation quality of VAEs.

There was a time when GANs were only able to generate low-resolution images stably, but this changed completely when the Progressive Growing of GANs [3] study was put forth. This paper also introduced a bunch of novel hacks (apart from the progressive growing of course) which were motivated from image/signal/graphics processing point of view. This research was especially important because it showed that by using a number of simple tricks GANs could be made to generate `1024 x 1024` resolution images without using any complicated mathematical concepts which is exactly where the GAN research was headed in general.

A strikingly similar parallel can be drawn with VAEs' current state. Complicated hacks such as the Vector Quantization or Hierarchical latent space [6] or Fitting a prior to a learned posterior [4] are introduced to increase the generated resolution and to improve the quality of the VAE images. As a part of this study, first of all, we would like to evaluate if the graphics-based techniques introduced by ProGANs are still transferable to VAEs. Post this initial study, we plan to brainstorm new techniques that build upon the obtained results and are simple in nature targeted to solve the aforementioned problem/s.

In nutshell, the main goal here is to obtain generated sample quality even better than the current SOTA GAN without using vector quantization in VAEs especially with a focus on bringing the high-frequency details in the generated samples. Although, a very recent work, [5] is a formidable attempt at solving the same problem, they are still tied to the hierarchical latent space which could be simplified greatly and they still lack high-frequency details in the generated samples.

Datasets: FFHQ (standard), CelebA-HQ (standard), Imagenet (for the biggest scale experiments)

References:

[1] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.
[2] Kingma, Diederik P., and Max Welling. "Auto-encoding variational Bayes." arXiv preprint arXiv:1312.6114 (2013).
[3] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).
[4] Razavi, Ali, Aaron van den Oord, and Oriol Vinyals. "Generating diverse high-fidelity images with VQ-VAE-2." Advances in Neural Information Processing Systems. 2019.
[5] Vahdat, Arash, and Jan Kautz. "NVAE: A Deep Hierarchical Variational Autoencoder." arXiv preprint arXiv:2007.03898 (2020).
[6] Kingma, Diederik P., et al. "Improving variational inference with inverse autoregressive flow.(nips), 2016." URL http://arxiv. org/abs/1606.04934.

Learning multi-view registration

Pose estimation for a set of perspective images (or point clouds) is a critical building block for applications like HD Map Making and 3D Scene Reconstruction. Most multi-view registration methods follow a two-step procedure:

1. Estimate the relative poses between every pair of images (or point clouds)
2. Given the relative poses from step 1, estimate, up to a free transformation, the absolute poses for each image, a.k.a., pose-graph optimization

Recent works [1] [2] employ deep neural network for relative pose prediction. But not many works put emphasis on integrating neural networks to a multi-view registration pipeline.

Gojcic et. al.[3] proposed an end-to-end learnable pipeline for multi-view point clouds registration. The proposed pipeline solves the pose-graph optimization problem using the spectral relaxation algorithm. Spectral relaxation algorithm is an approximate method. It over-parameterizes rotation with the rotation matrix and optimize under relaxed constraints.

Instead of parametrizing 3D rotation with a rotation matrix, one might also parametrize with quaternions. The pose-graph optimization is then formed as a non-lieanr least square problem and can be solved with Gauss-Newton algorithm and its variants. [4]

There could be multiple directions on improving the current state-of-the-art learnable methods:

- **Pair-wise pose estimation**
  One can think of training a siamese neural network to output a distribution over the manifold of relative poses instead of a single optimal pose. The pose-graph optimization stage could benefit from the added information from a distribution.
- **Differentiable pose-graph optimization**
  Another direction is to implement a differentiable Gauss-Newton method and solving the pose-graph optimization problem in a less over-parameterized way than the spectral relaxation algorithm.

Datasets

There are many public datasets available for both perspective images and point clouds, including 3DMatch, ScanNet, Argoverse, FordAV

References

1. CLKN: Cascaded Lucas-Kanade Networks for Image Alignment↩
2. PointNetLK: Robust & Efficient Point Cloud Registration using PointNet↩
3. Learning multiview 3D point cloud registration↩
4. g²o: A General Framework for Graph Optimization↩

Mapping without registered imagery

To create High Definition (HD) Map that fully covers a big scene, for example, a junction or a wide road, traditional methods:

- either first aggregate all relevant images and then extract semantic features
- or first extract semantic features from individual images and then fuse extracted features

However, both methods rely heavily on accurate pose estimation for the input images.

- On one hand, accurate camera poses are difficult to maintain on a large scale due to bad calibration, GNSS errors, accumulated drifts, etc.
- On the the other hand, neural networks do not necessarily require perfectly aligned sources as inputs. [1] [2]

In this project, we will address the problem of extracting HD Map features from overlaping but not necessarily well-aligned perspective images or birds-eye-view images. More specifically, we will explore training neural networks to perform this task end-to-end.

Recent works propose methods learning semantic maps in birds-eye-view directly from multiple perspective images [3]. These works often assume that accurate camera poses are given. However, in real data, camera poses are not always as accurate as the poses provided in those small-scale benchmark datasets.

In this project, we will instead assume that the input images are only roughly aligned. This assumption comes from the fact that GNSS information are widely available. But it is a chalenging task even for the high-end GNSS positioning system that most mobile mapping vehicles equiped with to achieve center-meter level absolute accuracy.

NuScenes, Argoverse and Berkly interaction are possible datasets to be used.

References

1. OverlapNet: Loop Closing for LiDAR-based SLAM↩
2. AutoCorrect: Deep Inductive Alignment of Noisy Geometric Annotations↩
3. Predicting Semantic Map Representations From Images Using Pyramid Occupancy Networks↩

Improving efficiency and expressiveness of implicit 3D generative models

Generative modeling techniques allow us to address the shortcoming of supervised machine learning, i.e. requiring a large amount of labelled samples, because they enable the creation of realistic synthetic samples from the real data. As Sir Richard Feynman famously said, “what I cannot create, I do not understand”; not only do generative models have direct applications in image synthesis, they also allow us to learn complex real world phenomena entirely from self-supervised data. This may be a key to improving not just generative, but discriminative tasks as well. Majority of the existing Generative Modeling techniques are in 2D image generation space while the field of 3D deep learning is only taking off. There are many interesting areas in 3D deep learning where generative modeling could be very useful.

In particular, the implicit models for 3D scenes (or any form of data in general) [1, 2, 3, 4, 5, 6] are currently being researched quite a lot since they have the advantage that the model sizes do not scale with the number of samples of the scene, but with the complexity of the scene being modeled. Moreover, these representations are much closer to our reality in the sense that objects around us are continuous while we perceive discrete samples from them. However, these models are in their infancy and are far from being able to be deployed in applications. Thus the overall goal of this study would be to bring these implicit models into real-time/world applications.

The following research directions are being proposed for this study:

- Improving the efficiency of the implicit models:
  - Training efficiency: these implicit models currently need time of the order of minutes - hours to train on individual scenes. Although [7] has been an attempt to reduce this time, the fidelity of this shortly-trained model doesn't match the fully trained model. There is still a big scope of improvement and a plethora of research ideas to explore to make these models train faster (real-time training being the ultimate goal).
  - Inference efficiency: inference with these models is still far from being ready for real-time applications. The curent state-of-the-art NeRF [3] for instance, takes ~1-2 mins to render a single 1K image from a novel viewpoint. Researching and engineering techniques to make the rendering of these models real time is not only interesting, but has a number of applications too. A very recent work [8] is a good attempt at solving this bottleneck, but there is still a big scope of improvement.
- Improving the expressiveness of the implicit models:
  All these `implicit model` methods still train one model per scene which is highly in-efficient; in spite of being way more efficient than the sample based representations. Although [7, 1] have shown preliminary results with their hypernetworks, there is tremendous work that can be done in this space. Explicitly saying, the goal is to make these models generalize to multiple scenes for scenes that share content or otherwise. A formidable solution to this problem then opens up another really interesting question. Let's say we are able to generalize these models to multiple scenes, what do we get by interpolating between any two scenes that the models creates. This question paves way to solving the longest standing problem of generative modelling. The problem of compositionality. Being able to compose these implicit models will have innumerable applications in procedural generation.

Datasets: Synthetic-Nerf, Scannet (for efficiency direction and expressiveness direction in non content sharing case). Maria Sequence (for expressiveness direction in content sharing case).

References:

[1] Sitzmann, Vincent, et al. "Implicit Neural Representations with Periodic Activation Functions." arXiv preprint arXiv:2006.09661 (2020).
[2] Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." arXiv preprint arXiv:2006.10739 (2020).
[3] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." arXiv preprint arXiv:2003.08934 (2020).
[4] Niemeyer, Michael, et al. "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[5] Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[6] Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[7] Sitzmann, Vincent, Michael Zollhöfer, and Gordon Wetzstein. "Scene representation networks: Continuous 3d-structure-aware neural scene representations." Advances in Neural Information Processing Systems. 2019.
[8] Liu, Lingjie, et al. "Neural Sparse Voxel Fields." arXiv preprint arXiv:2007.11571 (2020).

Calibration of deep neural networks

Calibration of deep neural networks refer to the ability of accurately representing the true data distribution. E.g. in discriminative tasks such as classification, a calibrated network predicts classes with a confidence approximately matching the original class distribution. It has been extensively shown that some successful and widely used architectures suffer from overconfident predictions [1, 2, 3] which in turn hinders uncertainty estimation and related downstream tasks, e.g. active learning.

Multiple solutions have been proposed with various degrees of implementation complexity and/or probabilistic assumptions [1, 2, 3, 4, 5, 6]. In this master thesis, we aim at studying the phenomenon of miscalibration in semantic image segmentation, adapting and evaluating the existing approaches and developing a robust solution.

To this end, one can explore multiple directions of research: a probabilistic perspective on the target and predicted data distributions similar to [4, 5], a data oriented approach as in [6] or something new and original.

The evaluation of the developed method consist in comparing to existing works on established benchmarks in image classification (e.g. CIFAR-100, SVHN, etc.) and estimating improvements in downstream tasks, sensitive to miscalibrated predictions, such as in the multimodal adversarial segmentation presented in [7].

References

[1] Guo, Chuan, et al. "On calibration of modern neural networks." arXiv preprint arXiv:1706.04599 (2017).

[2] Kull, Meelis, et al. "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration." Advances in Neural Information Processing Systems. 2019.

[3] Zhang, Jize, Bhavya Kailkhura, and T. Han. "Mix-n-Match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning." arXiv preprint arXiv:2003.07329 (2020).

[4] Kristiadi, Agustinus, Matthias Hein, and Philipp Hennig. "Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks." arXiv preprint arXiv:2002.10118 (2020).

[5] Joo, Taejong, Uijung Chung, and Min-Gwan Seo. "Being Bayesian about Categorical Probability." arXiv preprint arXiv:2002.07965 (2020).

[6] Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in Neural Information Processing Systems. 2019.

[7] Kassapis, Elias, et al. "Calibrated Adversarial Refinement for Multimodal Semantic Segmentation." arXiv preprint arXiv:2006.13144 (2020).

Task driven learned image compression

Standard image compression techniques are targeted to humans, and focus on the reconstruction error between the original image and the compressed one [1,2]. Their goal is thus to create a visually pleasing image with less artifacts as possible given a compression budget.

In computer vision, images are not given to humans, but to CNNs. We have an opportunity to target image compression to the task the CNN is solving, and thus optimize for a completely different goal: CNNs performance at inference time.

This project tries to answer the following question: Can we target image compression to a specific task? Can we exploit the peculiarities of an application in order to push the boundaries of image compression?

For example: can we compress (or even erase) the sky in object detection for autonomous driving? How can we teach a compressor do that?

The goal of the project would be to design a learned image compressor that given a compression budget can exploit the feedback from a target task to keep performance unchanged.

A task driven compression should be able to generalize to different architectures, and not be overfitted on a specific CNN architecture. [3] performs joint learning of the compressor and the CNN, coupling the compression with a specific network. [4] computes per pixel bit-rates based on a content-weighted importance map, but it still targets visual quality as its main objective. [5] performs task specific compression, but it is specifically designed for facial analysis; while the proposed approach should be generic.

The project will initially focus on simple classification datasets like CIFAR, and will gradually move to more complex tasks such as object detection or instance segmentation.

References

[1] Agustsson et al., "Generative Adversarial Networks for Extreme Learned Image Compression", CVPR, 2019

[2] Minnen et al., "Joint Autoregressive and Hierarchical Priors forLearned Image Compression", NIPS, 2018

[3] Alex Golts and Yoav Y. Schechner, "Image compression optimized for 3D reconstruction by utilizing deep neural networks", *arXiv preprint arXiv:2003.12618* (2020)

[4] Li et al., "Learning Convolutional Networks for Content-weighted Image Compression", CVPR 2018

[5] Zhibo Chen, Tianyu He, "Learning based Facial Image Compression with Semantic Fidelity Metric" *Neurocomputing*, 2019

Few shot learning for object detection by using shape priors and low-level features

Few shot learning is still an open problem in Machine Learning, especially for complex vision tasks such as Object Detection.

There's very little research in few-shot learning for Object Detection [1, 2, 3, 4]. These methods are mostly based on metric learning and optimizing the latent space structure in a way that accommodates the few samples that are available for the new classes. However, these can never represent the true distribution of the new class.

As humans we know that there are properties in objects in general that help us identify new objects easily, even with a few samples. These properties relate to things that characterize objects, such as shapes, boundaries, texture etc. We hope that these are learned implicitly by convolutional neural networks, but not often do we make these explicit.

Other tasks in computer vision such as instance segmentation have benefited from information priors such as shape [5, 6] and boundaries [7]. Another work called "Learning to Segment everything" uses weight transfer in a partially supervise setting to learn to segment a large number of new classes easily [8].

Can we then Learn to Detect Everything?

With this thesis topic, we will investigate how we can induce shape priors and low level features of unknown object categories in a few-shot learning object detection setup, in order to enable reliable detection of new classes with low sample cost. We aim to improve detection of few-shot classes by giving the models a prior of their shapes and low-level features first. We will focus on datasets such as Pascal VOC and COCO, as well as novel benchmarks that include LVIS [9] .

References

[1] Schwartz, Eli, et al. "RepMet: Representative-based metric learning for classification and one-shot object detection." arXiv preprint arXiv:1806.04728 4323, 2018
[2] Wertheimer, Davis, and Bharath Hariharan. "Few-shot learning with localization in realistic settings." CVPR, 2019
[3] Few-Shot Object Detection https://www.researchgate.net/publication/317930531_Few-shot_Object_Detection
[4] Chen, Hao, et al. "Lstd: A low-shot transfer detector for object detection." Thirty-Second AAAI Conference on Artificial Intelligence, 2018
[5] Kuo, Weicheng, et al. "Shapemask: Learning to segment novel objects by refining shape priors." ICCV, 2019
[6] Yanzhao Zhou et al., "Learning Saliency Propagation for Semi-Supervised Instance Segmentation", CVPR 2020
[7] Fan, Qi, et al. "Commonality-Parsing Network across Shape and Appearance for Partially Supervised Instance Segmentation." arXiv preprint arXiv:2007.12387, 2020
[8] Hu, Ronghang, et al. "Learning to segment every thing.", CVPR, 2018
[9] Wang, Xin, et al. "Frustratingly Simple Few-Shot Object Detection." arXiv preprint arXiv:2003.06957, 2020

Important frame selection for efficient learning in videos.

Videos contain many frames. This makes the deep networks dealing with videos less efficient and computationally intensive.
There has been work to make the training efficient [1]. The focus of these works to separate space and time domain and use a lower resolution of space in a time domain and lower resolution of time in the space domain. Since in videos there are many frames with large overlaps, the keyframe information could be sufficient to perform computer vision tasks, like action recognition. But how do you select these keyframes?

A naive way could be an offline unsupervised clustering technique, like k-means computed on image features, that treats images as an unordered set of samples, thus disregarding temporal information.
However, we hypothesize that leveraging the temporal information is key for a more informative and efficient selection process. Moreover, we think that the key-selection process should be tailored to the recognition task at hand. Therefore, in this research, we would like to study the different approaches to encode and utilize the temporal information in the key-frame selection, while optimizing for the target task in an end-to-end fashion.

The following are several relevant works in the literature: Temporal shift module [1], dataset distillation [2], unsupervised deep clustering [3], video summarization [4] and Something-something [5] is a sample relevant dataset to evaluate the proposed method on.

References:

[1] Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7083-7093).

[2] Wang, Tongzhou, et al. "Dataset distillation." arXiv preprint arXiv:1811.10959 (2018).

[2] Sudhakaran, S., Escalera, S., & Lanz, O. (2020). Gate-Shift Networks for Video Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1102-1111).

[3] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. CVPR 2018.

[4] Mahasseni, Behrooz, Michael Lam, and Sinisa Todorovic. "Unsupervised video summarization with adversarial lstm networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.

[5] Goyal, Raghav, et al. "The" Something Something" Video Database for Learning and Evaluating Visual Common Sense." ICCV 2017.

Learning from structured noisy data

With the ever-increasing popularity and success of the highly-parameterized deep neural networks, the importance of possessing gigantic datasets is increasing at the same time. However, obtaining huge datasets imposes significant costs, especially on densely labeled images, for instance in tasks such as semantic segmentation for which a precise annotation process for an image might take up to several hours [1]. One possible direction to significantly decrease the costs is to develop machine learning models that are more robust in handling noisy labels [2-4] to enable learning from coarsely annotated labels or pseudo-labels making it more feasible to benefit from much more scalable processes of creating datasets.

Despite the abundance of methods tackling the noisy labels in simpler scenarios, like image classification [2-3], not many approaches are there to take the anticipated structure of dense labels, as e.g. in semantic segmentation, into account as a source of information to tackle the problem [4]. In this project, we aim to answer the following research question: How can we effectively and efficiently make use of scene structural and shape priors to identify and mitigate the noisy samples for the semantic segmentation problem? We will evaluate the performance of the proposed model on standard semantic segmentation datasets such as CityScapes, Mapillary Vistas, etc. by systematically introducing noise in them or by using pseudo-labels to train our models.

References

[1] Cordts et al. "The cityscapes dataset for semantic urban scene understanding." CVPR 2016.
[2] Hendrycks et al. "Using trusted data to train deep networks on labels corrupted by severe noise." Neurips 2018.
[3] Han et al. "Deep self-learning from noisy labels." CVPR 2019.
[4] Abid et al. "Improving Training on Noisy Structured Labels." arXiv preprint arXiv:2003.03862 (2020).