

CVPR 2018 - Summary

תוכן עניינים

CVPR 2018 - Summary	1
כללי	7
אירוע	8
18.06.2018	12
0830-1230 Tutorial: Motion Averaging: A Framework for Efficient and Accurate Large-Scale Camera Estimation in 3D Vision (Room 151 D-F)	12
PM 1330-1700 Tutorial: Optimisation in Multiple View Geometry: The L-infinity Way (Room 151 – ABC). 12	12
Full Day: 1st International Workshop on Deep Learning for Visual SLAM (Room 255 - C)	12
1340-1740 Workshop: Large-Scale Landmark Recognition: A Challenge (Room 251 B-C)	12
Full Day: DeepGlobe: A Challenge for Parsing the Earth through Satellite Images (Room 150 – G).....	12
0845-1730 Tutorial: A Crash Course on Human Vision (Room 155 C).....	12
0850-1010 Session 1-1C: 3D Vision I (Room 255)	13
0850 Orals (O1-1C)	13
2. [C10] Hybrid Camera Pose Estimation.....	13
3. [C13] A Certifiably Globally Optimal Solution to the Non-Minimal Relative Pose Problem	13
0934 Spotlights (S1-1C)	14
1. [C16] Single View Stereo Matching.....	14
5. [D6] PPFNet: Global Context Aware Local Features for Robust 3D Point Matching	14
1010-1230 Poster Session P1-1 (Halls C-E)	14
3D Vision.....	14
10. [F1] Spline Error Weighting for Robust Visual-Inertial Fusion	14
Image Motion & Tracking	15
33. [I4] End-to-End Flow Correlation Tracking With Spatial Temporal Attention	15
Low-level & Mid-level Vision.....	15
37. [I16] The Unreasonable Effectiveness of Deep Features as a Perceptual Metric	15
38. [I19] Local Descriptors Optimized for Average Precision.....	15
47. [K2] Multi-Image Semantic Matching by Mining Consistent Features.....	16
Machine Learning for Computer Vision	17
58.[L13] Decorrelated Batch Normalization	17
65.[M12] Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++	17
67. [M18] GAGAN: Geometry-Aware Generative Adversarial Networks	17
Object Recognition & Scene Understanding	17
84.[P3] Zero-Shot Visual Recognition Using Semantics Preserving Adversarial Embedding Network .	17

1230-1450 Poster Session P1-2 (Halls C-E)	17
Object Recognition & Scene Understanding	17
5. [B3] Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks ..	17
9. [B15] Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning.....	18
Machine Learning for Computer Vision	18
41. [G1] MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks.....	18
Low-level & Mid-level Vision.....	21
58.[I8] Matching Pixels Using Co-Occurrence Statistics.....	21
69.[J19] Learning a Discriminative Feature Network for Semantic Segmentation.....	21
3D Vision.....	21
77. [K21] Matryoshka Networks: Predicting 3D Geometry via Nested Shape Layers.....	21
80.[L8] End-to-End Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching ...	21
81. [L11] ICE-BA: Incremental, Consistent and Efficient Bundle Adjustment for Visual-Inertial SLAM	21
1450-1630 Session 1-2C: Machine Learning for Computer Vision II (Room 255)	22
1450 Orals (O1-2C)	22
1. [E10] Learning to Find Good Correspondences	22
2. [E13] OATM: Occlusion Aware Template Matching by Consensus Set Maximization.....	26
1548 Spotlights (S1-2C)	29
1. [E22] Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference	29
7. [F18] Learning Deep Descriptors With Scale-Aware Triplet Networks	29
1630-1830 Poster Session P1-3 (Halls C-E)	29
3D Vision.....	29
9. [H7] Estimation of Camera Locations in Highly Corrupted Scenarios: All About That Base, No Shape Trouble.....	29
21. [I21] Camera Pose Estimation With Unknown Principal Point	29
27. [J17] Learning Patch Reconstructability for Accelerating Multi-View Stereo.....	30
Machine Learning for Computer Vision	31
54.[N10] BPGrad: Towards Global Optimality in Deep Learning via Branch and Pruning	31
58.[N22] Domain Adaptive Faster R-CNN for Object Detection in the Wild	31
61. [O9] Lightweight Probabilistic Deep Networks	32
74. [Q4] Learning Time/Memory-Efficient Deep Architectures With Budgeted Super Networks	32
20.06.2018	34
0830-1010 Session 2-1C: 3D Vision III (Room 255)	34
0830 Orals (O2-1C)	34

1. [E8] Density Adaptive Point Set Registration.....	34
3. [E14] Im2Pano3D: Extrapolating 360° Structure and Semantics Beyond the Field of View	35
0928 Spotlights (S2-1C)	35
3. [F4] Tangent Convolutions for Dense Prediction in 3D	35
4. [F7] RayNet: Learning Volumetric 3D Reconstruction With Ray Potentials.....	35
5. [F10] Neural 3D Mesh Renderer.....	35
9. [F22] Beyond Gröbner Bases: Basis Selection for Minimal Solvers	35
1010-1230 Demos (Hall C)	35
Efficient Annotation of Segmentation Datasets With Polygon-RNN++	35
1010-1230 Poster Session P2-1 (Halls C-E)	36
From Orals	36
10. [C7] DenseASPP for Semantic Segmentation in Street Scenes, Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, Kuiyuan Yang	36
Object Recognition & Scene Understanding	36
3. [G12] Finding Beans in Burgers: Deep Semantic-Visual Embedding With Localization	36
3D Vision.....	36
63. [O16] Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs.....	36
65.[O22] ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans ..	38
73. [Q2] Learning Less Is More - 6D Camera Localization via 3D Surface Regression.....	38
74. [Q5] Feature Mapping for Learning Fast and Accurate 3D Pose Inference From Synthetic Images	38
1230-1450 Poster Session P2-2 (Halls C-E)	39
Image Motion & Tracking	39
17. [C17] Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking	39
Object Recognition & Scene Understanding	40
19. [D1] Efficient Large-Scale Approximate Nearest Neighbor Search on OpenCL FPGA.....	40
27. [E3] Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation	40
Machine Learning for Computer Vision	40
73. [K9] CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise.....	40
76. [K18] Structured Uncertainty Prediction Networks	41
78. [L2] Adversarial Feature Augmentation for Unsupervised Domain Adaptation.....	41
84.[L20] Joint Optimization Framework for Learning With Noisy Labels	41
91. [M19] Analyzing Filters Toward Efficient ConvNet	42
93. [N3] In-Place Activated BatchNorm for Memory-Optimized Training of DNNs.....	43
Object Recognition & Scene Understanding	43
100.[O2] Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking	43

102.[O8] Globally Optimal Inlier Set Maximization for Atlanta Frame Estimation	44
Applications.....	45
115.[Q3] Fast Monte-Carlo Localization on Aerial Vehicles Using Approximate Continuous Belief Representations	45
116.[Q6] DeLS-3D: Deep Localization and Segmentation With a 3D Semantic Map.....	46
1450-1630 Session 2-2C: Computational Photography (Room 255).....	47
10. [G2] Learning to Detect Features in Texture Images	47
1630-1830 Demos (Hall C)	47
Efficient Annotation of Segmentation Datasets With Polygon-RNN++	47
Semi-Dense, Event-Based Visual SLAM	47
Ultimate SLAM? Combining Events, Frames and IMU for Robust Visual SLAM.....	47
Real-Time Visual SLAM Using a Jointly Optimized, Compact Dense Code	47
1630-1830 Poster Session P2-3 (Halls C-E)	47
Low-level & Mid-level Vision.....	47
35. [L6] Latent RANSAC	47
39. [L18] Graph-Cut RANSAC.....	48
Object Recognition & Scene Understanding	48
51. [N10] ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices	48
56.[O3] Semantic Visual Localization	48
21.06.2018	50
0830-1010 Session 3-1A: Object Recognition & Scene Understanding IV (Ballroom)	50
11. [B16] CVM-Net: Cross-View Matching Network for ImageBased Ground-to-Aerial Geo-Localization	50
0830-1010 Session 3-1B: Analyzing Humans	52
1. [C6] Consensus Maximization for Semantic Region Correspondences.....	52
0830-1010 Session 3-1C: Applications (Room 255).....	52
2. [D18] Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics52	
1010-1230 Poster Session P3-1 (Halls D-E).....	53
Object Recognition & Scene Understanding	53
2. [E22] Show Me a Story: Towards Coherent Neural Story Illustration.....	53
4. [F4] Fast Spectral Ranking for Similarity Search.....	53
Machine Learning for Computer Vision	54
35. [H22] Stochastic Downsampling for Cost-Adjustable Inference and Improved Regularization in Convolutional Networks	54
43. [I16] Unsupervised Domain Adaptation With Similarity Learning.....	54
51. [J10] HydraNets: Specialized Dynamic Architectures for Efficient Inference	54

54.[J16] OLÉ: Orthogonal Low-Rank Embedding - A Plug and Play Geometric Loss for Deep Learning	54
58.[K2] Fast and Robust Estimation for Unit-Norm Constrained Linear Fitting Problems	54
1250-1430 Session 3-2B: Machine Learning for Computer Vision IV (Room 155)	54
Orals.....	54
1. [C1] MapNet: An Allocentric Spatial Memory for Mapping Environments	54
Spotlights.....	54
1. [C7] Generate to Adapt: Aligning Domains Using Generative Adversarial Networks	54
3. [C11] A PID Controller Approach for Stochastic Optimization of Deep Networks.....	54
4. [C13] “Learning-Compression” Algorithms for Neural Net Pruning	54
5. [C15] Large-Scale Distance Metric Learning With Uncertainty	55
11. [D5] Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions.....	55
1250-1430 Session 3-2C: Object Recognition & Scene Understanding V (Room 255).....	55
1. [D11] Learning Descriptor Networks for 3D Shape Synthesis and Analysis.....	55
1. [D17] Learning Compositional Visual Concepts With Mutual Consistency.....	55
1450-1630 Session 3-3B: Image Motion & Tracking (Room 155)	55
6. [H1] Real-World Repetition Estimation by Div, Grad and Curl	55
1450-1630 Session 3-3C: Machine Learning for Computer Vision VI (Room 255)	55
1450 Orals (O3-3C)	55
1. [H17] Feature Space Transfer for Data Augmentation.....	55
3. [H21] Detail-Preserving Pooling in Deep Networks	56
1534 Spotlights (S3-3C)	56
2. [I3] Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions	56
9. [I17] NISP: Pruning Networks Using Neuron Importance Score Propagation	56
12. [J1] 3D Semantic Segmentation With Submanifold Sparse Convolutional Networks	56
1630-1830 Poster Session P3-2 (Halls D-E)	57
Biomedical Image.....	57
2. [J7] An Unsupervised Learning Model for Deformable Medical Image Registration	57
5. [J13] CNN Driven Sparse Multi-Level B-Spline Image Registration	57
7. [J17] 3D Registration of Curves and Surfaces Using Local Differential Information	57
Machine Learning for Computer Vision	57
16. [K13] Spatially-Adaptive Filter Units for Deep Neural Networks	57
17. [K15] SO-Net: Self-Organizing Network for Point Cloud Analysis.....	57
20.[K21] Explicit Loss-Error-Aware Quantization for Low-Bit Deep Neural Networks.....	57
Applications.....	57
23. [L5] ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing.....	57
22.06.2018	58

0830-1230 Tutorial: Differential Geometry for Engineers (Ballroom H)	58
0840-1710 Workshop: Visual Odometry and Computer Vision Applications Based on Location Clues	59
1:30 pm - 2:10 pm, Keynote Talk: Ruigang Yang (Baidu Research & University of Kentucky)	59
2:10 pm - 2:50 pm, Keynote Talk: Anelia Angelova (Google Brain), Talk topic: Unsupervised Learning of Depth and Ego-motion using 3D Geometric Constraints	60
2:50 pm - 3:10 pm, Oral: A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration With Applications to UAV Geolocalization, Ahmed Nassar (IRISA Institute), Mohamed ElHelw (Nile University), Karim Amer (Nile University), Reda ElHakim (Nile University)	61
3:50 pm - 4:30 pm, Keynote Talk: Andreas Geiger (MPI & University of Tübingen), Talk topic: Semantic Visual Localization	62

כללי

מסמך זה מסכם את הרשמי של מכנס CVPR 2018 בSalt Lake City, Utah, USA בחודש יוני 2018.

קצת מספרים:

- הכנס ארך 5 ימים, מתוכם 3 הוקדשו לכנס המרכז ו 2 לסדנאות (workshops) ומדריכים (tutorials), והכיל כ-6500 משתתפים.
- הכנס המרכזי:
 - התקבלו 979 מאמרים (מתוך 3552 אשר הוגש – סיכוי של 1/3), מתוכם כ-300 הוצגו באופן פרונטלי: 70 orals אשר ארכו 12 דקות ו 224 spotlights אשר ארכו 4 דקות.
 - כל המאמרים הוצגו גם על ידי פוסטרים לאורך 8 מושבים של שעתיים כל אחד. אחד המחברים היה זמין ליד הפוסטר על מנת להסביר ולענות לשאלות במשך אחד מהמושבים.
 - בחלוקת גסה, כחצי מהיום הוקדש להצגות פרונטליות וחצי השני לפוסטרים.
 - באופן די מפתיע, מצאת את מושבי הפוסטרים מועלם ומלבדם שמעניינים אותו ולדון עם המחבר באופן חופשי יותר. עם עיקר בಗל שאפשר לבחור את המאמרים שמעניינים אותך ולדון עם המחבר בדרך כלל היה צריך לעמוד בתור.
 - בנוסף לתערוכת הפוסטרים, התקיימה תערוכה של כ-140 חברות מה תעשייה אשר הציגו חידושים, גיבועים ועבודים וגם חילקו מזכרות ☺
- סדנאות ומדריכים:
 - 21 מדריכים, 48 סדנאות, כל אחד באורך חצי יום או يوم שלם.
 - אורך הרצאה אופיינית: 40 דקות.

המטרה של המסמך זהה היא לספר סקירה קצרה של העבודות שענינו אותנו בכנס. בשום אופן לא מדובר בסקירה מקיפה או מעמיקה – זאת ניתן להציג על ידי התעמקות במאמרים עצמם.

הסיכון מכיל מספר שורות על כל עבודה אשר נכתבו באופן חופשי ולא פורמלי, יחד עם תוכנות והערות של. בחלק מהעבודות צימתי את הפוסטר (לא תמיד זכרתי לעשות זאת ☺). בחלק מהמקרים לא הבנתי את כל הפרטים, והשתדלתי לציין זאת בסיכום. מפאת קוצר הזמן ועומס המאמרים לא הספקתי להגיע לכל העבודות שענינו אותנו – במרבית המקרים עבדות אלו מסווגות על ידי שורה או שתיים אשר זיקתי מהתקציר (abstract).

לא צירפתי קישורים למאמרים המקוריים (יש הרבה מאמרים ומדובר בהרבה עבודה ☺ וחוץ מזה קל למצוא אותם באינטרנט). את כל המאמרים (וגם חלק מהסדנאות) ניתן למצוא באתר של CVF:

openaccess.thecvf.com

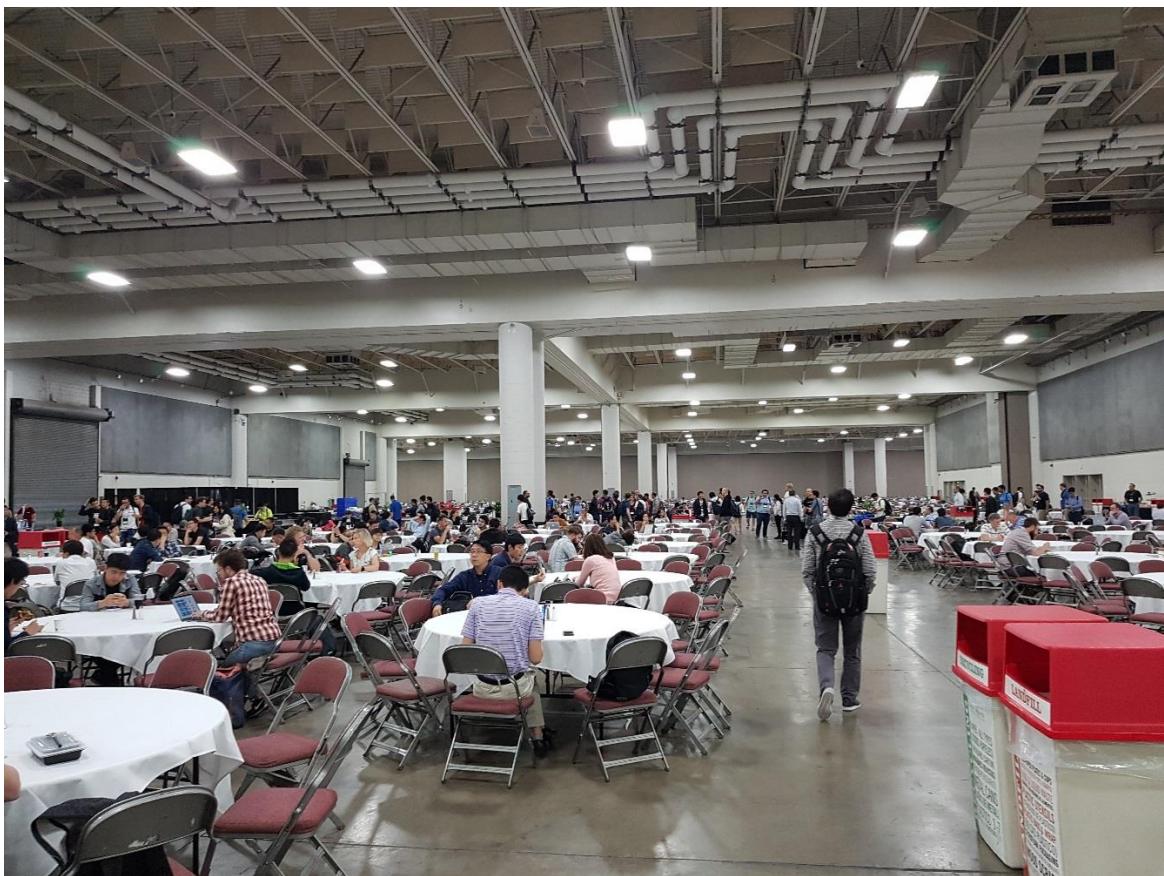
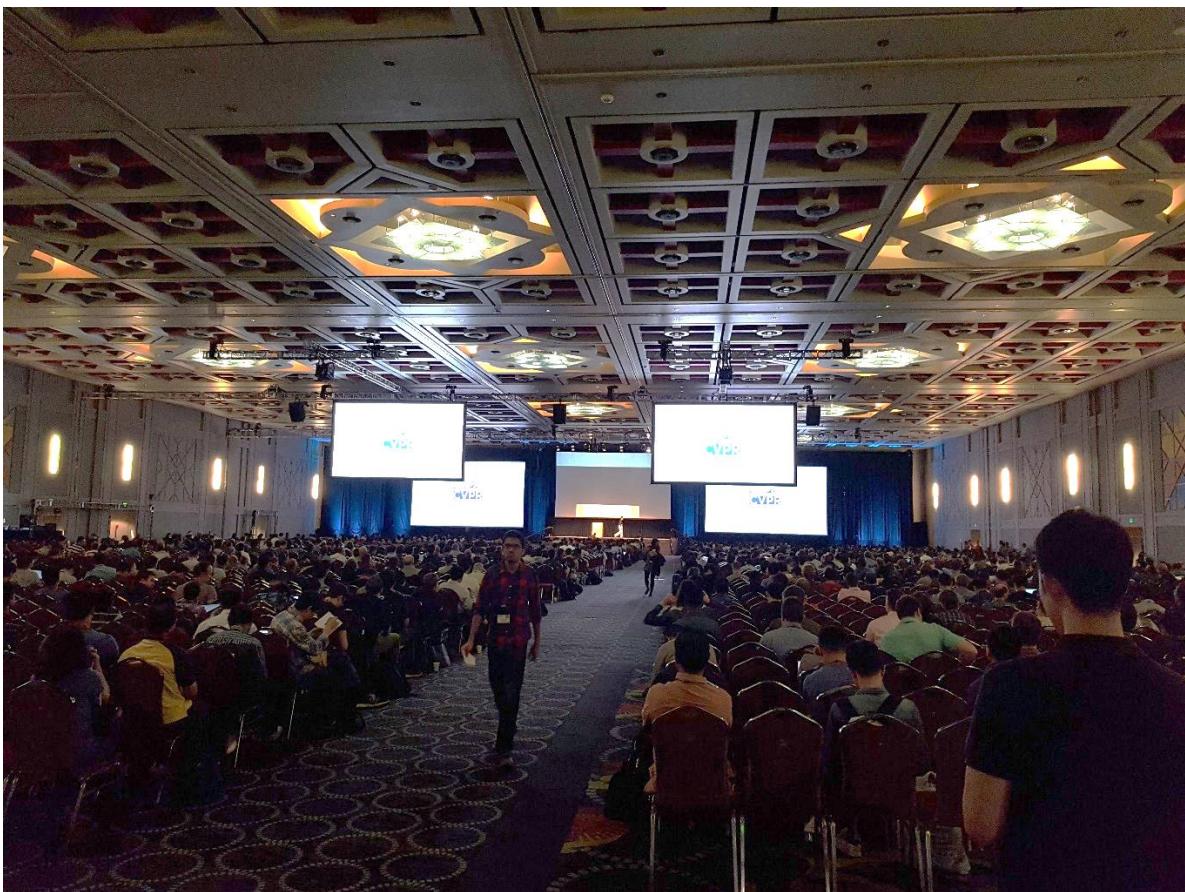
דבר אחר שאני כן מצורף להן 2 חוברות (pocket guides) אשר מסכימות את כל העבודות אשר הוצגו בכנס, אחת עבור הכנס המרכזי והשנייה עבור הסדנאות והמדריכים. באמצעות חוברות אלו תוכל לחפש כתורות מעניינות באופן מהיר, ולאחר מכן לחפש את המאמרים עצם באינטרנט:

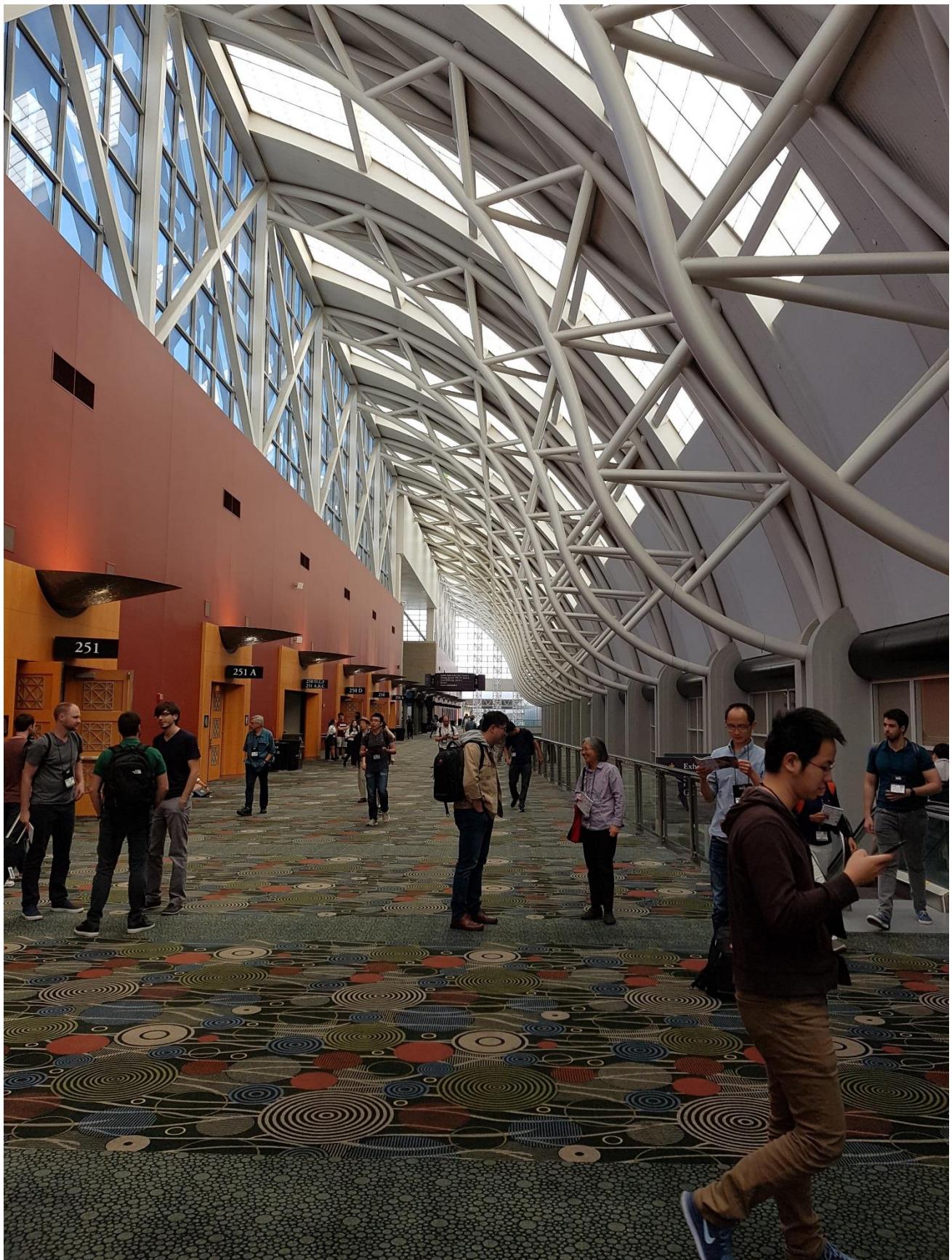


אני מקווה שתמצאו את המסמך הזה מעיל. אם אתם מעוניינים לדבר על אחד הנושאים, אם גיליתם טעות, יש לכם הערה או הארה – אתם יותר ממוזמנים ליצור איתי קשר במיל: moshes777@gmail.com או בנוייד: 052-6545920.

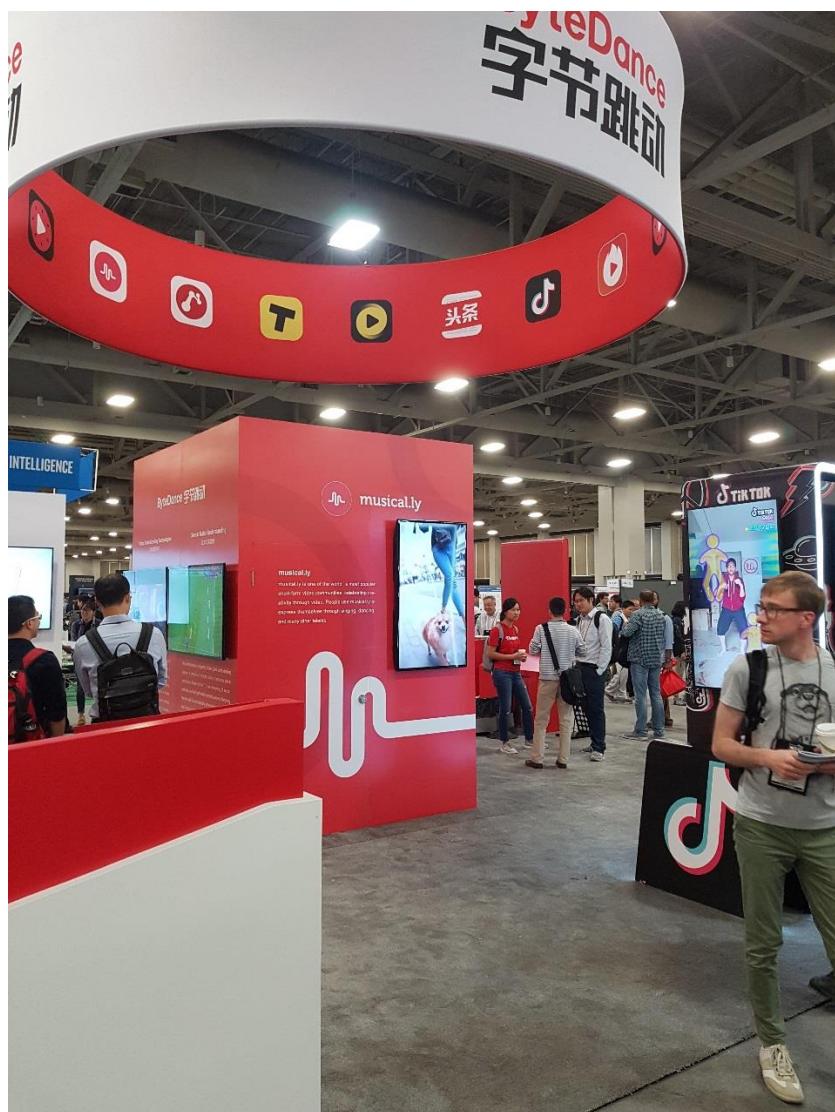
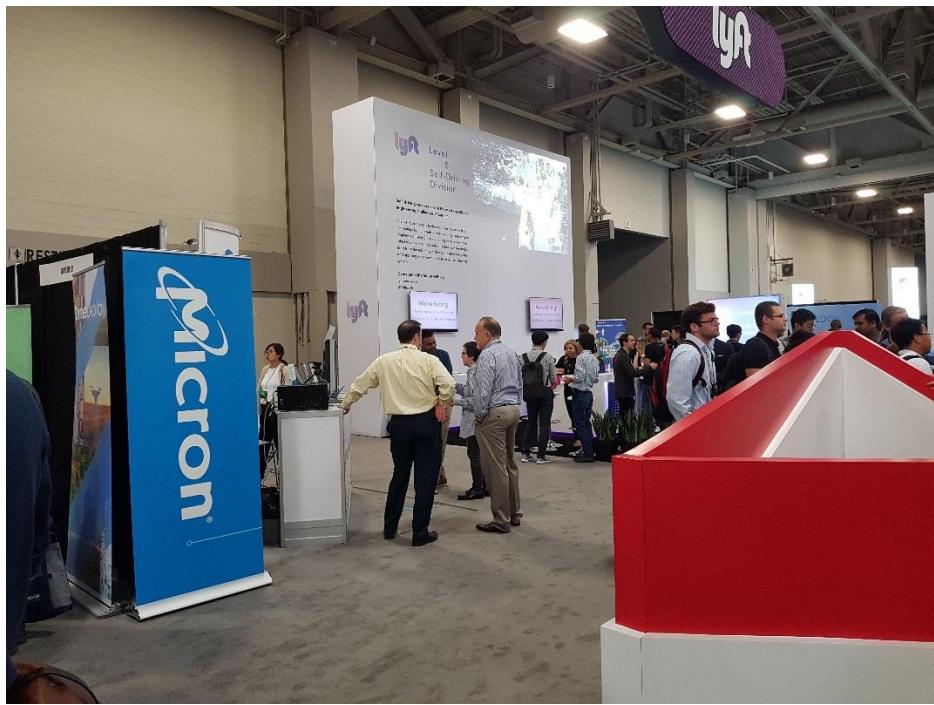
משה שלמאי

אוירה









18.06.2018

דណאות בהן השתתפתו:

0830-1230 Tutorial: Motion Averaging: A Framework for Efficient and Accurate Large-Scale Camera Estimation in 3D Vision (Room 151 D-F)

שיעור מודל מצלמה בגישה של Motion Averaging: בהינתן שערוך של תנועה יחסית של מצלמות (לאורך זמן), יש לשערך את הפרמטרים שלהן בכל נקודה בזמן.



Motion Averaging
A Framework for Eff

PM 1330-1700 Tutorial: Optimisation in Multiple View Geometry: The L-infinity Way (Room 151 – ABC)

הנושא המרכזי הוא שימוש בনורמת L_{infinity}, במקום בnormet L₂, בעיות אופטימיזציה של multiple view geometry. גישה זו אמורה לעזור בעיות לא-קמורות (non-convex), דוגמת bundle adjustment, בהם יש מספר מינימיות מקומיות. הTutorial כולל סקירה של החומר התאורטי הבסיסי והציגו של נושאים מתקדמים במחקר.



Optimization in
Multiple View Geom

דណאות מעניינות נוספת:

Full Day: 1st International Workshop on Deep Learning for Visual SLAM (Room 255 - C)

1340-1740 Workshop: Large-Scale Landmark Recognition: A Challenge (Room 251 B-C)

אתגר של גугл העוסק בזיהוי מבנים (landmark) במקומות שונים בעולם. נשמע מעניין ורלוונטי.

Full Day: DeepGlobe: A Challenge for Parsing the Earth through Satellite Images (Room 150 – G)

אתגר העוסק בניתוח תמונות לוין עבור שלוש משימות: זיהוי כבישים, זיהוי בניינים, וסגמנטציה/סיווג תכסית.

0845-1730 Tutorial: A Crash Course on Human Vision (Room 155 C)

מדריך בנושא ראייה אנושית: היבטים פיזיולוגיים ופסיכולוגיים, עם אוריינטציה ליישומי ראייה ממוחשבת.

19.06.2018

0850-1010 Session 1-1C: 3D Vision I (Room 255)

0850 Orals (O1-1C)

2. [C10] Hybrid Camera Pose Estimation

שערור מצלמה מתוך מודל 3D תוך שימוש בשיטות 3D-2D ו 2D-2D.

בהינתן מודל SfM Structure from Motion המטרה היא למצוא הזזה וסיבוב של מצלמות. דרך מקובלת לפתור את הבעיה היא באמצעות שיטות 3D-2D. יש מספר חסרונות (צריך לבצע טריינגולציה ועוד) ויתרונות – לא הספקתי לרשום. דרך נוספת היא לפתור באמצעות שיטות 2D-2D בלבד. יתרון: לא צריך לבצע טריינגולציה. חיסרון משמעותי – איטי מאד, בערך פי 1000 מאשר השיטה הקודמת, כיוון שצריך 6 נקודות לפחות לפתרון מינימלי ויחד עם לולאט RANSAC זה לוקח זמן רב.

השאלות העיקריות בעבודה זו: האם ניתן לשלב שיטות משתי השיטות? כיצד ניתן לשלב בתהליך מסווג RANSAC?

התשובה לשאלת הראושונה היא כן: המחברים הציעו 9 פתרונות מינימליים אפשריים אשר משלבים שיטות משני הסוגים: חלק 2D-2D וחילק 3D-2D.

זמן הריצה הוא בערך באותה ביצוע בין שתי השיטות המקוריות.

שילוב בלולאט RANSAC: H-RANSAC

1. בכל איטרציה דוגמים פותרן solver מתוך 2 השיטות המוצעות – הוסבר קצר באריכות, ולאחר מכן דוגמים שיטות הדורשים לפותרן.

2. תנאי עצירה

תוצאות על מידע אמיתי מראות שמקבלים יותרliers יתרכז בשיטה המוצעת. בנוסף מקבלים גםliers בפרק – מה שבדרך כלל לא קורה באחת מהשיטות המוצעות (לא הבנתי איזה מהם). מבחינה זמן ריצה: גם כאן הזמן הוא בערך באותה ביצוע בין השיטות הקודמות יחד עם RANSAC.

במפגש המסקנות נאמר כי השיטה המוצעת מכילה פתרונות ל4 בעיות של שערור מצלמה.

3. [C13] A Certifiably Globally Optimal Solution to the Non-Minimal Relative Pose Problem

העבודה עוסקת במציאת מיקום מצלמה יחסית בין 2 מצלמות מתוך סט נקודות 3D בעולם. יש מספר פתרונות מקובלים, אך הם לא מספקים הבטחה לפתור את הבעיה באופן אופטימלי ובזמן פולינומי.

קשה ידוע באופטימיזציה בתחום זה הוא שמדובר בבעיה לא לינארית. בעבודה זו פותחה רלקסצייה קמורה convex Quadratically Constrained Quadratic Program relaxation והזוכה לבעיה. הרעיון הוא לנסח את הבעיה בצורה של Semidefinite Program (SDP) והזוכה אינהQCQP) ואחר כך לבצע רלקסצייה לצורה של (SDP)冗余 redundant terms על מנת לשפר את הבדיקות של הרלקסצייה. דרך ההתחממות היא להשתמש באילוצים יתרים על מנת לשפר את הבדיקות של הרלקסצייה. בצורה פשוטה אפשר לקבל 21 אילוצים נוספים, ולאחר שימוש בטרייקים – מקבלים $255+270=525$ אילוצים נוספים (לא הבנתי אם מדובר בתוספת או בהכפלה של המספרים הללו). לאחר הוספה האילוצים הללו הרלקסצייה SDP היא הדוקה (באופן אמפירי). שלבי האלגוריתם (עם פתרון אופטימלי מובטח):

1. בהינתן הנתונים, לחשב את כל האילוצים

2. לפתור על ידי SDP

3. לפרק לוקטורים וערכים עצמיים eigen decomposition

4. לחשב את הפתרון לבעיה המקורית

5. לוודא אופטימליות עם dual certificate

בשוואה לשיטות הקיימות – רואים שהפתרונות שיטות קיימות מציאות לא תמיד אופטימליים.

מבחרת זמינים – בגלל שמדובר בSDP אז הזמן לא כל כך קצרים (המרצה הציג בעבר 1 שניה).

0934 Spotlights (S1-1C)

1. [C16] Single View Stereo Matching

Spatial depth \ עומק מתוך תמונה בודדת תוך שימוש באילוצים גיאומטריים. מוטיבציה לעובדה: 1. שימוש ב-algebraic methods. 2. stereo matching, transformers :matchings. השלבים:

1. שערור מפת disparity על ידי CNN.
 2. חישוב מפת עומק – לא הבנתי בדיק כיצד.
- תוצאות טובות יותר ביחס לשיטות קיימות.

5. [D6] PPFNet: Global Context Aware Local Features for Robust 3D Point Matching

למידת דסקריפטורים עבור השוואת 2 עניינית נקודות 3D. (מודול מבנים ?)

השאלה היא כיצד למצוא שיוכים בין 2 ענייניות נקודות 3D. גישות קיימות עושות שימוש בדסקריפטורים מקומיים 3D. הגישה המוצעת עובדת ישירות על ענייני הנקודות, ללא שימוש בדסקריפטורים מקומיים. לא הבנתי הכל.

1010-1230 Poster Session P1-1 (Halls C-E)

3D Vision

10. [F1] Spline Error Weighting for Robust Visual-Inertial Fusion

מערכת לשערור מסלול של מיקום מצלמה תוך שילוב מידע מתמונות ומדידים אינרציאליים. פתרון עובד על רצף (במובן של אינטגרל לעומת סכום) ולאו דווקא על מיקומים בודדים, ולכן מתחאים במיוחד לצלמו מוסוג rolling shutter shutter video, וכאן מתחאים במיוחד לצלמו מוסוג consumer consumer cameras (בדברי המציג הרבה המצלמות הרכביות הנקראות consumer cameras). בהן התמונה מתקנית על ידי סוג של אינטגרציה רציפה בזמן).

מסוג זה.

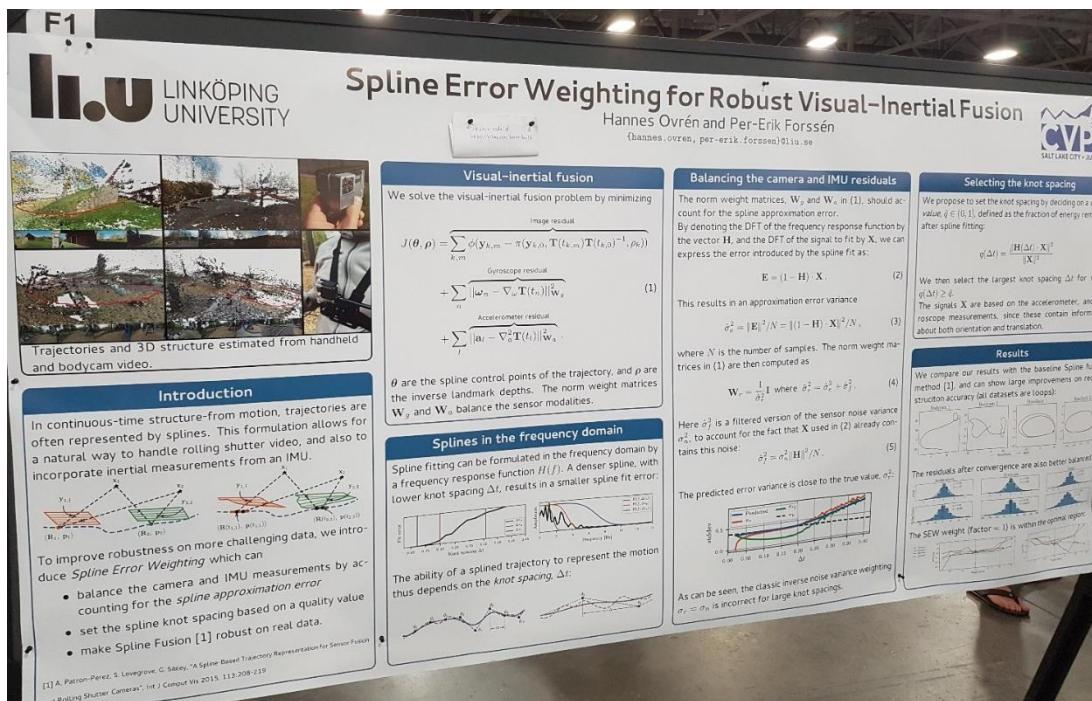


Image Motion & Tracking

33. [I4] End-to-End Flow Correlation Tracking With Spatial Temporal Attention

Low-level & Mid-level Vision

37. [I16] The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

שימוש Deep Features לשם הבנה חזותית \ תפיסתית מוביל לתוצאות טובות יותר מאשר שיטות תפיסה חזותית קיימות.

38. [I19] Local Descriptors Optimized for Average Precision

למיידת descriptors לשם מקסום AP. הרעיון המרכזי הוא לנפח את בעית דוגר, כך שיש חلون שאלתה query ויש אוסף של חלונות אחר אשר חלק (קטן) מהן דומה לchlון השאלתה (מסתכל על אותו איזור בעולם, אך מנקודות מבט \ תנאי צילום שונים) וחלק גדול לא דומה distractors. הרעיון המוצע הוא ללמידה דסקריפטור כזה אשר בהינתן ממד דמיון \ מרחק (לדוגמא מרחק אוקלידי) ממין את הדסקריפטורים של החלונות הנ"ל כך שהדומים מופיעים בראש הרשימה והשונים לאחריהם. ממד מקובל להערכת טיב דירוג הוא AP. בעבודה זו מציגים פונקציית מחיר המנסה ללמידה דסקריפטורים אשר ימקסמו את AP ישירות. לדברי המחברים גישה זו שונה מגישות קודמות כיון שאינה מכילה חלק מהונדי ידנית (manually). למשל HardNet מתמקד בדוגמה השלילית הקשה ביותר בלבד, מtorrent מחשבה שקריטריון זה יוביל למיידת דסקריפטור טוב.

רעיון נוסף המופיע בעבודה הוא שיטה לבחירת דוגמאות שליליות: המחברים חוששים ממקרים בהם יש דפוסים חוזרים בתמונה כך שיש חלונות שליליים שבאמת דומים לחلون השאלתה. על מנת להימנע מחלונות כאלה, הם מציעים להשתמש במדד דמיון קלאסי HOG על מנת לייצג כל החלונות בסיס הנטונום. לאחר מכן מבצעים ניתוח אשכולות (clustering), ודוגמים לבחור דוגמאות שליליות מסוימות שונות (אם הבנתי נכון).

שאלתי את המחבר לגבי התוצאות של רשותות HardNet ו-L2Net אשר הוצגו במאמר: במאמר שלו ברוב המקרים L2Net הוביל לתוצאות (מעט) טובות יותר מאשר HardNet – הfork מהה שנכתב בHardNet. המחבר אמר שהוא לא עשה שימוש ברשותות שפורסמו אל אימן את כל הרשותות בלבד על A split של HPatches, ולכן יכולם להיות הבדלים קטנים ביחס למופיע במאמרים המקוריים.

Local Descriptors Optimized for Average Precision

Kun He¹ Yan Lu² Stan Sclaroff¹
¹Boston University ²Honda Research Institute USA

Learning Local Features

(feature based) computer vision pipelines:

Feature Descriptors for vision pipelines
or the Feature Matching stage

Learning Feature Matching Performance

This nearest neighbor retrieval w/ binary relevance metric: Average Precision (AP)

Hashing as Tie-Aware Learning to Rank

AP

• Directly reuse TALR descriptors: reduce to TALR by distance quantization

"Learning to Rank" View

- Most existing methods: local ranking with triplets
- Optimization issues (hard negative mining, sampling)
- Ours: listwise ranking
- Direct optimization, no complex heuristics

True match ✓ False match ✗

Increasing descriptor distance

Triplet Listwise

Task-Specific Improvements

- Geometric Alignment:** Spatial Transformer module [2]

42x42 → Conv → Conv → Conv → 6-DOF Affine → Bilinear Sampling → 32x32

- Label Mining** on HPatches dataset [3]
- Cluster patches to mine in-sequence hard negatives

Experiment

- UBC Phototour / Brown dataset: patch verification

Method	Train	Notredame	Liberty	Yosemite	Notre Dame
SIFT	128	29.84	3.82	5.65	1
MatchNet (CVPR'15)	128	7.04	11.47	3.06	
TFeat-M* (BMVC'16)	128	7.39	10.31	3.06	
TL-AS-GOR (ICCV'17)	128	4.80	6.45	1.95	5
DC-2ch2at* (CVPR'15)	512	4.85	7.20	1.90	
L2Net+ (CVPR'17)	128	2.46	4.7	0.72	2
HardNet+ (NIPS'17)	128	2.28	3.25	0.57	
DOAP+	128	1.54	2.62	0.43	
DOAP-ST+	128	1.47	2.29	0.39	1

- RomePatches [3]: patch retrieval

Method	Coverage	Dim.	Train	Test
SIFT	51x51	128	91.6	87.9
AlexNet-conv3	99x99	384	81.6	79.2
PhilipNet (arXiv'14)	64x64	512	86.1	81.4
CKN-grad (ICCV'15)	51x51	1024	92.5	88.1
DOAP	51x51	128	95.8	88.4
Binary DOAP	51x51	256	95.2	86.8

- HPatches [3]: patch verification/retrieval, image matching 116 image sequences (76 train, 40 test), 2.5M patches

Method	DiffSeq	SameSeq	Viewpt	Illum	Easy
RootSIFT	56.02%	63.35%	SIFT	53.48%	53.48%
SIFT	78.48%	78.48%	RootSIFT	53.39%	53.39%
DOAP-Lb	78.48%	78.48%	DOAP-Lb	52.95%	52.95%
TFeat-M* Lb	83.56%	83.56%	TFeat-M* Lb	52.79%	52.79%
HardNet-Lb	84.21%	84.21%	HardNet-Lb	52.70%	52.70%
L2Net-Lb	87.69%	87.69%	L2Net-Lb	52.67%	52.67%
DOAP-Lb	88.05%	88.05%	DOAP-Lb	52.64%	52.64%
DOAP-ST-Lb	88.05%	88.05%	DOAP-ST-Lb	52.64%	52.64%
DOAP	93.93%	93.93%	DOAP	52.64%	52.64%
DOAP-ST	94.45%	95.70%	DOAP-ST	52.64%	52.64%
DOAP-ST-LM	95.70%	95.70%	DOAP-ST-LM	52.64%	52.64%

DiffSeq SameSeq Viewpt Illum Easy

Patch Verification mAP [%] Patch Retrieval mAP [%]

References

[1] Y. Tian, B. Fan, F. Wu, L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space, CVPR 2017
[2] M. Jaderberg et al. Spatial Transformer Networks, NIPS 2015
[3] M. Paulin et al. Local Convolutional Features with Unsupervised Training for Image Retrieval, ICCV 2015
[4] V. Balntas*, K. Lenc*, A. Vedaldi, K. Mikolajczyk, HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR 2017

Local Descriptors Optimized for Average Precision

Kun He¹ Yan Lu² Stan Sclaroff¹
¹Boston University ²Honda Research Institute USA

Learning Local Features

(based) computer vision pipelines:

Feature Descriptors for vision pipelines
or the Feature Matching stage

Learning Feature Matching Performance

This nearest neighbor retrieval w/ binary relevance metric: Average Precision (AP)

Hashing as Tie-Aware Learning to Rank

AP

• Directly reuse TALR descriptors: reduce to TALR by distance quantization

"Learning to Rank" View

- Most existing methods: local ranking with triplets
- Optimization issues (hard negative mining, sampling)
- Ours: listwise ranking
- Direct optimization, no complex heuristics

True match ✓ False match ✗

Increasing descriptor distance

Triplet Listwise

Task-Specific Improvements

- Geometric Alignment:** Spatial Transformer module [2]

42x42 → Conv → Conv → Conv → 6-DOF Affine → Bilinear Sampling → 32x32

- Label Mining** on HPatches dataset [3]
- Cluster patches to mine in-sequence hard negatives

Experiments

- UBC Phototour / Brown dataset: patch verification

Method	Train	Notredame	Liberty	Yosemite	Notre Dame
SIFT	128	29.84	3.82	5.65	1
MatchNet (CVPR'15)	128	7.04	11.47	3.06	
TFeat-M* (BMVC'16)	128	7.39	10.31	3.06	
TL-AS-GOR (ICCV'17)	128	4.80	6.45	1.95	5
DC-2ch2at* (CVPR'15)	512	4.85	7.20	1.90	
L2Net+ (CVPR'17)	128	2.46	4.7	0.72	2
HardNet+ (NIPS'17)	128	2.28	3.25	0.57	
DOAP+	128	1.54	2.62	0.43	
DOAP-ST+	128	1.47	2.29	0.39	1

- RomePatches [3]: patch retrieval

Method	Coverage	Dim.	Train	Test
SIFT	51x51	128	91.6	87.9
AlexNet-conv3	99x99	384	81.6	79.2
PhilipNet (arXiv'14)	64x64	512	86.1	81.4
CKN-grad (ICCV'15)	51x51	1024	92.5	88.1
DOAP	51x51	128	95.8	88.4
Binary DOAP	51x51	256	95.2	86.8

- HPatches [3]: patch verification/retrieval, image matching 116 image sequences (76 train, 40 test), 2.5M patches

Method	DiffSeq	SameSeq	Viewpt	Illum	Easy
RootSIFT	56.02%	63.35%	SIFT	53.48%	53.48%
SIFT	78.48%	78.48%	RootSIFT	53.39%	53.39%
DOAP-Lb	78.48%	78.48%	DOAP-Lb	52.95%	52.95%
TFeat-M* Lb	83.56%	83.56%	TFeat-M* Lb	52.79%	52.79%
HardNet-Lb	84.21%	84.21%	HardNet-Lb	52.70%	52.70%
L2Net-Lb	87.69%	87.69%	L2Net-Lb	52.67%	52.67%
DOAP-Lb	88.05%	88.05%	DOAP-Lb	52.64%	52.64%
DOAP-ST-Lb	88.05%	88.05%	DOAP-ST-Lb	52.64%	52.64%
DOAP	93.93%	93.93%	DOAP	52.64%	52.64%
DOAP-ST	94.45%	95.70%	DOAP-ST	52.64%	52.64%
DOAP-ST-LM	95.70%	95.70%	DOAP-ST-LM	52.64%	52.64%

DiffSeq SameSeq Viewpt Illum Easy

Patch Verification mAP [%] Patch Retrieval mAP [%]

References

[1] Y. Tian, B. Fan, F. Wu, L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space, CVPR 2017
[2] M. Jaderberg et al. Spatial Transformer Networks, NIPS 2015
[3] M. Paulin et al. Local Convolutional Features with Unsupervised Training for Image Retrieval, ICCV 2015
[4] V. Balntas*, K. Lenc*, A. Vedaldi, K. Mikolajczyk, HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR 2017

47. [K2] Multi-Image Semantic Matching by Mining Consistent Features

omidat maayanim dillim robuskim leshem bicutu shiokim sumantim bin tamonot.

Machine Learning for Computer Vision

58.[L13] Decorrelated Batch Normalization

שיפור של BN, אשר מלבד הפחתת ממוצע וחולקה בסטיית תקן מוסיף גם הלבנה (whitening) תוך שימוש בZCA. אם אני מבין נכון, BN עובד על כל ערוץ בנפרד, בעוד השיטה המוצעת עובדת על כל הערוצים יחד ודואגת לכך שייהי חסר קורלציה.

65.[M12] Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++

סגןנטציה סמי-אוטומטית: מפעיל מסמן מלבן חום ורשות מחשבת פוליגון חום מדויק יותר. יכול לעזור בתהילcis פענוח.

67. [M18] GAGAN: Geometry-Aware Generative Adversarial Networks

GAN המתמחב בגיאומטריה על מנת ליצור תמונות ריאליסטיות.

Object Recognition & Scene Understanding

84.[P3] Zero-Shot Visual Recognition Using Semantics Preserving Adversarial Embedding Network

למידת ייצוג (embedding) סמנטי באמצעות רשת מתחילה.

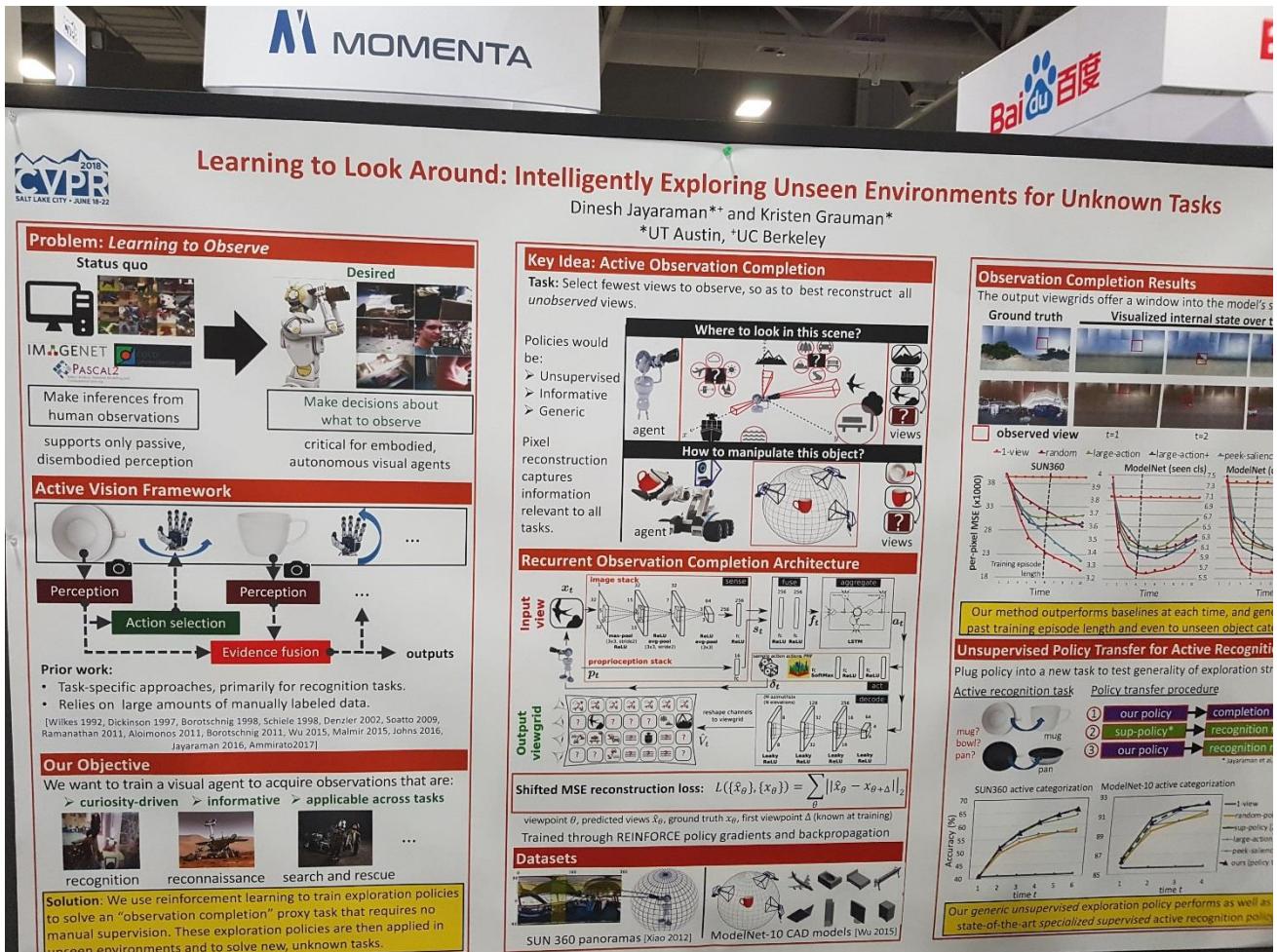
1230-1450 Poster Session P1-2 (Halls C-E)

Object Recognition & Scene Understanding

5. [B3] Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks

חקירת סביבה חדשה באופן עצמאי. לדעתי השם קצת מטעה ובפועל בוצע משהו מעט שונה. המטרה בעבודה זו היא למידת מדיניות policy learning כללי עבור עניה גנריית, ולאחר מכן ביצע כוונון עדין fine tuning / transfer learning (באופן דி פשוט): יש מצלמה שאפשר עבור למידת מדיניות ממשימה ספציפית. הדוגמא שהוצגה היא חקירת סצנה (באופן דி פשוט): יש מצלמה שאפשר להציג קצת לצדים, והמטרה היא ללמידה מדיניות ההזזה של המצלמה כך שתכסה את כל נקודות המבט על הסצנה עם מינימום ההזזה. הדוגמא הזה אמורה להיות ממשימה גנריית. לאחר מכן בוצע transfer learning למשימה אחרת – זיהוי אובייקט: מקבלים תמונה ראשונית של אובייקט (לדוגמא ספל קפה המצלום מלמעלה ונראת קצת כמו צלחת או קערה), ואפשר להציג את המצלמה על מנת לzechות את האובייקט (באמצעות מבט מהצד ניתן להזות בקלות שמדובר בספל ולא בקערה). גם כאן, המטרה היא ללמידה של הציגת מצלמה כך שהאובייקט יזוהה עם מינימום ההזזה.

העבודה עצמה נראה לי די פשוטה ולא מרשימה. רעיון מעניין יכול להיות בהקשר של ניוטו: לדוגמא רובוט מתעורר במקומות מסוימים, וצריך לחקור את הסביבה על מנת להגיע אליו (שנתון על ידי תמונה, למשל) באופן יעיל.



9. [B15] Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning

למידת רשת למשימות ذיהוי, סיווג וסמן-מציה סמנטית באופן weakly supervised.

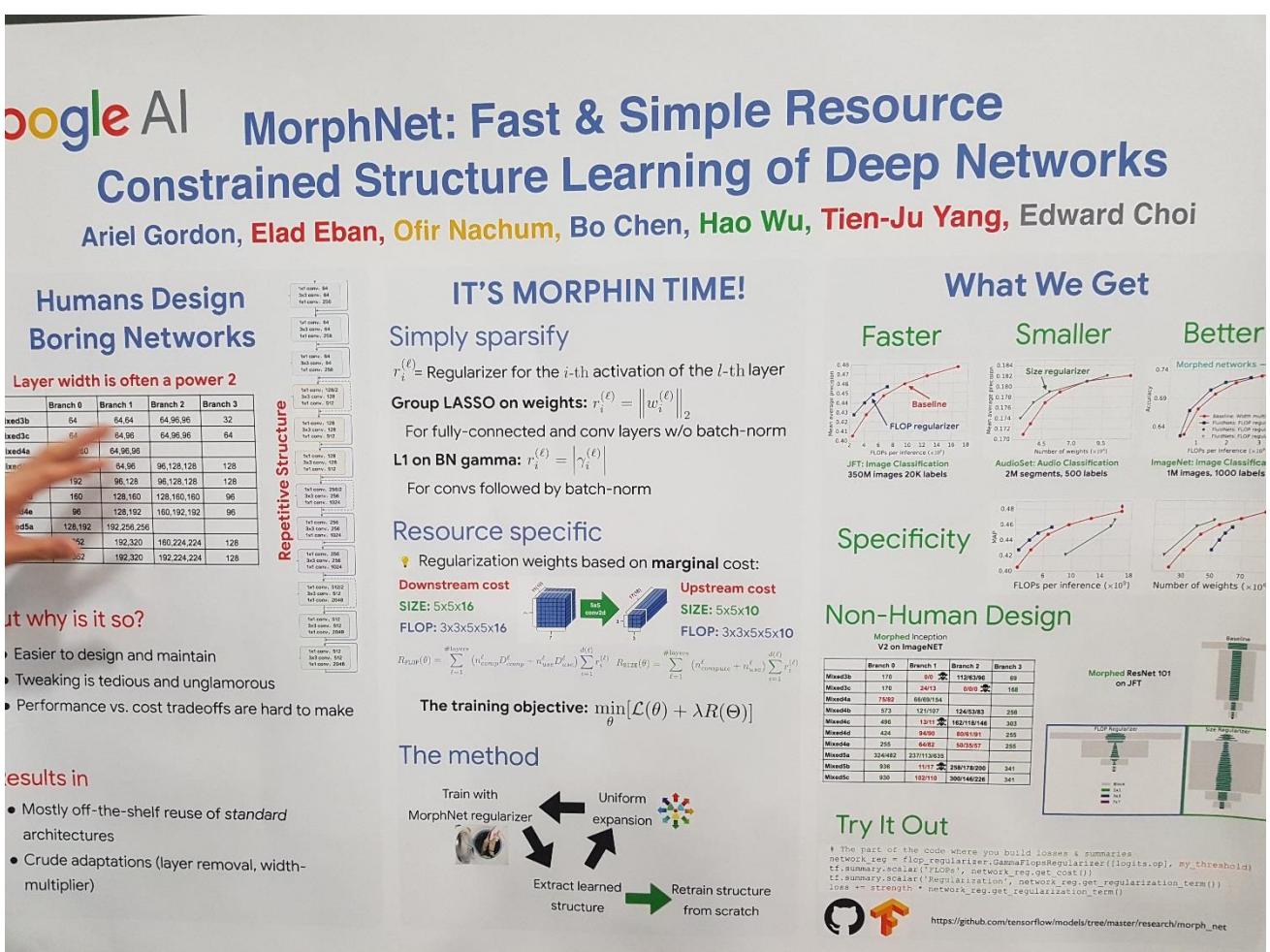
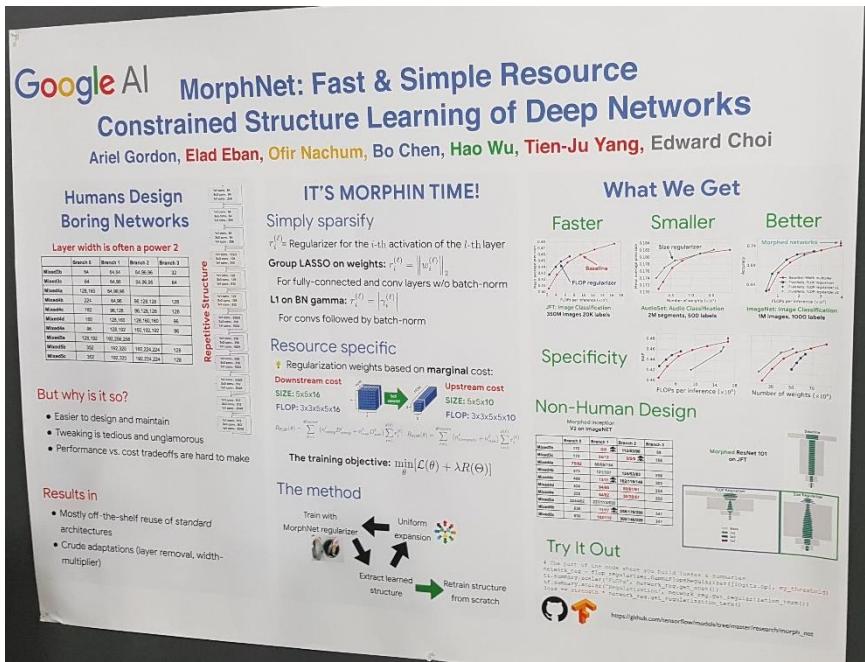
Machine Learning for Computer Vision

41. [G1] MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks

עבודה של גול – שיפור ארכיטקטורת רשת באופן אוטומטי. הרעיון הוא להוסיף איברי רגולריזציה על האקטיבציות של הרשת (בשונה מהגישה הסטנדרטית בה הרגולריזציה היא על המקדים), כך שאפשר להציג אקטיבציות מסוימות, כאשר האקטיבציה 'מתה' – לא ציריך יותר את המסתן שיצר אותה ואפשר להעלים את המקדים שלו (בשבבota konvolוציה כל אקטיבציה מקושרת לרמת עומק בודדת של השכבה).

בעבודה ניתוחו את ההשפעה של הריגת אקטיבציות על שכבות אחרות במעלה upstream ובמורד downstream. הרשת, והגדרו 2 ייעול אפשריים: 1. חיסכון במספר המקדים (זיכרון), או 2. חיסכון במספר הפעולות (FLOPS). העבודה מאפשרת לשנות על tradeoff בין זיכרון לFsops.

תהליך האימון הוא איטרטיבי: לומדים רשת בסיס, מאמנים עם פונקציית מחיר מקורית + איברי רגולריזציה מוצעים ובסיום האימון מסירים מושנים המקיים לאקטיבציות עם ערך נמוך מס' מסוים. לאחר מכן ניתן לעצור או לבצע איטרציה נוספת. במידה ומבצעים איטרציה נוספת, המחברים מצאו שכך להגדיל את מספר המקדים ברשת בפקטור של כ-1.25 על מנת שלא ייווצרו צווארי בקבוק בראשת (לא ברור לי מדוע לא לעשות זאת על ידי הקטנת ערך הסף כך שפחות שכבות יعلמו).



Simple Resource Learning of Deep Networks

Chen, Hao Wu, Tien-Ju Yang, Edward Choi

IN TIME!

What We Get

ation of the l -th layer

$$= \|w_i^{(l)}\|_2$$

layers w/o batch-norm

orm

on marginal cost:



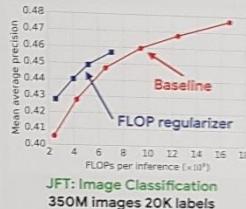
$$R_{\text{SIZE}}(\theta) = \sum_{\ell=1}^{\# \text{layers}} (n_{\text{compute}}^{\ell} + n_{\text{use}}^{\ell}) \sum_{i=1}^{d(\ell)} r_i^{(\ell)}$$

$$\min[\mathcal{L}(\theta) + \lambda R(\Theta)]$$

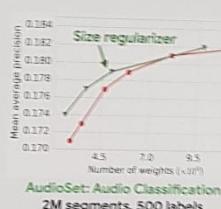
Uniform
expansion

Retrain structure
from scratch

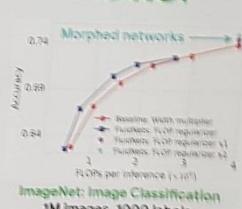
Faster



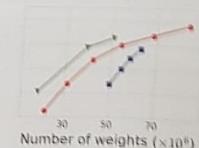
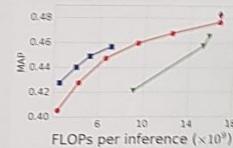
Smaller



Better



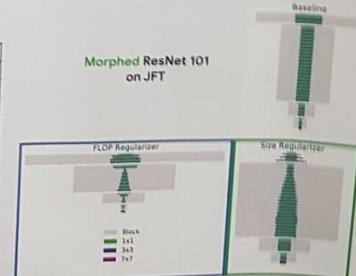
Specificity



Non-Human Design

Morphed Inception V2 on ImageNET				
	Branch 0	Branch 1	Branch 2	Branch 3
Mixed3b	170	0/0	112/63/96	69
Mixed3c	170	24/13	0/0/0	168
Mixed4a	75/82	66/69/154		
Mixed4b	573	121/107	124/53/83	258
Mixed4c	496	13/11	162/118/146	303
Mixed4d	424	94/90	80/61/91	255
Mixed4e	255	64/82	50/55/57	255
Mixed5a	324/482	237/113/635		
Mixed5b	936	11/17	258/178/200	341
Mixed5c	930	102/110	300/146/226	341

Morphed ResNet 101
on JFT



Try It Out

```
# The part of the code where you build losses & summaries
network_reg = flop_regularizer.GammaFlopsRegularizer([logits.op], my_threshold)
tf.summary.scalar('FLOPs', network_reg.get_cost())
tf.summary.scalar('Regularization', network_reg.get_regularization_term())
loss += strength * network_reg.get_regularization_term()
```



https://github.com/tensorflow/models/tree/master/research/morph_net

Low-level & Mid-level Vision

58.[I8] Matching Pixels Using Co-Occurrence Statistics

סוג של ממד דמיה המבוסס על סטטיסטיות גלובליות של סמיוטות בין פיקסלים, ומתמקד בעיקר בטקסטורות דומות ולא דוקא בדמיה גיאומטרית.

69.[I19] Learning a Discriminative Feature Network for Semantic Segmentation

למיידת מאפיינים לצורך סגמנטציה סמנטית תוך שימוש בשתי תתי-רשתות לשם 1. זהות גבולות, 2. צמצום מרחק (במרחב היזוג) בין דוגמאות שונות מסוימתה המחלקה.

3D Vision

77. [K21] Matryoshka Networks: Predicting 3D Geometry via Nested Shape Layers

שחזור מודל 3D עם פרטיהם המקוריים מתוך תמונה 2D – במאמר מופיעות תמונות יפות.

שיטת קודמות עובדות עם שכבות קונבולוציה 3D (שכבות הקונבולוציה הסטנדרטיות הן 2D), ומנסota ללמידה יצוג עבור ופיקסלים – לכל וкосל ערך אחר. הייצוג הזה מוגבל כיוון שדורש הרבה מאוד פרמטרים בראשת וכאן ניתן לקבל במצב נפח כולל של עד 32^3 ופיקסלים (אני מבין שנפח גדול יותר דורש רשת גדולה שאינה מעשית). השיטה המוצעת עשויה שימוש בשכבות קונבולוציה 2D וכן מאפשרת להשתמש ברשתות קיימות (המחברים אמרו שאצליהם ResNet הובילה לתוצאות הטובות ביותר). הרעיון המרכזי הוא לשזר בבת אחת את כל מידע העומק עבור קרן מסויימת העוברת דרך כל פיקסל.

80.[L8] End-to-End Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching

למיידת קצה לקצה של גלאי + descriptor עבור תמונות 3D.

נתון מודל CAD של אובייקט. באמצעות המודל מייצרים מפות עומק עבור נקודות מבט שונות (ערכים אופיינים: עצמים בגודל $3 \times 3 \times 4$, נקודות מבט משתנה ב-20-30 מעלות + הזזה). מגדרים דוגמא חיובית על ידי שינוי קטן בנקודות המבט על אותו העצם. מכנים את 2 תמונות העומק החביבות לרשת R-CNN אשר: 1. מציע איזורי עניין keypoints (זהו הגלאי) ו- 2. מחלצת דסקריפטור עבור כל איזור עניין.

פונקציית המהיר מכילה 2 איברים: 1. Contrastive loss, 2. איבר שנותן ציון חיובי למספר ההתאמות הנכונות.

זמן ריצה: בערך ms 140 לעיבוד תמונה בגודל 640x480.

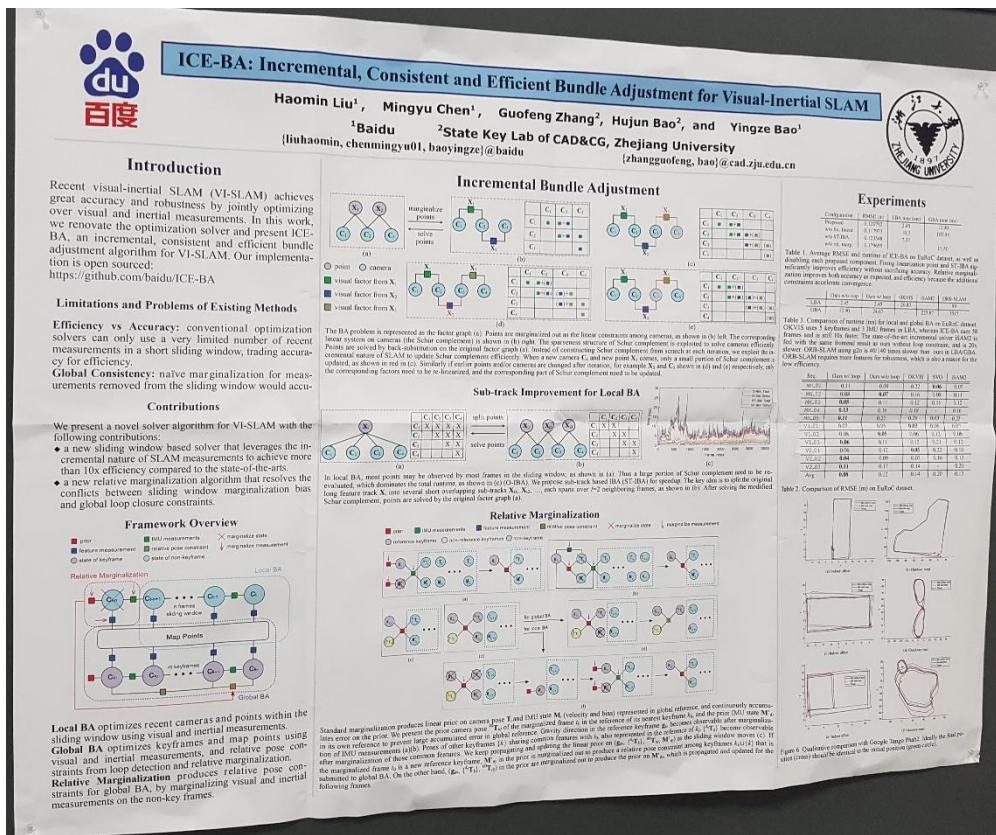
81. [L11] ICE-BA: Incremental, Consistent and Efficient Bundle Adjustment for Visual-Inertial SLAM

היתוך מידע אינרציאלי וחוזקי בSLAM בצורה איטרטיבית עם יעילות חישובית גבוהה.

בשיטת קיימות יש טריד-אוף בין דיק להמירות. השיטה המוצעת משתמש בחלוון רץ המכיל ח מצלמות אחרונות, ובנוסף משתנה (סקלר) המכיל את ההיסטוריה (שקדמה לח המצלמות). לדברי המציג, השיטה זו נקראת marginalization, והיא דומה מאוד ל-EKF, אלא ש-EKF מכיל חלוון רץ של דוגמה אחת בלבד.

אפשרות נוספת לשיפור הדיק המוצגת בעבודה זו היא שימוש בלולאות, בהן העצם עושה מסלול סגור וחוזר לנקודת השניה בה בעבר. שימוש באילוץ זהה מוביל לתוצאות הטובות ביותר ביום. ללא האילוץ מקבלים תוצאות דומות לטובות ביותר ביום.

בשני המקרים (עם ו ללא אילוץ לולה) זמן הריצה של השיטה המוצעת קצר משמעותית מהשיטות הקיימות, בערך פי-10-20.



1450-1630 Session 1-2C: Machine Learning for Computer Vision II (Room 255)

1450 Orals (01-2C)

1. [E10] Learning to Find Good Correspondences

מציאת שיכים בין זוג תמונות: בהינתן מספר קטן של התאמות אפשריות ונתוני מצלמה אינטראקטיביים, הרשות מחשבת מספר גדול יותר של שיכים טובים' ומחשבת את הערך Essential Matrix.

נראה מרשימים מאוד. העבודה עוסקת בבעיה של מציאת Inliers טוביים לשם חישוב essential matrix: בהינתן סט שיכים אפשריים, צריך מגננו לבחור את השיכים הטובים ויסון חריגים באופן יעיל. לשם חישוב פתרון אפשרי לessential matrix דרושים 4 שיכים - 4 זוגות של נקודות שמנחים שיש התאמה ביניהם, כאשר נקודת מוגדרת על ידי זוג קואורדינטות ע,א. בשיטה זו מכנים את כל השיכים לרשות אשר מוציאה ציון לכל שיור. הציון מחושב תוך התחשבות בכל שאר השיכים ולכן הוא גלובלי. לאחר מכן מוחשבים באמצעות ל השיכים ווקטור הציונים (כלומר לא מסכנים שיכים עם ציון נמוך, אלא משתמשים בכל השיכים וממשקלים אותם לפי הציונים) את האxivים (הכל באמצעות רשת עמוקה).

עם זאת, יכולה להיות בעיה של יציבות נומרטיות בעקבות גזירת האxivים (back propagation), ולכן תחיליה. מאENNIM עם פונקציית מחיר המזהה חריגים, ולאחר הוכנסות ראשונית מוסיפים את חישוב האxivים.



Learning

Kwang Moo Yi^{1,*} Eduard

¹Visual Computing Group

³Sony Imaging Products & Solutions

Outlier Re

Contributions

We solve **sparse correspondence** with **deep networks**. Classical pipeline: (a) find putative matches (e.g. SIFT); (b) find inliers (e.g. RANSAC); (c) retrieve camera motion.

Our approach:

- **Input:** correspondences. **Output:** weights.
- Unordered data → **multi-layer perceptrons**.
- **Global context** from non-parametric units.
- **Hybrid loss:** joint classification & regression.
- **State of the art** results on indoors/outdoors.

Collecting the Ground truth



Indoors: RGB-D

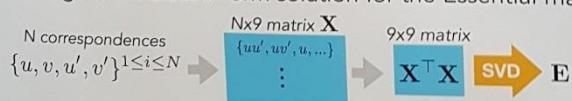


Outdoors: SfM

Can't have pixel-to-pixel correspondences. We propose to use **only the pose as ground truth**. We can recover it with off-the-shelf SfM.

Learning with regression

- 8-point algorithm: closed-form solution for the Essential matrix:



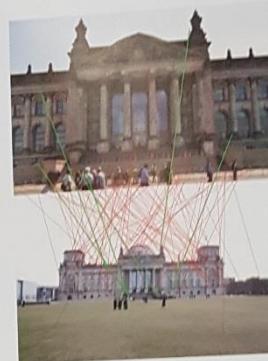
- Problem: **weak to outliers**. Solution: **weighted 8-point**, using the weights from the network: $\mathbf{X}^T \text{diag}(\mathbf{w}) \mathbf{X}$. Fully differentiable.

$$\mathcal{L}_e(\Phi, \mathbf{x}_k) = \min \left\{ \| \mathbf{E}_k^* \pm g(\mathbf{x}_k, \mathbf{w}_k) \|^2 \right\}$$

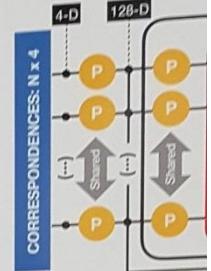
Learning with classification

- Learning outlier rejection implicitly by regressing the pose is too hard. **Network does not converge**.
- Solution: create **training labels from epipolar constraints**. How? threshold over the symmetric epipolar distance. Noisy but good enough!
- Loss: standard binary cross-entropy.

$$\mathcal{L}_x(\Phi, \mathbf{x}_k) = \frac{1}{N} \sum_{i=1}^N \gamma_k^i H(y_k^i, S(o_k^i))$$



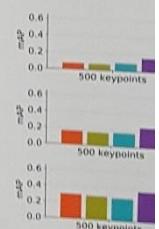
$$\mathbf{R}, \mathbf{t}$$



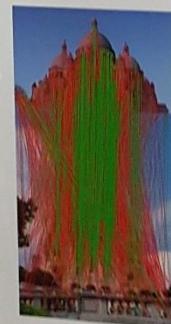
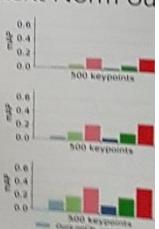
- **Input:** N correspondences
- **Problem:** input invariant. Not feasible for SfM.
- **Solution (PointNet)**: Deep network:
- Each point is processed independently. PointNet solution.
- **Our solution:** efficient.

Ablation: Invariant

- Classification reduces error. Does best: $\mathcal{L}_c(\Phi)$



- PointNet-style context norm outperforms Context Norm ou



EPFL UVIC
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
TU SONY
Graz

Learning to find good correspondences

Kwang Moo Yi^{1,*}, Eduard Trulls^{2,*}, Yuki Ono³, Vincent Lepetit⁴, Mathieu Salzmann², Pascal Fua²

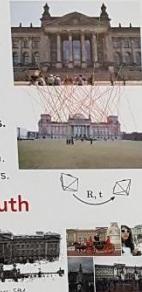
¹Visual Computing Group, University of Victoria ²Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne
³Sony Imaging Products & Solutions Inc. ⁴Institute for Computer Graphics and Vision, Graz University of Technology (*: equal contribution)

Contributions

solve sparse correspondence with deep works. Classical pipeline: (a) find putative matches (e.g., SIFT); (b) find inliers (e.g., RANSAC); (c) retrieve camera motion.

Our approach:
Input: correspondences. **Output:** weights.
 Unordered data → multi-layer perceptrons.
Global context from non-parametric units.
Hybrid loss: joint classification & regression.
State of the art results on indoors/outdoors.

Collecting the Ground truth

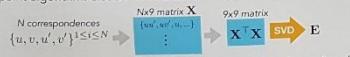


Indoor: RGB-D Outdoor: SM

Can't have pixel-to-pixel correspondences. We propose to use only the pose as ground truth. We can recover it with off-the-shelf SfM.

Learning with regression

- 8-point algorithm: closed-form solution for the Essential matrix:



- Problem: weak to outliers. Solution: weighted 8-point, using the weights from the network: $\mathbf{X}' \text{diag}(\mathbf{w}) \mathbf{X}$. Fully differentiable.

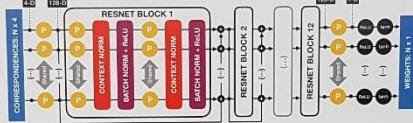
$\mathcal{L}_e(\Phi, \mathbf{x}_k) = \min \left\{ \| \mathbf{E}_k^T \pm g(\mathbf{x}_k, \mathbf{w}_k) \|^2 \right\}$

Learning with classification

- Learning outlier rejection implicitly by regressing the pose is too hard. Network does not converge.
- Solution: create training labels from epipolar constraints. How? threshold over the symmetric epipolar distance. Noisy but good enough!
- Loss: standard binary cross-entropy.

$\mathcal{L}_c(\Phi, \mathbf{x}_k) = \frac{1}{N} \sum_{k=1}^N \gamma_k^i H(y_k^i, S(o_k^i))$

Outlier Rejection Network



- Input: N correspondences $\{u_i, v_i, u'_i, v'_i\}_{1 \leq i \leq N}$. Output: N weights.
- Problem: input data is unordered. Output should be permutation-invariant. Not feasible with e.g. convolutional or fully-connected layers.
- Solution (PointNet): Multi-layer, weight-sharing perceptrons.
- Deep network: 12 resnet-style blocks. Still very small!
- Each point is processed individually! We need contextual information.
- PointNet solution: global feature, pooled with a second network.
- Our solution: embed into the feature maps with Context Normalization.

Context Normalization

- Simple, non-parametric normalization. Given features $o_i^l \leq N$ at layer l :

$$CN(o_i^l) = \frac{(o_i^l - \mu^l)}{\sigma^l}$$

$$\mu^l = \frac{1}{N} \sum_{i=1}^N o_i^l, \quad \sigma^l = \sqrt{\frac{1}{N} \sum_{i=1}^N (o_i^l - \mu^l)^2}$$

- Similar to BN/LN, but nothing is learned. Same operation for training/inference.
- Operates separately over image pairs:



- In image stylization: Instance Norm.

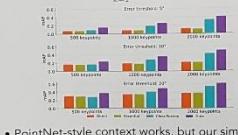
Qualitative results

- Top: RANSAC. Bottom: Our Drawing inliers only. Pictures they are below the ground truth distance threshold, and in red.



Ablation: Loss & Context

- Classification required to converge. Hybrid loss does best: $\mathcal{L}_c(\Phi) = \sum_{k=1}^K (\alpha \mathcal{L}_c(\Phi, \mathbf{x}_k) + \beta \mathcal{L}_e(\Phi, \mathbf{x}_k))$

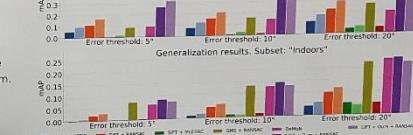


- PointNet-style context works, but our simple Context Norm outperforms it on this problem.



Evaluation

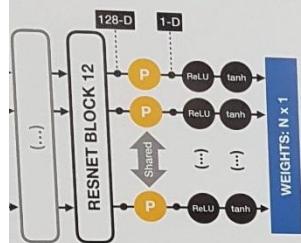
- Datasets: indoors (SUN3D) and outdoors (YFCC100+SFm).
- Our models are trained on a single sequence from each.
- Baselines: sparse (RANSAC variants, GMS) & dense (G3DR, DeMoN).
- Metric: angular error between ground-truth & estimated R/T. Determine accuracy by thresholding at varying values & compute mAP. Generalization results. Subsets: "Outdoors"



- Outdoors: great improvements. Indoors: still better than dense SoA.
- For testing we do not need differentiability! We run ours (one forward pass) then RANSAC. Improves performance (2x) and speed (17x).

Good correspondences

Incent Lepetit⁴, Mathieu Salzmann², Pascal Fua²
 Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne
 Computer Graphics and Vision, Graz University of Technology (*: equal contribution)



Output: N weights.
 Should be permutation-
 or fully-connected layers.
 Using perceptrons.
 Very small!
 Add contextual information.
 Through a second network.
 with Context Normalization.

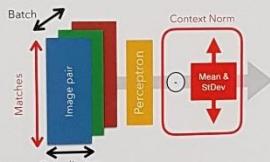
Context Normalization

- Simple, non-parametric normalization.
 Given features \mathbf{o}_i^l at layer l :

$$CN(\mathbf{o}_i^l) = \frac{(\mathbf{o}_i^l - \mu^l)}{\sigma^l}$$

$$\mu^l = \frac{1}{N} \sum_{i=1}^N \mathbf{o}_i^l, \quad \sigma^l = \sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{o}_i^l - \mu^l)^2}$$

- Similar to BN/LN, but nothing is learned.
 Same operation for training/inference.
- Operates separately over image pairs:

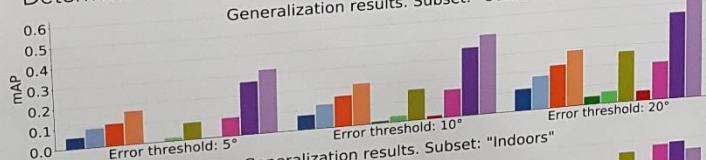


- In image stylization: Instance Norm.

Evaluation

- Datasets:** indoors (SUN3D) and outdoors (YFCC100+SFM).
- Our models are trained on a **single sequence from each**.
- Baselines:** sparse (RANSAC variants, GMS) & dense (G3DR, DeMoN).
- Metric:** angular error between ground-truth & estimated R/T.
 Determine accuracy by thresholding at varying values & compute mAP.

Generalization results. Subset: "Outdoors"



Generalization results. Subset: "Indoors"



- Outdoors:** great improvements. Indoors: still better than dense SoA.
- For testing we do not need differentiability! We run ours (one forward pass) then RANSAC. Improves performance (2x) and speed (17x).

Qualitative results

- Top: RANSAC. Bottom: Ours. Same input.
- Drawing **inliers only**. Pictured in green if they are below the ground truth epipolar distance threshold, and in red otherwise.

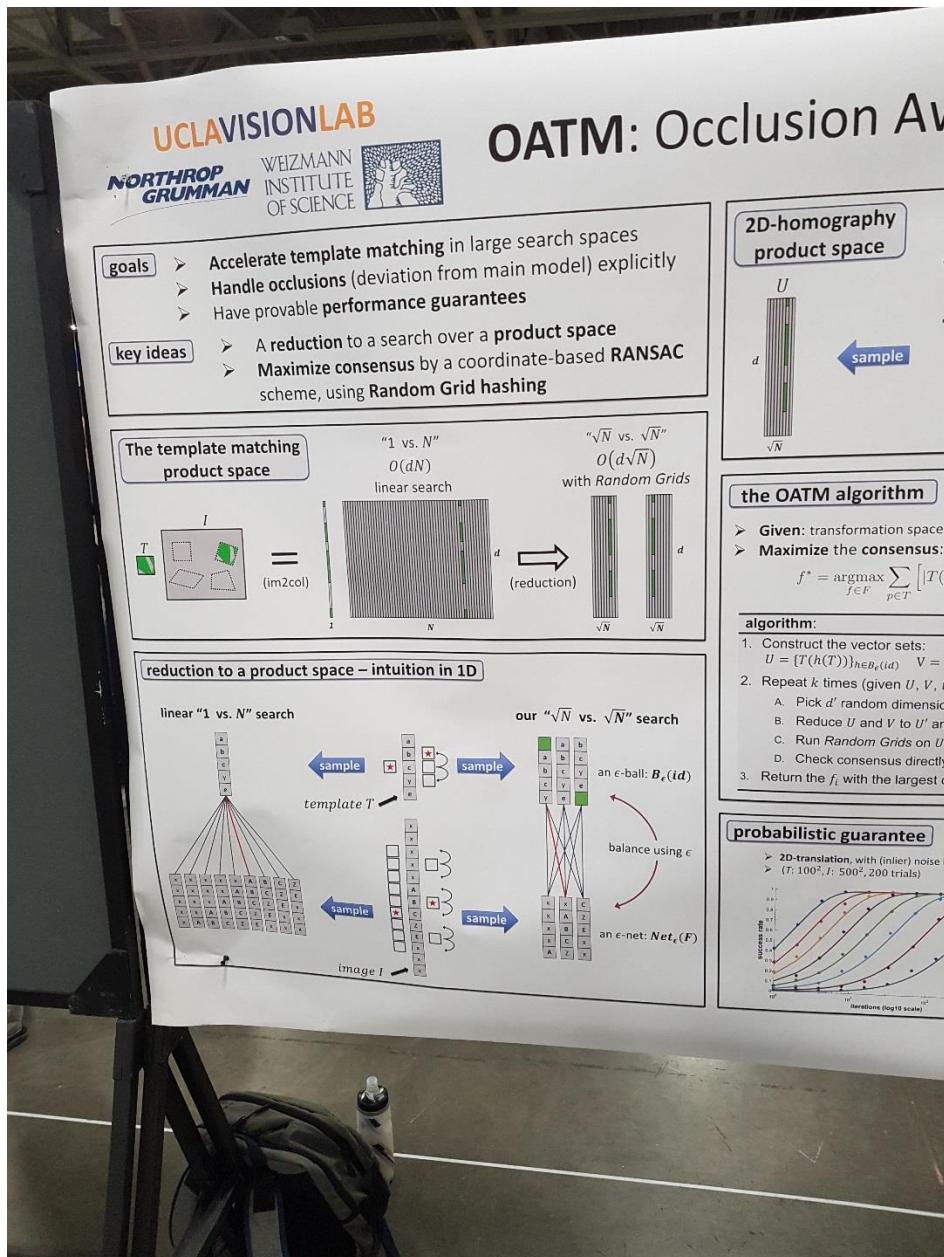


2. [E13] OATM: Occlusion Aware Template Matching by Consensus Set Maximization

Template Matching ייעיל, ללא למידה עמוקה.

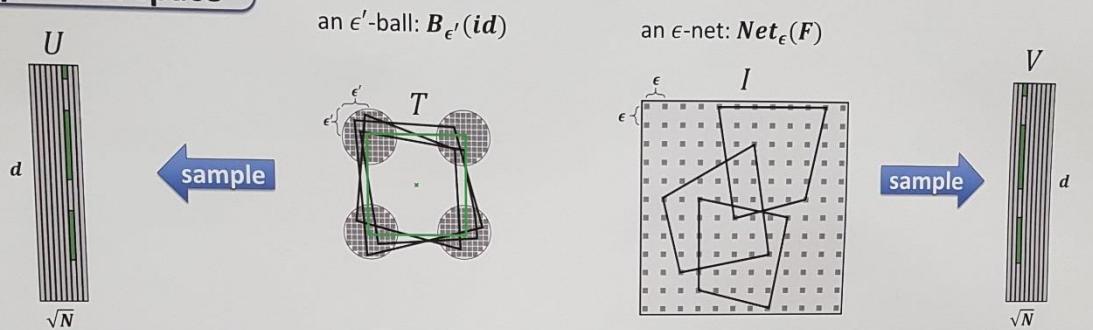
שיטת ליביצוע template matching בצורה ייעילה, ובנוסף תוך התחשבות בהסתירות occlusions (החידוש העיקרי) הוא החישוב הייעיל, התחשבות בהסתירות היא בונוס).template matching מבצע במרחב התמורות (לדוגמא התמורות אפשריות או הומוגרפיות). אם רוצים לחשב את כל האפשרויות של התמורות מגעימם למספרים גדולים מאוד, ולא מעשיים. בעבודה זו: 1. עושים שימוש בשיטת חיפוש מתוחכמת המפחיתה את הסיבוכיות $O(dN\sqrt{d})$ (ולא $O(dN^2)$) כאשר d הוא מספר הפיקסלים בטמפלט N והוא מספר האפשרויות לטמפלט כזה בתמונה (אולי הכוונה למספר התמורות האפשריות?). 2. תקונה נוספת היא התמודדות עם הסתרות – מבוצע באמצעות דגימת חלק מהפיקסלים בטמפלט (ולא כל הטמפלט) ושלוב בลอואת RANSAC.

בסוף הוסבר על שימוש בHPatches על מנת לבדוק את הביצועים. כיוון שמדובר בשיטת השוואת טמפלט לא נעשה שימוש במצבית שכן קרוב, אלא הרכיבו תמונה באמצעות קולאז' של חלונות ואחר כך ביצעו template matching בין חלון השאלתה לבין הקולאז'. המציג אמר שזו שיטה נאיבית, ולמרות זאת הגיעו לתוצאות טובות. שיטות מתקדמות יותר צפויות להוביל לתוצאות טובות אף יותר.



Occlusion Aware Template Matching

2D-homography product space



the OATM algorithm

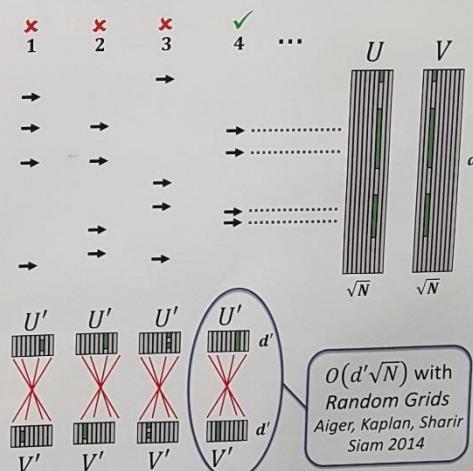
- Given: transformation space F; threshold t
- Maximize the consensus:

$$f^* = \operatorname{argmax}_{f \in F} \sum_{p \in T} [|T(p) - I(f(p))| \leq t]$$

algorithm:

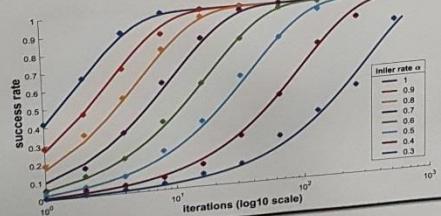
- Construct the vector sets:
 $U = \{T(h(T))\}_{h \in B_{\epsilon}(id)}$ $V = \{I(f(T))\}_{f \in Net_{\epsilon}(F)}$
- Repeat k times (given U, V, t) to obtain $\{f_i\}_{i=1}^k$
 - Pick d' random dimensions out of $1, \dots, d$
 - Reduce U and V to U' and V'
 - Run Random Grids on U' and V'
 - Check consensus directly on U and V
- Return the f_i with the largest consensus set

Iterations



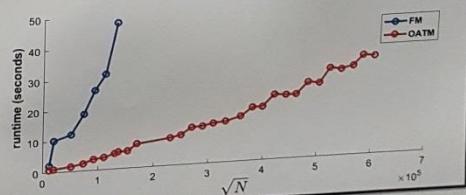
probabilistic guarantee

- 2D-translation, with (inlier) noise level 5 GLs
- ($T: 100^2, I: 500^2, 200$ trials)



scalability

- 2D-affine, with (inlier) noise level 5 GLs
- Template Dim: [32] image dims: [100 : 100 : 3000]



Matching

Simon Korman UCLA/WIS
Mark Milam NG
Stefano Soatto UCLA



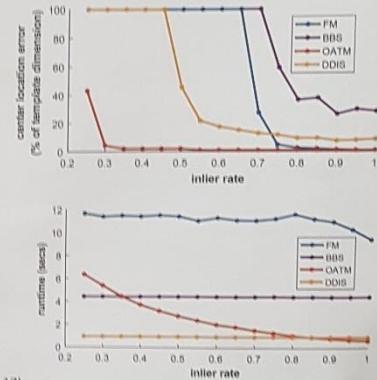
template matching benchmark

- 2D-affine, with (inlier) noise level of 5 GLs
 - A variety of template/image sizes:
- | | | |
|---------------|---------------|---------------|
| T3: 16 × 16 | T2: 32 × 32 | T3: 64 × 64 |
| H1: 160 × 160 | I2: 320 × 320 | I3: 640 × 640 |
- "Fast-Match" (FM) Korman, Reichman, Tsur, Avidan (CVPR 13)

	template-image sizes									
	T1-I1	T1-I2	T1-I3	T2-I1	T2-I2	T2-I3	T3-I1	T3-I2	T3-I3	
FM	0.09	0.13	NA	0.05	0.05	0.09	0.02	0.01	0.03	
	12.22	25.37	NA	4.35	7.78	32.07	1.33	1.90	11.61	
OATM	0.07	0.10	0.13	0.02	0.04	0.04	0.01	0.02	0.13	
	0.15	0.18	0.39	0.53	0.76	1.73	0.51	0.64	1.01	

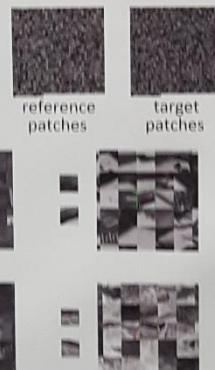
robustness to occlusion

- 2D-affine, with (inlier) noise level 5 GLs
- T2: 32 × 32 and I2: 320 × 320



"Best-Buddies Similarity" (BBS) Dekel, Oron, Rubinstein, Avidan (CVPR 15)
"Deformable-Diversity Similarity" (DDIS) Talmi, Mechrez, Zelnik-Manor (CVPR 17)

Hpatches dataset



method	viewpoint seqs			illumination seqs		
	Easy	Hard	Tough	Easy	Hard	Tough
BRIEF [7]	25.6	6.9	2.4	20.5	5.9	2.0
ORB [23]	36.4	11.1	3.7	28.9	8.8	3.2
SIFT [19]	59.4	30.6	15.3	52.6	26.1	13.3
TE-R [5]	58.9	35.5	19.0	48.5	28.6	15.6
DDIS [24]	58.6	36.0	20.2	50.7	30.0	17.0
RSIFT [3]	64.0	35.2	18.5	57.1	36.2	15.9
OATM	72.7	49.2	32.1	43.3	29.3	19.7

"HPatches" Balntas, Lenc, Vedaldi, Mikolajczyk (CVPR 17)

1548 Spotlights (S1-2C)

1. [E22] Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference

קונטיזציה לשם אימון וחיזוי ב Fixed Point – יכול להתאים למימוש ב-FPGA'.

המטרה: לדוח רשותות גדולות ליישומי זמן אמיתי, באמצעות ביצוע קונטיזציה. הקונטיזציה מתבצעת באמצעות התמירה אפינית (הגביר והזזה). ניתן לבצע גם על ווקטורים שלמים. האימון מתבצע עם קונטיזציה, כאשר יש גם פונקציות מחיר שדווגת לכך שההתוצאות יהיו דומות לרשות עט נקודת צפה. יש שתי אפשרויות לבצע קונטיזציה: מוקדים וקטיביות (לא ברור לי מה עושים בעבודה זו). לפि הבנתי ניתן לעשות זאת באמצעות tensorflow.tensors.

7. [F18] Learning Deep Descriptors With Scale-Aware Triplet Networks

בחינה של פונקציות מחיר וגישות שונות ללמידה descriptors טוביים.

קשה מאוד להבין את המרצה – מבטא אס'יטי גראע ...

העבודה עוסקת ב triplet networks – רשותות המכילות 3 ענפים עם אותה הארכיטקטורה. המרצה הציג בעיה מסוימת שקיימות ברטשות טריפלט ולא ברטשות סיאמיות, והסביר כיצד התמודד אליה (אולי – השתמש ברטשות סיאמיות לצורך מסויים?). בחלק של התוצאות הוצגה השוואת L2Net HardNet על בסיס נתונים HPatches ונאמר כי הרשות המוצעת הציגת תוצאות טובות יותר.

1630-1830 Poster Session P1-3 (Halls C-E)

3D Vision

9. [H7] Estimation of Camera Locations in Highly Corrupted Scenarios: All About That Base, No Shape Trouble

שערוֹן מיקום מצלמה בבעיה SFM עם רעש חזק.

העבודה עוסקת בשחזור מיקומי מצלמות (רבים) מתוך ידיעת וקטורי הcyton המצלמים מצלמה אחת לאחרת pairwise directions (pairwise directions), ללא ידע על המרחק האבסולוטי. יש אוסף של וקטורי cyion שכאה, כל אחד מקשר בין זוג מצלמות, מהם ניתן לבנות גרפ בוקודקווידים (vertices) הם המצלמות והקשתות (edges) הם (edges) (להבנתי כאן מעוניינים למצוא את האxitments בין כל המצלמות – אבל אני לא בטוח, אולי רוצים למצוא מיקומים מוחלטים בין כל המצלמות). וקטורי cyion הילו בדרך כלל מתקיים בשל חישוב מקדים לחישוב essential matrix. הבעיה היא שבדרך כלל יש חריגים outliers רבים המפריעים לפתרון – וכן מוצאה שיטה לזייה חריגים הילו. הרעיון הוא להשתמש באילוץ גיאומטרי של משולשים בגרף. בהינתן שלשה של מצלמות בה יש קו ראייה בין כל זוג מצלמות, השלשה הזאת תופיע כמשולש סגור בגרף (וקטורי cyion שליהם יוצרים משולש). אם בפועל המשולשים הילו לא סגורים – יכול להיות שיש חריג. בעבודה זו הוצג ממד סטטיסטי הבודק כמה משולשים יש בעיה עם החסרים pairwise distance – pairwise direction (כל זוג מצלמות – זוג מצלמות – יכול להשתתף במסולשים רבים) על ידי סיכום משוכל של כל המשולשים המכילים אותו. הוצגו 2 שיטות לביצוע השקלול. הוצג שיפור מרשימים מאוד ביחס לשיטות קודמות (עשרות אחוזים).

21. [I21] Camera Pose Estimation With Unknown Principal Point

שערוֹן מודל מצלמה כאשר principal point אינו ידוע.

למעשה שני דברים אינם ידועים: 1. Principal point (הנקודה בה האופטי חותך את מישור התמונה) – בדרך כלל במרכז התמונה, אך לא תמיד. לדוגמה כאשר ממצאים חיתוך crop (לא ידוע) לתמונה. 2. אורך מוקד focal length.

הוֹצָג פתרון מינימלי minimal solution לשחזור מיקום מצלמה 3D מתוך זוג תמונות 2D המשמש ב 4.5 נקודות: הפתרון דורש 9 מושוואות, וכל נקודה מספקת 2 מושוואות (עבור x וy), ולכן עבור אחת מהנקודות משתמשים רק במושואה אחת (חצי נקודת). בנוסף פותרים גם עבור עיוות רדיאלי radial distortion – אך על ידי רלקסציה ולא באופן מדויק (הפתרון ללא עיוות הוא מדויק).

Camera Pose Estimation with Unknown Principal Point

Viktor Larsson
Lund University
Lund, Sweden
viktorl@maths.lth.se

Zuzana Kukelova
Czech Technical University
Prague, Czech Republic
kukelzuz@fel.cvut.cz

Yingqiang Zheng
National Institute of Informatics
Tokyo, Japan
yqzheng@nii.ac.jp

Overview

In most of existing camera pose estimation solvers, the principal point is assumed to be in the image center. Unfortunately, this assumption is not always true, especially for asymmetrically cropped images. In this paper, we develop the first exactly minimal solver for the case of unknown principal point and focal length by using four and a half point correspondences ($P4.5PFuv$). We also present an extremely fast solver for the case of unknown aspect ratio ($P5PFuv$). The new solvers outperform the previous state-of-the-art in terms of stability and speed. Finally we also consider the case of both unknown principal point and radial distortion.

Contributions

- New polynomial constraints for unit aspect ratio and zero skew.
- First minimal $P4.5PFuv$ solver (unit aspect ratio and zero skew).
- Extremely fast $P5PFuv$ solver (zero skew).
- Solver for unknown principal point and radial distortion using 7 points ($P7PFuv$).

Unknown Focal Length and Principal Point ($P4.5PFuv$)

We want to estimate pose (R, t) , focal length f and principal point $\mathbf{x}_0 = (a_0, n_0)$

$$\begin{pmatrix} \mathbf{x}_1 - \mathbf{x}_0 \\ f \end{pmatrix} \approx \begin{bmatrix} I & 0 \\ 0 & 1 \end{bmatrix} (R\mathbf{X}_1 + t). \quad (1)$$

Minimal with 4.5 point correspondences

Previous work by Triggs [7]

- Non-minimal solver using 5 point correspondences.
- Minimizes $\|R\mathbf{X}_1 + t - \mathbf{x}_1\|^2$.
- Sensitive to noise due to non-minimal parameterization.

Zero Skew and Unit Aspect Ratio

The constraints for a camera matrix to have zero skew and unit aspect ratio are well known

Theorem 1 [Triggs [1], Heyden [2]].

The camera matrix P has zero skew and unit aspect ratio if and only if

$$\begin{aligned} P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 \\ \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_4 \end{bmatrix}, & \det[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3] \neq 0 \quad (2) \\ \mathbf{p}_1 \times \mathbf{p}_2 = \mathbf{p}_1 \times \mathbf{p}_3 = 0, & \quad (3) \\ \mathbf{p}_1 \times \mathbf{p}_2 = \|\mathbf{p}_1\|^2 = \|\mathbf{p}_2\|^2 = 0. \quad (4) \end{aligned}$$

If only (3) holds the camera has non-unit aspect ratio.
Although formulated differently, the constraints (3) and (4) are equivalent to the ones used in Triggs [7].

New Camera Matrix Constraints

- Non-zero constraint in (2) is difficult to handle
- Ignoring it introduces false positives (and redundant constraints)
- Computing solutions of (3) and (4) with respect to (2) yields 5 polynomial constraints of degree 5.
- Shows the same set of the new constraints:

$$\begin{aligned} p_1p_2p_3p_4 + p_1p_2p_3^2 - p_1^2p_2p_3 = -p_1^2p_2p_4 + p_1p_2p_3p_4 + p_1^2p_3p_4 + p_1p_2p_3^2 = 0 \\ p_1p_2p_3^2 + p_1p_2p_4^2 - p_1^2p_3p_4 = -p_1^2p_3p_4 + p_1p_2p_3p_4 + p_1^2p_3p_4 + p_1p_2p_3^2 = 0 \\ \text{Holds for any constraints satisfying (3), (4) and (2)} \end{aligned}$$

New solver ($P4.5PFuv$)

• Nullspace parametrization for square matrix

• 4.5 points \rightarrow 9 linear constraints \rightarrow 2 unknowns after fixing scale:

$$\begin{cases} \mathbf{x}_1 R \mathbf{X}_1 - P \mathbf{X}_1 = 0 \\ \mathbf{x}_1 R \mathbf{X}_1 - P \mathbf{x}_1 = 0 \end{cases} \implies P = N_0 + \alpha_1 N_1 + \alpha_2 N_2$$

• Constrains new solvers using automatic generator from Larson et al [4]

• New constraints give smaller template and fewer solutions

New solver ($P5PFuv$)

We also consider only zero skew cameras (i.e. unknown aspect ratio as well)
 $\rightarrow 3$ points \rightarrow 10 linear constraints \rightarrow 1 unknown after fixing scale.

- Inserting subpace parametrization into (3) \rightarrow univariate quartic equation

- Closed form solution or Jenkins method (companion matrix)

Radial Distortion with Unknown Principal Point

We also consider radial distortion using one-parameter division model.
The problem is natural with 3 point correspondences.

Experiments

Experiment with Shifted Principal Points

Experiment with Shaded Principal Points and Radial Distortion

Code is available at www.maths.lth.se/~viktor/

References

[1] B. Triggs, Photometric stereo computer vision, a perspective survey, 1991.

[2] D. H. Ballard, Generalizing the Hough transform to detect arbitrary shapes, 1981.

[3] J. C. Oliensis and T. Pavlidis, Efficient and robust estimation of camera parameters, 1991.

[4] V. Larsson, Z. Kukelova, and Y. Zheng, Efficient and robust camera pose estimation with unknown principal point, 2013.

[5] V. Larsson, Z. Kukelova, and Y. Zheng, Camera pose estimation with unknown principal point and radial distortion, 2014.

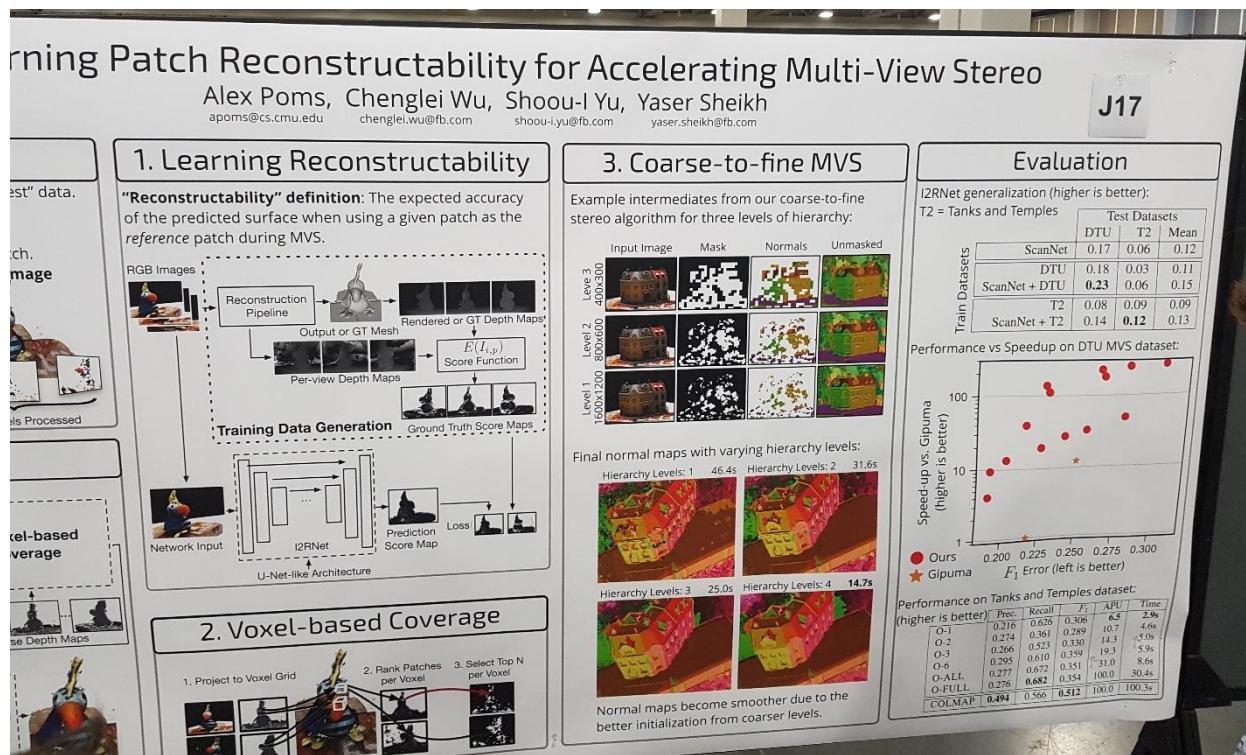
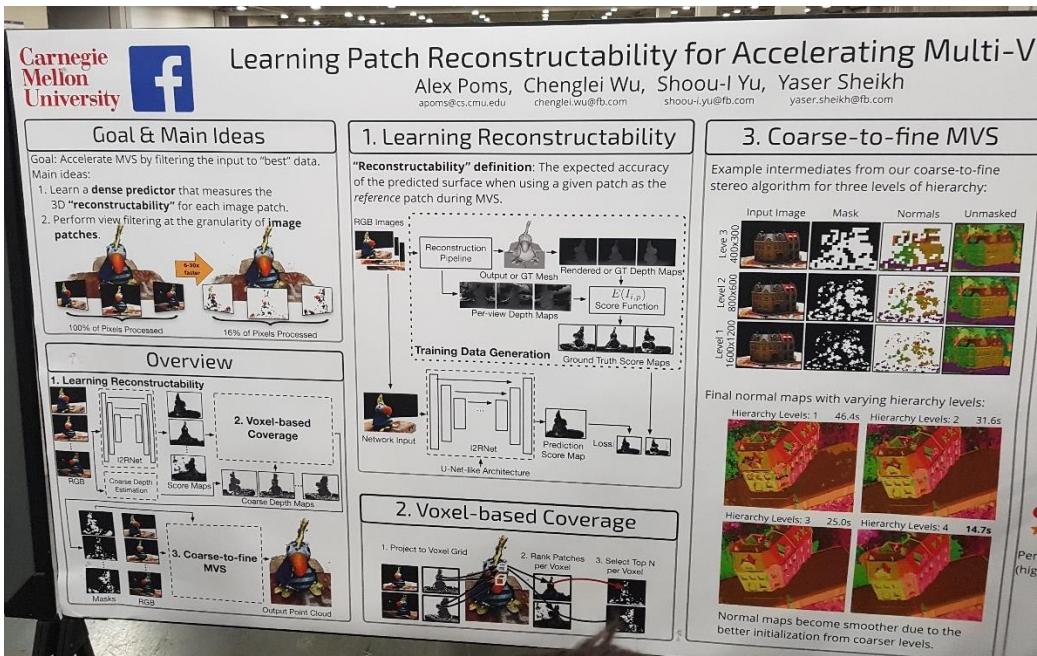
[6] V. Larsson, Z. Kukelova, and Y. Zheng, Camera pose estimation with unknown principal point and radial distortion, 2014.

[7] B. Triggs, A feature-based approach to visual servoing, 1996.

27. [J17] Learning Patch Reconstructability for Accelerating Multi-View Stereo

שיטת למידויים מאפיינים אשר יובילו לשחזור מודל מצלמה טוב יותר. מעוניין – לבדוק!

העבודה עוסקת בבעיה של שחזור תלת מימדי מתוך תמונות מסוימות מבעט multi view 3D reconstruction כאשר מנחים שמייקם המצלמות ידועה וצריך לשחזר רק את מיקום העצם בעולם (רעיון מעוניין הוא להכפיל את השיטה לפתרון מיקום המצלמות והעצמים במקביל M^{Sf}). לדברי המציג השיטות הקיימות כבדות מבchnה חישובית, והשיטה המוצעת מהירה משמעותית. הרעיון הוא שיש יתרונות redundancy רבים: אוטם היפיקסלים מופיעים במספר גדול של תמונות, ואפשר להשתמש רק בחלק קטן 'איכות' מהתמונות עבור כל נקודה בעולם. לשם כך אימנו רשת שטייצר מפת ציונים – עבור כל פיקסל מה החסודות לשחזור מוצלח של הנקודה 3D המתאימה לו בעולם תוך שימוש בתמונה הזו. לאחר מכן מחשבים מפות עומק גסות, מחלקים את המרחב לווקסלים, ומשייכים כל חלון (16x16) לווקסל. בכל ווקסל בוחרים רק מספר קטן של חלונות עבור השחזור בתבסס על הציונים שהרשף יצרה.



Machine Learning for Computer Vision

54.[N10] BPGrad: Towards Global Optimality in Deep Learning via Branch and Pruning

שיטת אופטימיזציה חדשה המתקדמת לכיוון של חיפוש אופטימום גלובלי ולא מקומי. לטענת המחברים מובילה לביצועים טובים יותר מאשר Adam, Adagrad ועוד.

58.[N22] Domain Adaptive Faster R-CNN for Object Detection in the Wild

שיפור הרובוטיות של רשת Faster RCNN כך שתבצע טוב יותר במספר דומיינים.

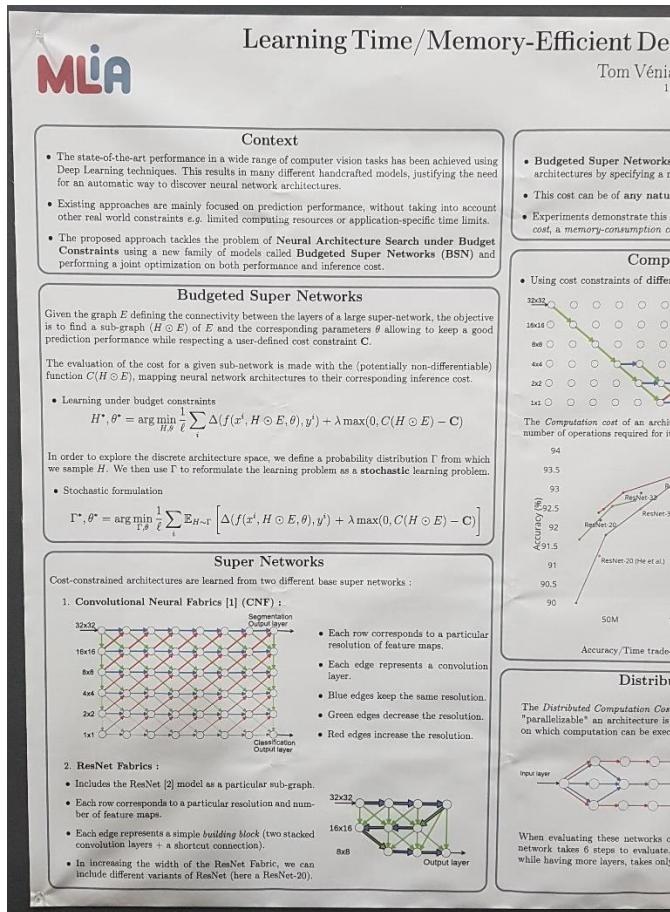
61. [09] Lightweight Probabilistic Deep Networks

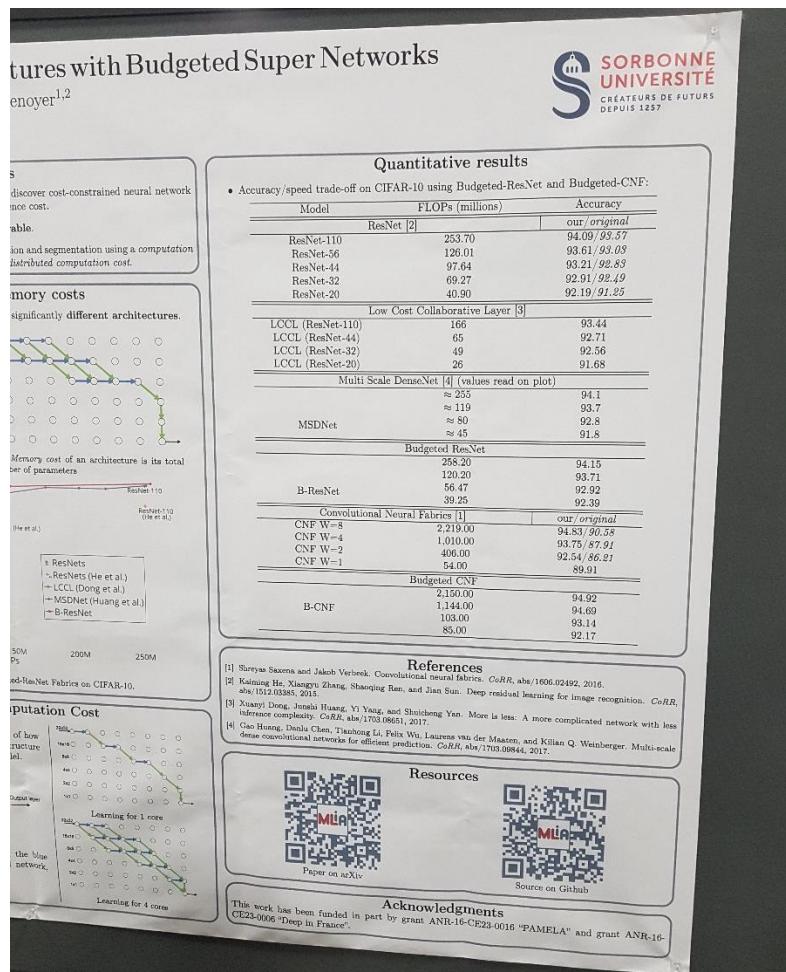
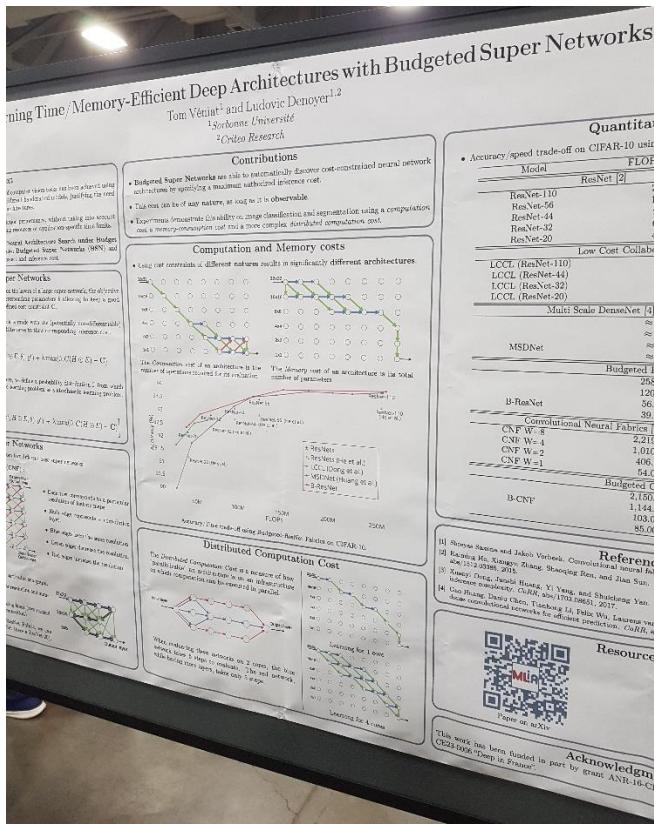
רשות המוציאים ציון הסתברותי (ולא דטרמיניסטי). נשמע נחמד – רק אם יש זמן.

74. [Q4] Learning Time/Memory-Efficient Deep Architectures With Budgeted Super Networks

למייד רשות עם זמן חישוב מהיר \ חתימת זיכרון קטנה.

רעיון קצר דומה ל-MorphNet של גול. משתמשים ב-CNF – Convolution Neural Fabrics – סוג של גרף המגדיר משפחה של ארכיטקטורות אפשריות, כאשר המטריה היא לומודת תרשים (ארכיטקטורה ספציפית) אשר ימקסם מטרות מסוימות דוגמת מינימום זיכרון או מינימום FLOPS. להבנתו עושים שימוש ב-reinforcement learning על מנת לומוד ארכיטקטורה מתאימה.





20.06.2018

0830-1010 Session 2-1C: 3D Vision III (Room 255)

0830 Orals (O2-1C)

1. [E8] Density Adaptive Point Set Registration

התאמת (צפופה?) של ענני נקודות באמצעות שימוש במבנה האמייתי (ממנו נדגמו ענני הנקודות) כמשתנה חופשי.

נתון ענן נקודות המתאר סצנה. הענן דגום באופן לא אחיד כך שחלקים מסוימים מהסצנה מותאים על ידי מספר גדול של נקודות וחלקים אחרים על ידי מספר קטן או כליל לא. המטרה היא לבצע ריגיסטרציה בין מספר ענני נקודות שכאלן. הגישה היא הסתברותית: לתאר כל ענן נקודות על ידי מודל הסתברותי המתאר את המיקום של כל הנקודות בענן ואת נקודות המבט (מצלמה – סיבוב והזזה) על מנת שייה אפשר לעננים שצולמו מנקודות מבט אחרות. המודל בו נעשה שימוש הוא Gaussian Mixture Model GMM עם 200 רכיבים.

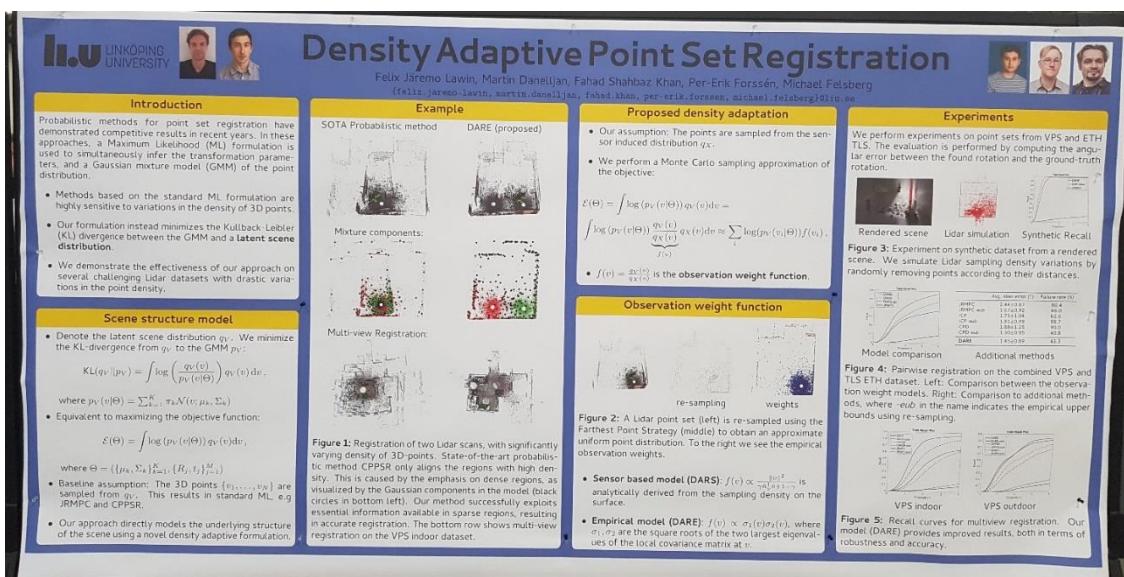
בעבודות קודמות הריכבים התמקדו בעיקר באיזורי הדגומים בצורה צפופה, ולכן הריגיסטרציה הייתה טובה דזוקא באיזורים אלו ופחות טובה עבור איזורים אחרים.

בעבודה זו הניחו שהפיגול האמייתי של הנקודות בסצנה הוא אחיד, וענן הנקודות הוא דגימה של הפיגול האמייתי אחריות, הפיגול האמייתי הוא משתנה חופשי latent אותו אנו מעריכים לשערך. אנו מעריכים למינן את הפיגול האמייתי – ולא את ענן הנקודות הנמדד – באמצעות GMM. המידול מתבצע על ידי הקטנת KL Divergence בין הפיגול האמייתי לבין ה-GMM.

הרעין המוצע הוא למשקל את הגאוסיאנים כך שאיזורים עם צפיפות גבוהה ימודלו על ידי מספר קטן של גאוסיאנים, ואילו איזורים עם צפיפות נמוכה ימודלו על ידי מספר גדול: המשקל הופכי לצפיפות.

הציגו 2 שיטות לחישוב המשקל. שיטה אחת המנסה לשערך את מודל הפיזיקלי של החישון, ומתוכו להבין כיצד החישון דגם את ענן הנקודות מתוך הפיגול האמייתי. לשם כך צריך מודל של רכישת תמונה על ידי LIDAR ומודל של 3D הסצנה. בפועל המציג אמר שלא עבד טוב, ככל הנראה כיון שהמודל האמייתי הוא מורכב בעוד המודל שלהם פשוט. השיטה השנייה לשערוך הцеיפות היא אמפירית: עברור כל נקודה לקחו את 10 השכנים הקרובים וחישבו את הפיזור של הקבוצה הזאת, אם היא מוקובצת צפופה, נתנו משקל נמוך, וההיפך. בפועל חישוב הפיזור נעשה על ידי SVD: שמו את הקואורדינטות של 10 הנקודות במטריצה, חישבו SVD, ולקחו את 2 הערכים הסינגולריים הגבוהים ביותר (σ_2, σ_1). נפרט מעת: נסמן על ידי q את הפיגול הנמדד ועל ידי f את מודל ה-GMM המשער את המודל האמייתי. מניחים שהפיזור הוא $\sigma_2 \cdot \sigma_1$. מכאן ניתן למצער את ה- $f = \frac{q_x}{q_y}$ – וcutout את המודל האמייתי. מנגנון שפהיזור הוא $\sigma_2 \cdot \sigma_1$.

אפשר להבין את המשוואות המופיעות בפוסטර.



3. [E14] Im2Pano3D: Extrapolating 360° Structure and Semantics Beyond the Field of View

לא סיכמתי את כל ההרצאה, רק רעיון מעניין אחד: כיצד להציג \לייצג מידע תלת מימדי? שתי אפשרויות מקובלות הן שימוש במפות עומק (חישרין – תליי חזק בנקודת המבט) ונורמלים למשטחים (חישרין): לא ניתן לשחרר מפות עומק under determined problem (השיטה המוצעת היא להשתמש במשוואת מישור המכילה למעשה שני חלקים: 1. נורמל למשטח, 2. מרחק מראשית הצירם).

0928 Spotlights (S2-1C)

3. [F4] Tangent Convolutions for Dense Prediction in 3D

סגןנטציה סמנטית לסצנות 3D גדולות. עושים שימוש בtangents Tangent convolutions היא למצוא את המשיק למשטח לכל נקודה בתמונה (?). היתרונות בעבודה הם שמדובר בחישוב יעיל – עבור 125K נקודות לוקח בערך 2.1 שניות עבור עיבוד מקדים + חיזוי ובערך 2GB זיכרון.

4. [F7] RayNet: Learning Volumetric 3D Reconstruction With Ray Potentials

מערכת לשחזור תלת מימדי מבוססת קרניים (rays) ולמידה עמוקה. עושים שימוש בMarkov Random Field MRF שמשתמש בפונקציית המחר. הMRF עשו לאחר הרשת העמוקה על מנת ללמד את כל הסצינה – תוצאות MRF משתתפות בפונקציית המחר. מושג בשפה הנפוצה – מקשר בין ווקסלים שונים בנפח באמצעות אלומטרים, וכן מאפשר להציג לתוצאות טובות יותר מאשר רשתות מבוססות למידה עמוקה בלבד אשר לא מתחשבות באילומטרים.

5. [F10] Neural 3D Mesh Renderer

שחזור (רנדור) מודל תלת מימדי של רשת mesh מתוך תמונה 2D. המציג הציג 3 אפשרויות לרנדור תלת מימדי באופן כללי: וקסלים, ענן נקודות mesh, ולדבריו להשייע יש יתרונות על שתי השיטות האחרות.

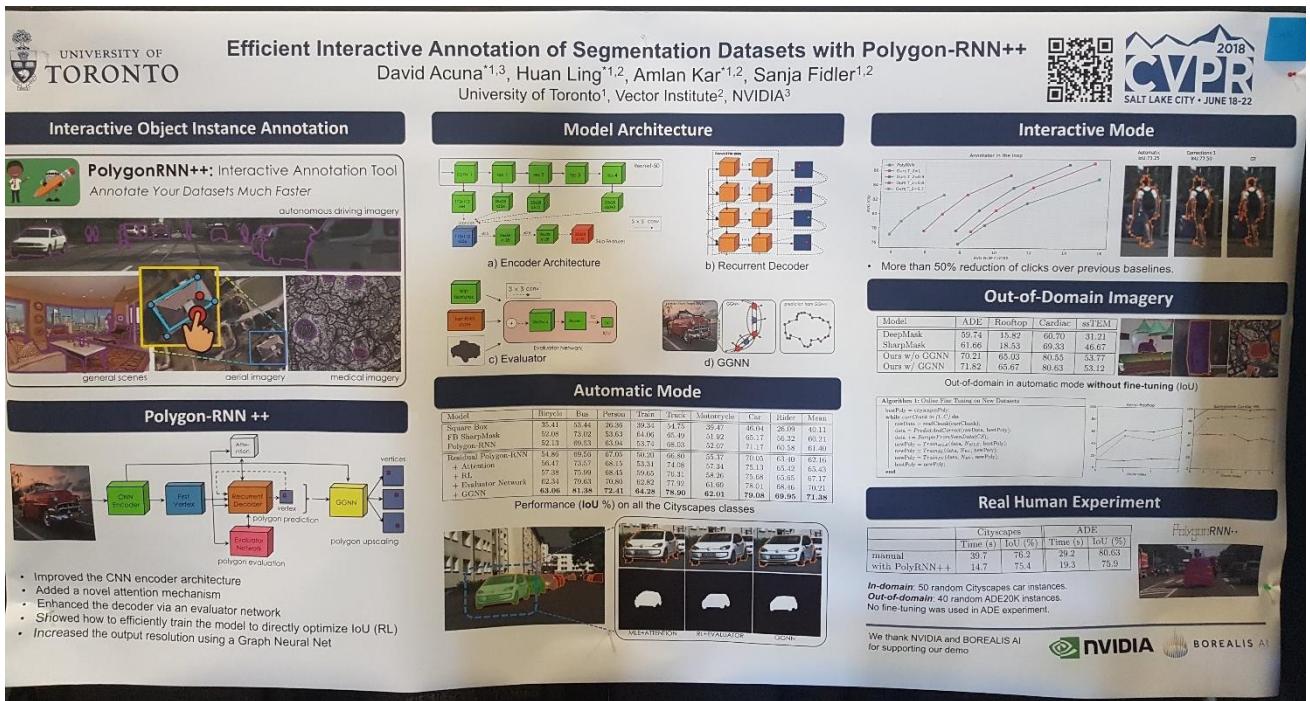
9. [F22] Beyond Gröbner Bases: Basis Selection for Minimal Solvers

בראייה ממוחשבת צריך לפתרו משוואות פולינומיות. שיטות גנריות בדרך כלל אורכות זמן רב וכן יש פתרונות פשוטים לכל בעיה minimal solvers. הזמן חשוב כיון שMRIIZIM בתוך לולאת RANSAC עם איטרציות רבות. דרך לקבלת solution minimal היא באמצעות פירוק לערכים עצמיים (?) ודורש בחירות בסיס. יש בסיס מקובל – Grobner basis. בעבודה זו מציעים בסיס אחר. לא הבנתי הכל.

1010-1230 Demos (Hall C)

Efficient Annotation of Segmentation Datasets With Polygon-RNN++

חישוב פוליגון חום באופן אוטומטי על ידי רשת RNN.



1010-1230 Poster Session P2-1 (Halls C-E)

From Orals

10. [C7] DenseASPP for Semantic Segmentation in Street Scenes, Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, Kuiyuan Yang

אגמנטציה באמצעות מסנן חדש עם פירמידות כיווצים (אני מבין שהמסנן אינו נלמד).

Object Recognition & Scene Understanding

3. [G12] Finding Beans in Burgers: Deep Semantic-Visual Embedding With Localization

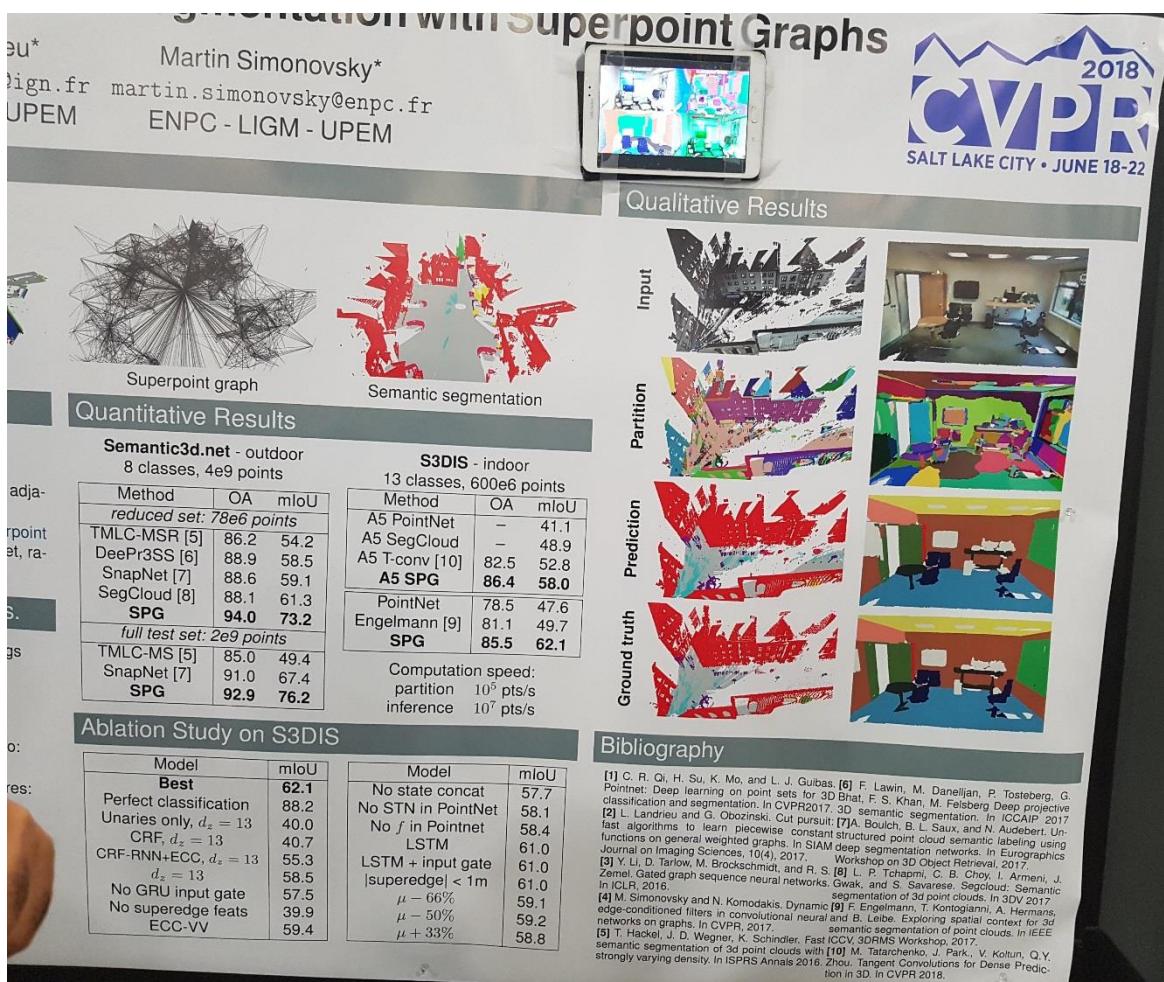
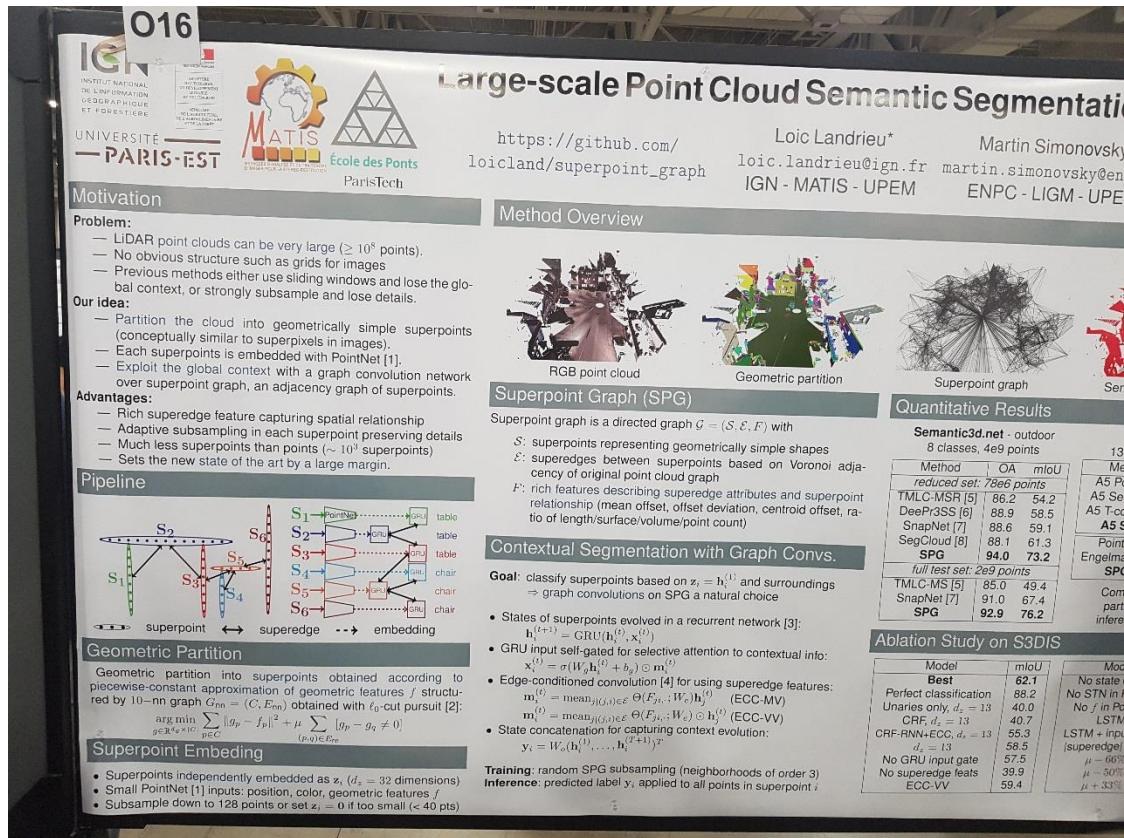
רשת המקבלת תמונה וטסקט המתאר את העצם המבוקש ומחזירה מפה חום המסמנת את המיקום המשוער של העצם.

3D Vision

63. [O16] Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs

אגמנטציה סמנטית של ענן נקודות גדול תוך שימוש ביצוג יעיל של הנקודות בענן והקשר ביניהן.

המטרה היא לבצע אגמנטציה סמנטית על ענן נקודות המכיל מאות מיליוני נקודות. הרעיון הוא לעבור לייצוג דלייל משמעותית, על ידי איחוד של נקודות הדומות מבחינה גיאומטרית למספר נקודות super points. האיחוד מתבצע על ידי שיטה קלאסית, נקרא *geometric partition*. לאחר מכן יוצרים גרף המחבר את כל הsuper נקודות שיש ביניהן קו ראייה (סוג של טריינגולציה ?), ומכו尼斯ים את הגרף לרשת RNN הלומדת קשרים סמנטיים בין הנקודות וביצעת אגמנטציה סמנטית. זמן ריצה של 2-3 שניות למיפוי חדר.



65.[Q22] ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans

שיטת להשלמת מידע חסר וביצוע סגמנטציה סמנטיבית עבור סריקות תלת ממדיות.

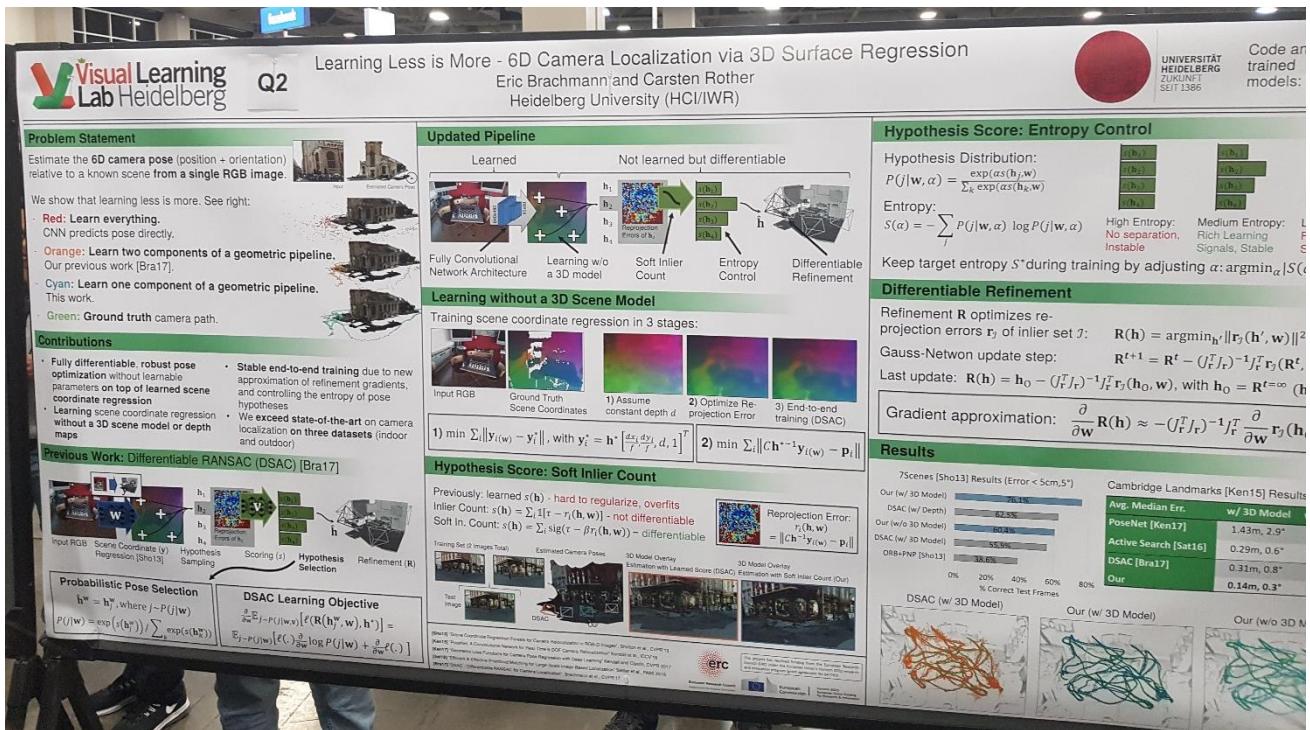
73. [Q2] Learning Less Is More - 6D Camera Localization via 3D Surface Regression

שחזור 6DOF של מצלמה מתוך תמונת RGB בודדת, באמצעות רשת הולומד מיפוי בין המידע בתלת ממד שבסצנה. יש סרטון בפייסבוק – [youtube](#)

מדובר על מקרה בו יש מספר (2) תמונות של סצנה שמהם לומדים, ולאחר מכן מקבלים תמונה שלישית של אותה סצנה מנוקודת מבט שונה וצריך לשער את המצלמה שלה. התהילך משלב רשות נוירונים בתוך pipeline של שחזור מצלמה: רשת אחת מהלצת נקודות עניין keypoints, רשת שנייה מבצעת RANSAC (מדובר בגרסה גלמודת בשם DSAC שהיא עבודה קדמתה של המחבר). אחר כך מחשבים פתרון אפשרי באמצעות פוטרן PnP (משהו זהה – לא הבנתי עד הסוף). בסופו של דבר מקבלים את המצלמה.

סיכום:

- החליף מספר חלקים של pipeline שחזור מצלמה ברשתות עמוקות.
- פתרון ייעודי עבור כל סצנה – צריך למדו מחדש כל סצנה.
- גודל הסצנה בעבודה זו מוגבל למבנה יחיד די גדול – אי אפשר להכניס שכונה שלימה.

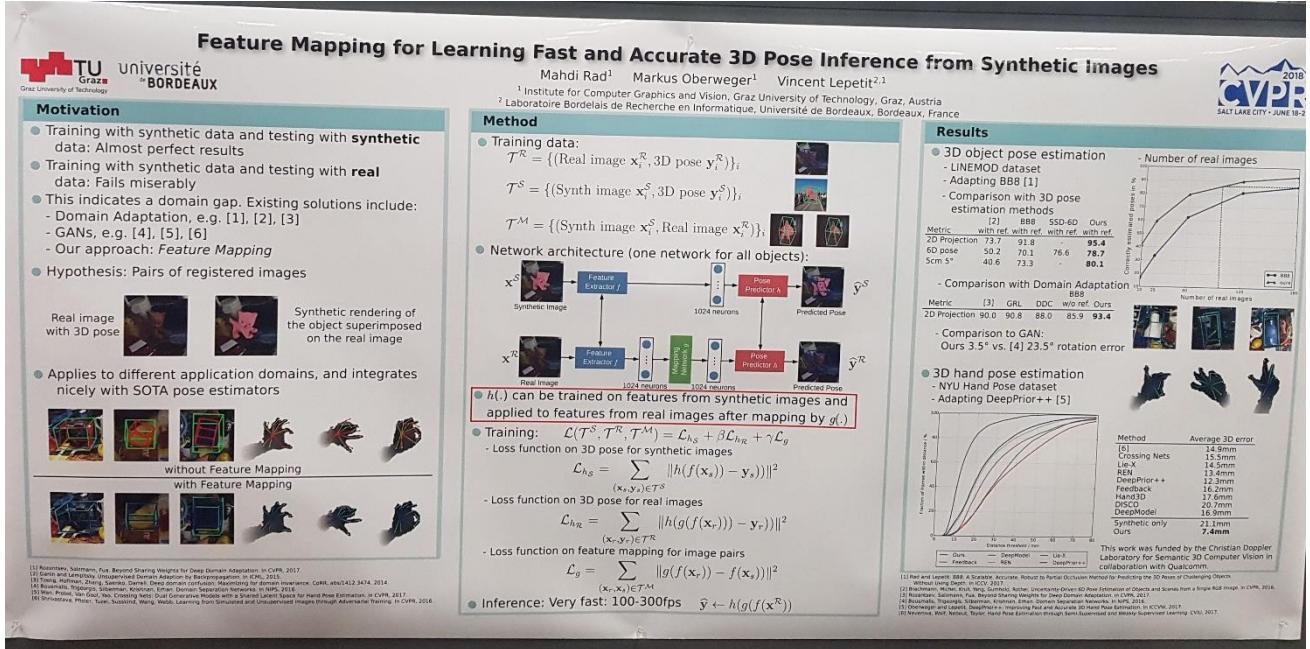


74. [Q5] Feature Mapping for Learning Fast and Accurate 3D Pose Inference From Synthetic Images

אימון באמצעות תמונות סינטטיות, וחיזוי על תמונות אמיתיות, תוך שימוש רשת להעברת מאפיינים מתמונות אמיתיות למאפיינים מתמונות סינטטיות. היישום במאמר הוא שערור קופסה חוסמת לעצמים, אבל הרעיון להעביר בין דומיין סינטטי לאמתי נראה לי חזק מאוד ויכול לעזור בהרבה יישומים.

סת האימון מורכב ממודל 3D של עצמים – מרנדרים אוטם ליצירת תמונות 2D ואחר כך מלבישים על רקע אקריאי (אם אני זוכר נכון הרקע נלקח מimagenet) – זה המידע הסינטטי. בנוסף צריך תמונה אמיתית של עצם, כמו למשל הקלט הוא זוג תמונות של אותו העצם, אחת סינטטית (מודל מרנדר על גבי רקע אקריאי) ואחת אמיתית. מכנים את התמונות לרשת לחילוץ מאפיינים (רשת שונה לכל תמונה). לאחר מכן מכנים את המאפיינים של התמונה האמיתית לרשת נוספת שמרתה להעביר אותן לדומיין של המאפיינים הסינטטיים על ידי פונקציית מחריך.

המזהurat מרחק אוקלידי. לאחר מכן המאפיינים מוכנסים לרשף המשערת מנה pose על ידי שערוך קופסה חוסמת. זמן ריצה מהיר מאוד 100-300 fps.

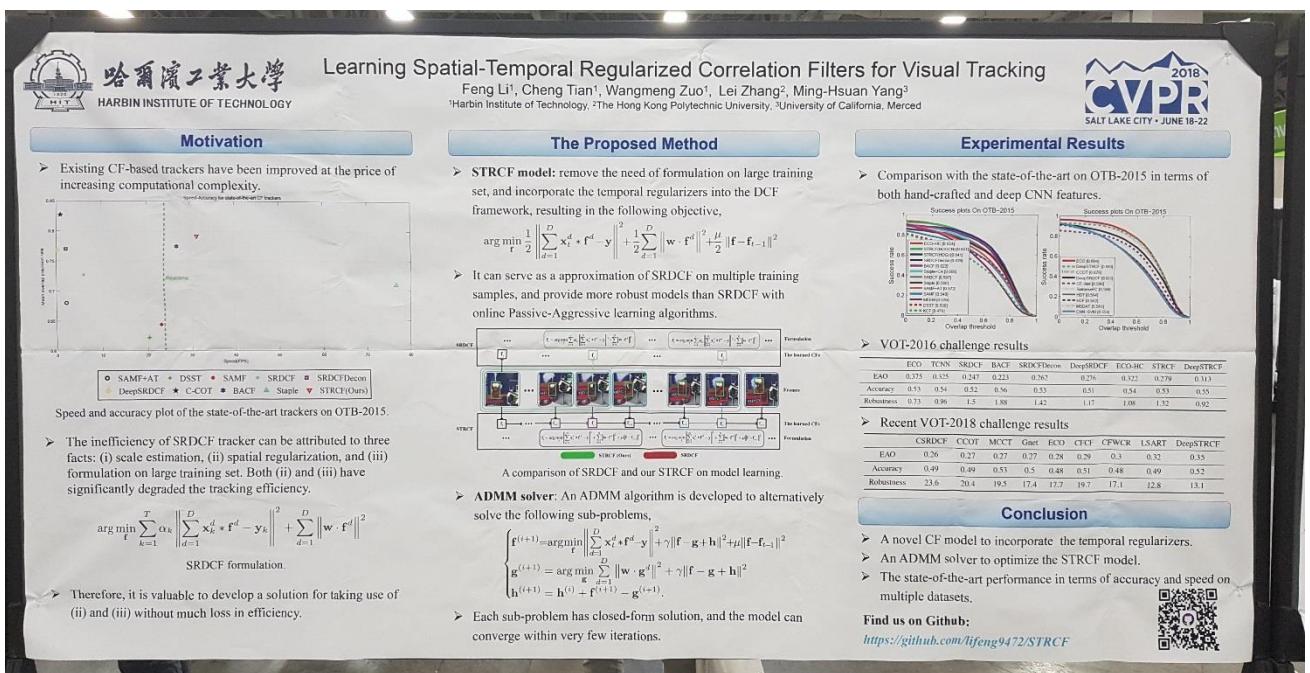


1230-1450 Poster Session P2-2 (Halls C-E)

Image Motion & Tracking

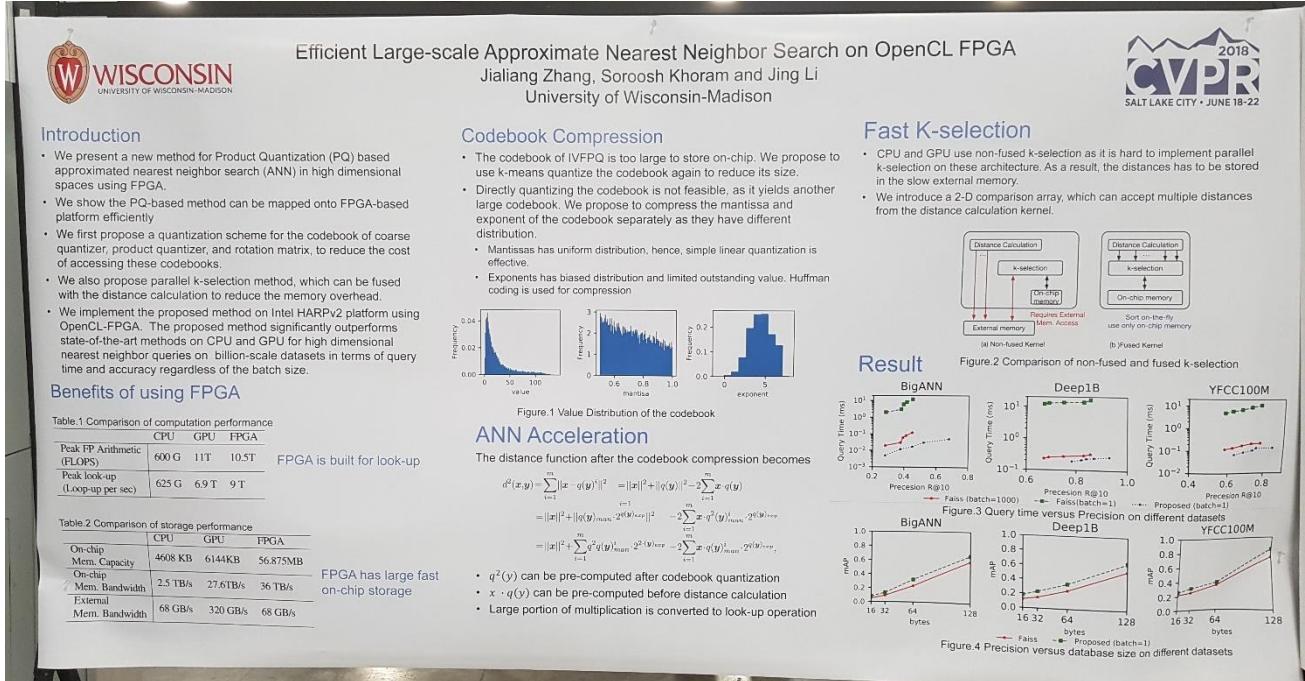
17. [C17] Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking

ואריציה מתקדמת של DCF, עם רגולריזציה מרחבית וזמןית.



Object Recognition & Scene Understanding

19. [D1] Efficient Large-Scale Approximate Nearest Neighbor Search on OpenCL FPGA



27. [E3] Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation

Machine Learning for Computer Vision

73. [K9] CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise

עבודה של פיסבוק ומיקרוסופט. למידת מסוכג באמצעות דוגמאות אימון עם תגיות מורעשות לשם שיפור יכולת הכללה והקטנת התלות במספר הדוגמאות המתויגות.

הרעיון הוא לאמן מסוכג עבור עשרות אלפי מחלקות, כאשר רק חלק קטן מהמחלקות יש מידע מתויג. תחילת מאנים רשות ללמידה יציג למחלקות שלמות: וקטור מאפיינים יחיד המיצג מחלקה שלימה. אחר כך מאנים מסוכג על ידי מזעור המרחק האוקלידי בין היזוג של הדוגמא הנוכחית לייזוג של המחלקה.

בנוסף, מעוניינים לבצע transfer learning בין המחלקות המתויגות למחלקות הלא מתויגות. לשם כך מעודדים את המסוכג לחקק את רמת הווודאות שלו במהלך חישוב כתף. רעיון דומה לעידוד הווודאות מופיע ב-202].

קשה לי עם הרעיון זהה של הגברת הווודאות בהמה שוחבים מראש כיון שלא נראה שיש לזה בסיס כלשהו למציאות – ומה אם הרשות ביצעה התכנסות ראשונית על משווה לא נכון וכעת נחקק את הווודאות של האימון הזה? עם זאת, המחברים צינו שמדובר ברעיון די מוקובל בתחום של unsupervised learning.

K9 Microsoft CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise

Motivation: Addressing **LABEL NOISE** is critical for learning image classifiers from images collected from noisy sources (e.g., Internet). Common approaches include:

- 1) Manually verify all the labeled images – **too much work**
- 2) Verify some labeled images and propagate the verification from human labelers for **every class – still not very scalable**
- 3) Methods without using any human supervision (e.g. one-class SVM) – **not very effective**

Key Idea

- 1) Address label noise with **transfer learning** to reduce need for human supervision, making image classifier learning scalable
- 2) Select representative class prototypes with **attention**

Task

Define verification label $l = \begin{cases} 1 & \text{if the image is relevant to its class label} \\ 0 & \text{if the image is mislabeled} \end{cases}$

Domain Adaptation: Learn from manual verification for some categories. Predict correctness of class labels for other categories.

Network Architecture

Query Encoder f_q : $v^q \rightarrow \phi^q \rightarrow r(v^q)$

Reference Set Encoder f_s : The "waffle" reference set. $h_1, h_2, h_3, \dots, h_n$. $h = \sum \alpha_k h_k$. $u_i = \tanh(W_h + b)$. $a_i = \exp(u_i^T u) / \sum_i u_i^T u$. Class Representation ϕ_c .

Application 2. Image Classification with Label Noise

$L_{weighted} = [\sum_i^n \max(0, \cos(\phi_q^i, \phi_c^i)) H(x_i, c_i)] / N$

Experimental Results

Image Classification

Food-101N Dataset

- 310k images we collected from web
- 60K v-labels for all 101 classes

Top-1 Classification Accuracy

number of held-out classes, n	Use CleanNet and exclude v-labels in n/101 classes	Train on noisy data (baseline)
18	0.6048	0.5839
26	0.7289	0.6960
34	0.7395	0.6935
42	0.7400	0.6935
50	0.7400	0.6935
58	0.7400	0.6935
66	0.7400	0.6935
74	0.7400	0.6935
82	0.7400	0.6935
90	0.7400	0.6935
98	0.7400	0.6935

Examples of Similarity Prediction

Cheese Plate (no images verified for training): 0.5048

Ramen (no images verified for training): 0.7289

Garlic Bread (no images verified for training): 0.6306

Label Noise Detection

On Food-101N dataset, use CleanNet and exclude v-labels in n/101 classes.

Averaged label noise detection error rate

Method	Verification	Validation Acc. Top-1 (Top-5) %
None	v-labels for all 2.4M images	67.76 (83.75)
all-1000	All 1000 classes with 250 v-labels per class	58.48 (79.76)
semantic-308	Randomly selected 308 classes with 250 v-labels per class that share a common 250-v-label hyperplane in WordNet with 250 v-labels per class	60.24 (81.21)
random-308	Randomly selected 308 classes with 250 v-labels per class	60.27 (81.27)
random-118	Randomly selected 118 classes with 250 v-labels per class	60.41 (81.03)
dogs-118	118 dog classes with 250 v-labels per class	59.45 (80.22)

Source code and Food-101N

[KuangHuei.github.io/CleanNetProject](https://github.com/KuangHuei/CleanNetProject)

76. [K18] Structured Uncertainty Prediction Networks

חיזוי מודל הסטברוטי (גאוסי) עם קוווארנס מלא, ולא אלכסוני כפי שצבע בעבר.

78. [L2] Adversarial Feature Augmentation for Unsupervised Domain Adaptation

GAN עם אוגמנטציה על הפיצרים המוחשיים על ידי הרשות על מנת לשפר יכולת הכללה.

84.[L20] Joint Optimization Framework for Learning With Noisy Labels

אימון רשות לשינוי תמונות תוך למידה דרך תיקון תוצאות שגויות במהלך האימון.

תחליה לומדים מסווג באופן רגילטור שימוש בcross entropy loss, ולאחר התוכניות הראשונית משנים את התగיות כך שכל תגית תהיה פילוג הסטברוטי ולא ערך ייחיד, כאשר הפילוג הוא softmax בموقع הרשות – ככלומר מעודדים את הרשות להחזק את הווודאות בהשהיא חושבת.

קשה לי עם הרעיון של עידוד הרשות לחזק הווודאות בחיזויים קודמים, ואני לא מאמין בגישה זו.

Joint Optimization Framework for Learning with Noisy Labels

DAIKI TANAKA DAIKI IKAMI TOSHIHIKO YAMASAKI KIYOHARU AIZAWA
THE UNIVERSITY OF TOKYO

PROBLEM

Deep neural networks trained on large-scale datasets have exhibited significant performance in image classification. Many large-scale datasets are collected from websites, however they tend to contain inaccurate labels that are termed as noisy labels. Training on such noisy labeled datasets causes performance degradation because DNNs easily overfit to noisy labels.

EXPERIMENT SETTING

We used the following datasets as noisy labeled datasets.

- Symmetric Noise CIFAR-10**: we reassigned $r\%$ labels in CIFAR-10 to random one-hot vectors.
- Asymmetric Noise CIFAR-10**: we reassigned $r\%$ labels in CIFAR-10 to class-dependent one-hot vectors as described in [4].
- Clothing1M**: 1M images of 14-class clothing obtained from online shopping websites. About 62% labels are correct.

OBSERVATION

Following [2], to examine the effect of the learning rate (lr) and the noise rate (r) on the training loss and the test accuracy, we trained the network on Symmetric Noise CIFAR-10 with the cross entropy loss.

Test accuracy curve with different learning rates. Test accuracy gradually decreases when the learning rate is low ($\text{lr}=0.02$). Conversely, test accuracy remains high at the end of training when the learning rate is high ($\text{lr}=0.2$).

Training loss curve with different noise rates. At the end of training with a low learning rate, the value of training loss is close to 0 even if the error rate is 0.9. In contrast, in the early phase of training with a high learning rate, an increase in the noise rate increases training loss.

JOINT OPTIMIZATION

In noisy supervised c -class classification problem setting, we have a set of n training images $X = \{x_1, \dots, x_n\}$ with noisy labels $Y = \{y_1, \dots, y_n\}$. Generally, the optimization problem is determining the network parameters θ to minimize the objective function \mathcal{L} . To address noisy labels, we formulate the problem as the joint optimization of network parameters θ and noisy labels Y as follows:

$$\min_{\theta} \mathcal{L}(\theta, Y, X) \rightarrow \min_{\theta, Y} \mathcal{L}(\theta, Y|X).$$

Our proposed loss function is constructed by three terms as follows: $\mathcal{L}(\theta, Y|X) = \mathcal{L}_c(\theta|X) + \alpha \mathcal{L}_p(\theta|X) + \beta \mathcal{L}_e(\theta|X)$, where $\mathcal{L}_c, \mathcal{L}_p, \mathcal{L}_e$ denotes the classification loss and two regularization losses, respectively, and α and β denote hyper parameters.

- Cross Entropy Loss**:
$$\mathcal{L}_c = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c s_{ij} \log s_j(\theta, x_i)$$

where s denotes the output of the c -class softmax layer.

- Distribution Matching Loss**:
$$\mathcal{L}_p = \sum_{j=1}^c p_j \log \frac{p_j}{s_j(\theta, X)}$$

where p denotes a prior distribution, which is a distribution of classes among all training data, and s denotes the mean probability.

- Entropy Minimizing Loss**:
$$\mathcal{L}_e = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c s_{ij} \log s_j(\theta, x_i)$$

In our proposed learning framework, network parameters θ and noisy labels Y are alternatively updated as shown in the following Algorithm.

```

Algorithm Alternating Optimization
for  $i \leftarrow 1$  to  $\text{num\_epochs}$  do
    update  $\theta^{(t+1)}$  by SGD on  $\mathcal{L}(\theta^{(t)}, Y^{(t)}|X)$ 
    update  $Y^{(t+1)}$  by  $Y_t = s(\theta, x_t)$ 
end for

```

RESULTS

Test accuracy of different baselines on Symmetric Noise CIFAR-10

method	test accuracy (%)
noise rate (%)	0 10 30 50 70 90
Cross Entropy Loss	best 93.5 91.0 88.4 85.0 78.4 41.1
	last 93.4 87.0 72.2 55.3 36.6 20.4
Our Method	best 93.4 92.7 91.4 89.6 85.9 58.0
	last 93.6 92.9 91.5 89.8 86.0 58.3

Test accuracy of different baselines on Asymmetric Noise CIFAR-10

method	test accuracy (%)
noise rate (%)	10 20 30 40 50
Cross Entropy Loss	best 91.8 90.8 90.0 87.1 77.3
	last 89.8 85.4 81.0 75.7 70.5
Forward [1]	best 92.4 91.4 91.1 90.3 83.8
	last 91.7 89.7 88.0 86.4 80.9
CNN-CRF [3]	best 92.0 91.5 90.7 89.5 84.0
	last 90.3 86.6 83.6 79.7 76.4
Our Method	best 93.2 92.7 92.4 91.5 84.6
	last 93.2 92.8 92.4 91.7 84.7

Test accuracy on Clothing1M

method	acc. (%)
Cross Entropy Loss	68.94
Forward [1]	69.84
Cross Entropy Loss (reproduced)	69.15
	last 66.76
Our Method	best 72.16
	last 72.23

(Upper) The images with the top-2 and the bottom-2 probabilities of T-shirt whose labels are reassigned from Hoodie to T-shirt (Lower) The images with the top-2 and the bottom-2 probabilities of Hoodie whose labels are reassigned from T-shirt to Hoodie.

REFERENCES

[1] G. Hinton et al., "Deep Neural Networks for Large-Scale Image Recognition", NIPS, 2012.

[2] A. Krizhevsky et al., "Understanding deep learning requires rethinking generalization", ICML, 2017.

[3] A. Vedaldi, "Toward Robustness: Label Noise in Training Deep Convolutional Neural Networks", NIPS, 2012.

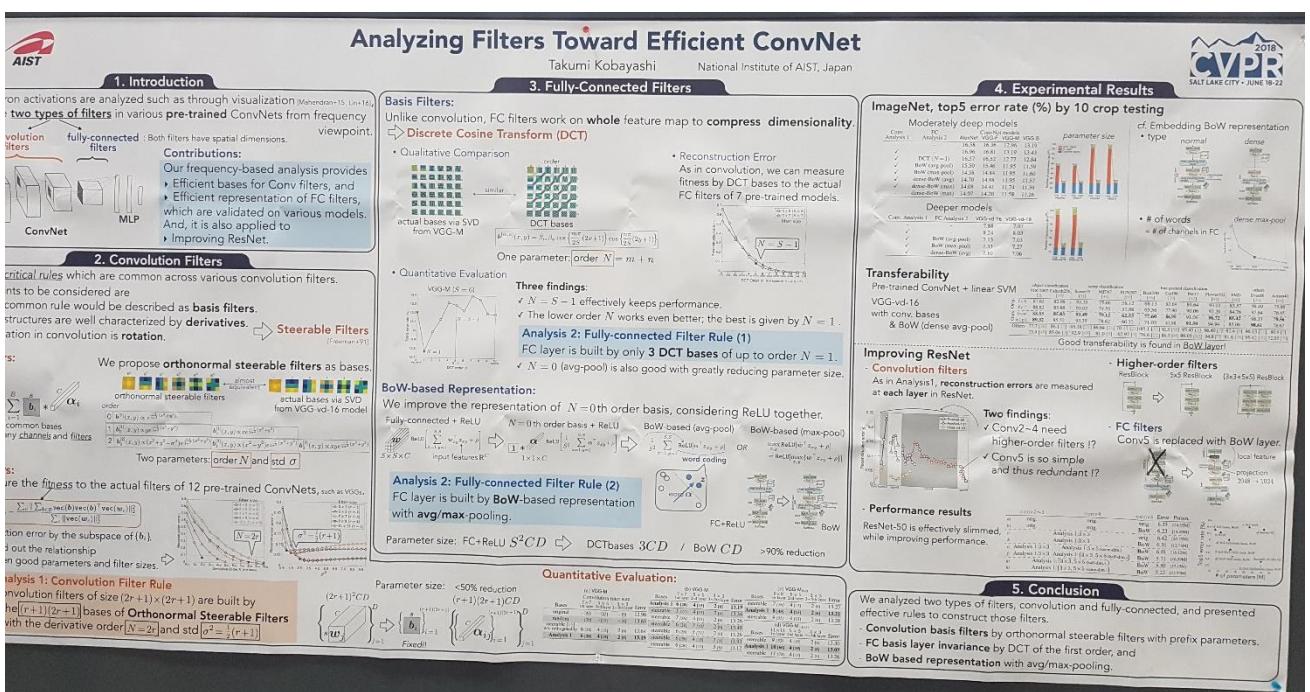
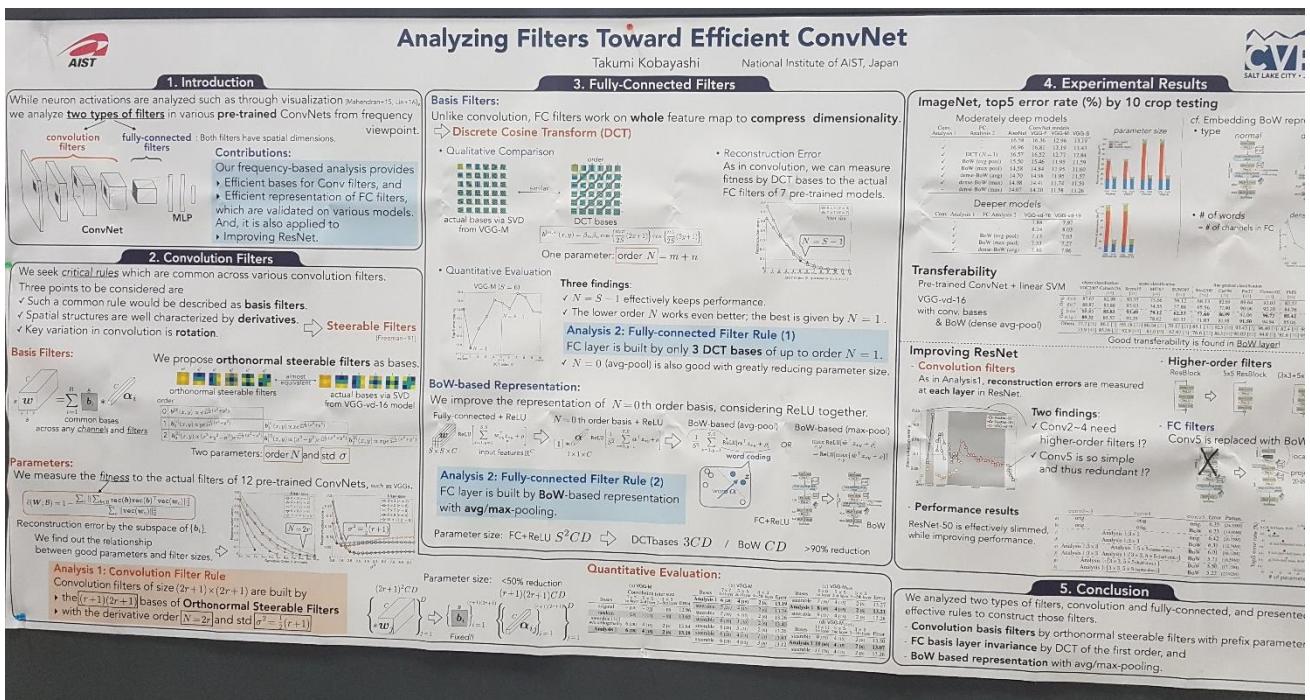
91. [M19] Analyzing Filters Toward Efficient ConvNet

ניתוח מקדמי הרשת (בשונה מניתוח אקטיבציות הרשת השכיח יותר) על מנת להשתמש ביצוג יעיל יותר.

מדובר בבחירה של בסיס אורטונורמלי לייצוג שכבות קונבולוציה (על ידי steerable filters) ושכבות FC (על ידי DCT), כך שבוחרים מספר קבוע של פונקציות הבסיס מראש ומיצגים באמצעות את שכבות הרשת.

הרעיון מזכיר לי מאוד פירוק פורייה, SVD, PCA ועוד – גם שם מדובר בפירוק לבסיס אורטונורמלי, בו ניתן לבחור איברים מובילים וכן לצמצם את הייצוג של הנתונים. ההבדל המשמעותי הוא שהשיטות אלו עובדות לאחר שיש לנו את כל הנתונים, וכעת אנו רוצים למצוא ייצוג מקרוב יעיל יותר, בעוד השיטה המוצעת עשויה זאת מראש: בוחרת ייצוג ייעיל אחר וכך מקרבבת באמצעותו את כל השכבות.

רעיון נחמד, מימוש מטלב יועלה לאינטרנט בעtid.



93. [N3] In-Place Activated BatchNorm for Memory-Optimized Training of DNNs

שכבה חדשה המבוצעת BN + אקטיבציה, ובכך מפחיתה ממשמעותית את חתימת הזיכרון באימון, בתמורה לעלייה קטינה בזמן החשבוב.

Object Recognition & Scene Understanding

100.[O2] Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking

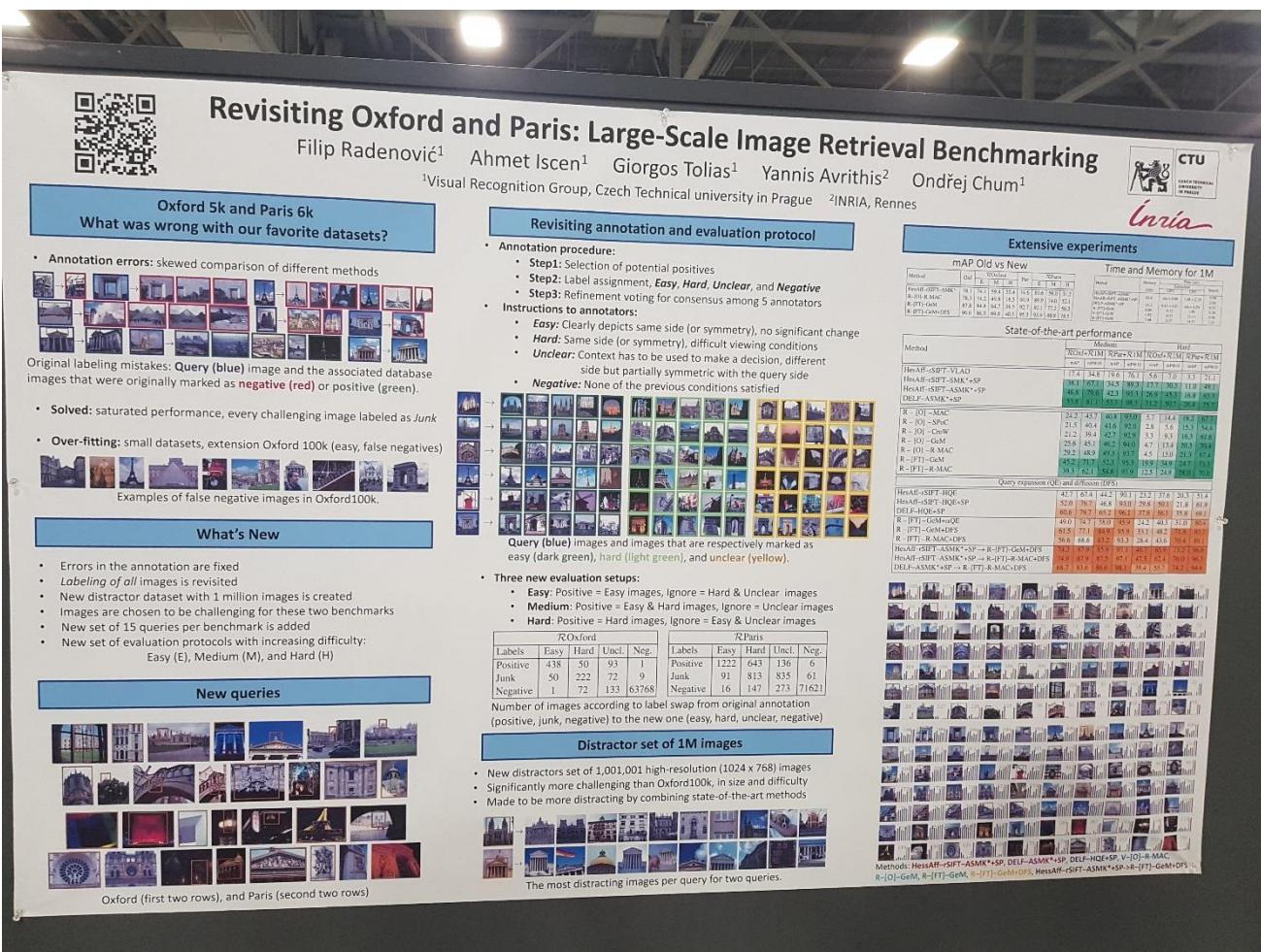
עדכן תוצאות שגויים והגדרת פרוטוקולים אחידים לבדיקה על 2 בסיסי הנתונים הנ"ל. סקירה של תוצאות עדכניות.

רלוונטי מאד למקרים בהם אין ייחוש התחלתתי. חלק מהמחברים היו שותפים ב-HardNet.

העבודה זו עוסקת בשיפור בסיס הנתונים Oxford ו Paris המשמשים לבחינת מושגות של אחזור תמונות מתוך בסיס נתונים גדול. לדברי המחברים עבדות עדכניות הגינו לביצועים טובים מאוד על בסיס הנתונים הללו (90% ויותר), ולכן יש הרגשה שהם למעשה פטורים ולא מתגררים מספיק. לאחר השיפור המוצע התווסף רמות קושי חדשות בהן יש מקום רב לשיפור. בין היתר בוצעו הדברים הבאים:

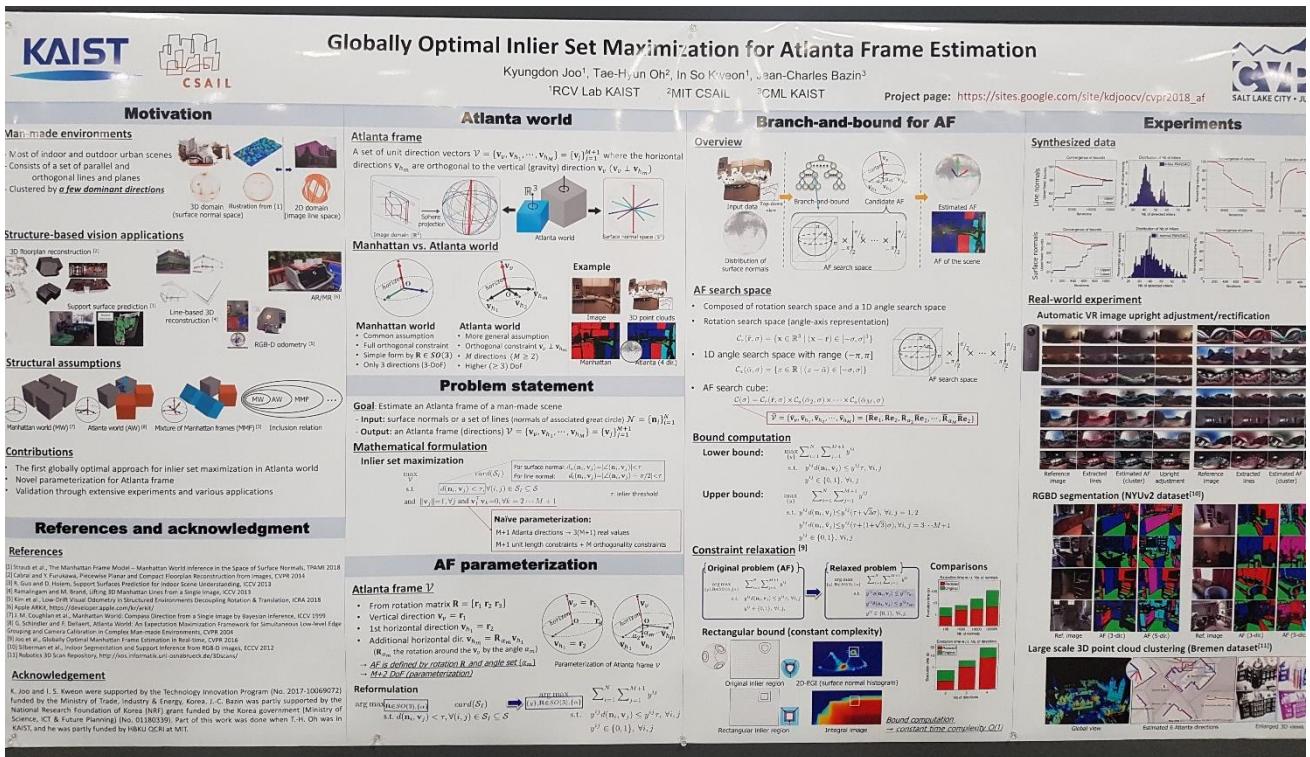
- תיקון תיוגים שגוים.
- חלקה לשולש רמות קושי.
- הוספה M1 דוגמאות שליליות קשות.

בנוסף בוצעה השוואת מספר שיטות קלאסיות וمبוססות למידה על בסיס הנתונים חדש. באופן מפתיע נראה שלמרות ששיטות מובוססות למידה טובות יותר עברו הרמה הקלה, שיטות קלאסיות מציאותיות יותר ברמה הקשה. לא הכרתי את השיטות מובוססות הלמידה שהוא הזכיר, ולאחר ששאלתי הסביר שמדובר בשיטות המציגות תמונה שלימה באמצעות דסקריפטור בלבד, ואילו בשיטות הקלאסיות עשו שימוש בدسקריפטורים לוקלים. חשוב לציין שלא בדקנו דסקריפטורים לוקלים מובוסטי' למידה דוגמת HardNet. כששאלתי בנושא שהוא בטוח שישמש HardNet מוביל לתוצאות טובות יותר מאשר הדסקריפטורים הקלאסיים, כיוון שהוא אמפירית שעומדת על מגוון משימות במאמר המקורי (בינהן גם אחזור תמונות על בסיס הנתונים הללו). כיוון שהמציג הוא אחד ממחברי HardNet אני נוטן משקל מוגבל לדבריו.



102.[08] Globally Optimal Inlier Set Maximization for Atlanta Frame Estimation

היא שיטה לייצוג מבנים באמצעות גובה וקווים אונכים. בעבודה זו מוצגת שיטה לשערוך AF עם המבניתה למצוא אופטימום גלובלי. נחמד אבל לא רלוונטי.



Applications

115.[Q3] Fast Monte-Carlo Localization on Aerial Vehicles Using Approximate Continuous Belief Representations

שיטת מהירה לשערוך מצלמה (מתחרה SLAM) בסביבה עם ענן נקודות צפוף, חישון عمוק, וחוויות גסות.

לא למדידה. המטריה היא לשערוך מיקום עצמי של רחפן תמונות של חישון عمוק זול. הדוגמאות שהוצעו הינו עבר סביבת *indoor*, לא יודע כיצד יעבד עבור סביבת *outdoor*.

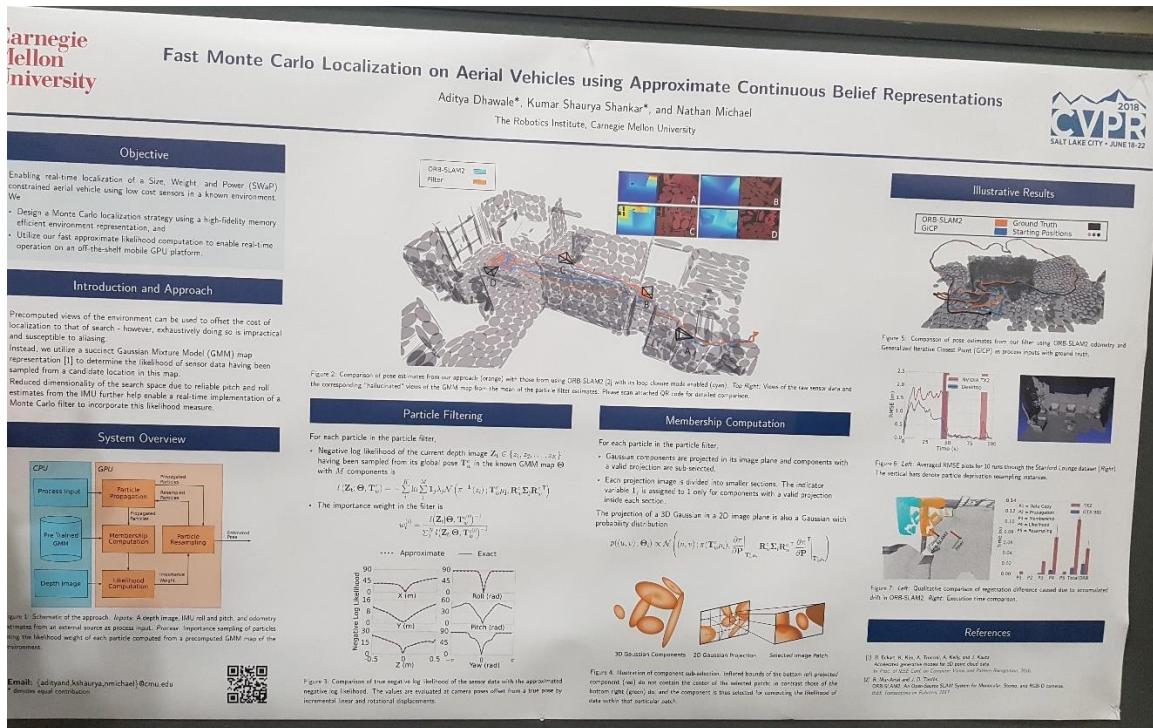
נתון ענן נקודות המתאר אזור מסוים. תחיליה מודלים את ענן הנקודות באמצעות GMM על מנת להפחית משמעותית את רמת הסיבוכיות (בעבודה זו השתמשו ב-2024 אוסיאנים). במהלך הריצה מפעלים מסנן חלקיקים על מנת לשערר את המיקום בכל רגע. המציגים הרואו שימוש ב-GMM עם 2024 רכיבים מספיק טוב בשבייל להגעה לרמת הדיק שרצאו במקורה שבדקו (במקרים אחרים צריך לבחון שוב).

המערכת רצה על רחפן בזמן אמיתי, עושים שימוש בשביב Tegra של Nvidia בשם Jetson (?). משתמשים ב-GPU כדי שיכול לחשב במקביל פתרונות עבור מסנן חלקיקים (ולא אഗל שיש רשת עמוקה או משה זהה).

שאלתי לגבי סוג חישון העומק, אמרו שבדקו הרבה רכיבים, גם *redSense* (?), לא זוכר במה השתמשו בסופו.

חישוב GMM בוצע באמצעות פונקציה של *scipy*.

עושים שימוש גם ב-SLAM על מנת לשערר זווית Roll Pitch (אמור שאפשר לסרוך על ה-SLAM לשערר זווית אלו), אך שמסנן החלקרים צריך לשערר פחות נעלמים.

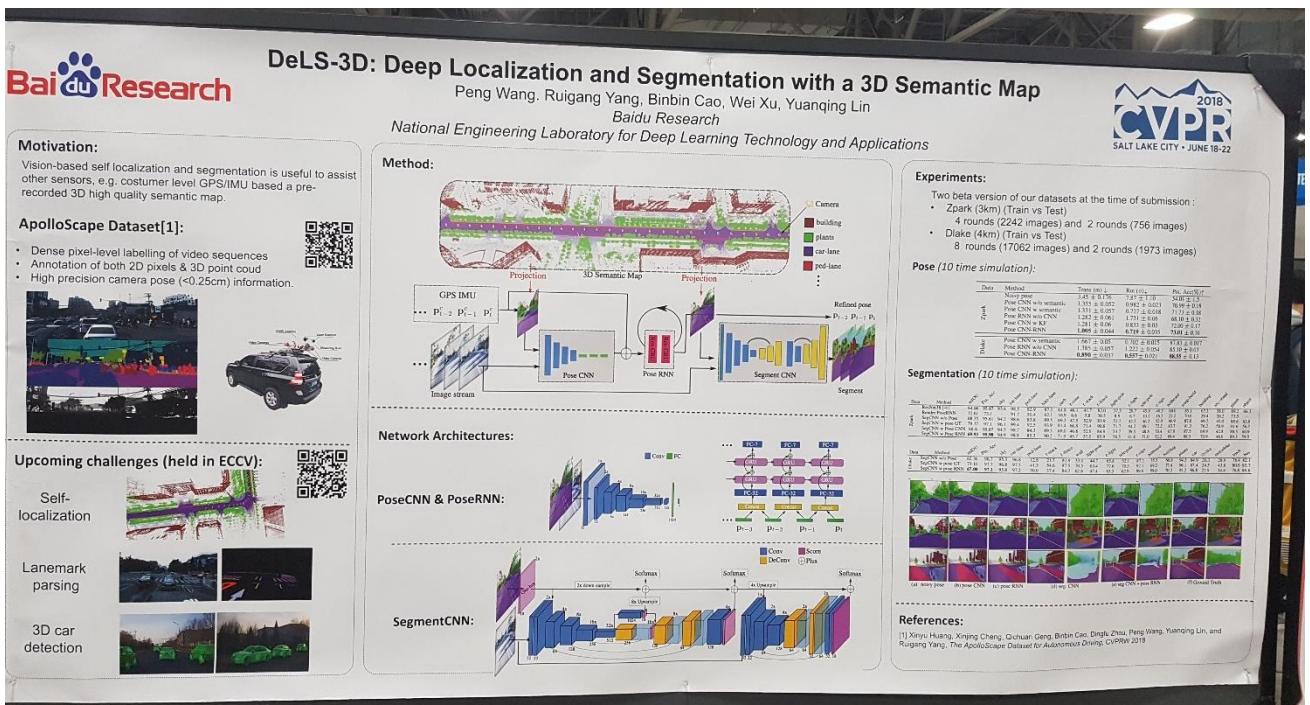


116.[Q6] DeLS-3D: Deep Localization and Segmentation With a 3D Semantic Map

שערוק מצלמה + ענן נקודות 3D סמנטיקי תור היתוך מידע מתונות, GPS, IMU

קבלת מיקום מצלמה (6DOF) ראשוני GPS/IMU. חישוב מפה סמנטיקת 3D ראשונית מתמונה ומיקום מצלמה. טויב מיקום מצלמה באמצעות רשת عمוקה ומפה סמנטיקת. התוצאות במדוע זמן באמצעות RNN. טויב מפה סמנטיקת 3D באמצעות מיקום מצלמה מותוקן.

בשmu מעניין ורלוונטי מאוד – לבדק! לא היה אף אחד בעמדה.



1450-1630 Session 2-2C: Computational Photography (Room 255)

10. [G2] Learning to Detect Features in Texture Images

למידת גלי מאפיינים בתמונות עם טקסטורה דומה (שחוורת על עצמה) ללא מאפיינים בולטים. פונקציית המחריף מורכבת משני איברים: אחד המעודד דירוג טוב, ושני המעודד שאים מוחנים (מצחיר את הרעיון של שילוב מפות דמיון בפונקציית המחריף).

גלי מאפיינים קיימים בדרך כלל מתוכנים עבור תמונות טבעיות, ולא עבור תמונות עם חזירות רבות (טקסטורה), ובעובדה זו מציגים גלי מאפיינים המתאים לתמונות עם טקסטורה. נעשה שימוש ברשות עמוקה על מנת ליצור מפת חום, כאשר השאים (גם מקסימיות וגם מינימיות – כל אחד בטקסטורות שונות) במפת החום הם המאפיינים המגולמים. עם זאת, השימוש בגלי זה בלבד לא מספיק טוב – לא הספקתי לכתוב. בעובדה זו הוצג שיפור בצורה של UIDOD השאים במפת החום להיות מוחנים \ חדים. שימוש בשיטה זו הוביל לשיפור במספר נקודות העניין keypoints המגולמות. בנוסף, התוצאות של המציג מראות שאימון עבור טקסטורות ספציפיות מוביל לשיפור בגיןUIDOD על טקסטורות.

1630-1830 Demos (Hall C)

Efficient Annotation of Segmentation Datasets With Polygon-RNN++

חישוב מלבן חום למטרת תיאוג סמנטי – יכול לעזור לתיאוג \ לפענוח תמונות במהירות.

Semi-Dense, Event-Based Visual SLAM

Ultimate SLAM? Combining Events, Frames and IMU for Robust Visual SLAM

שילוב מידע מצלמה רגילה + מצלמה מבוססת מאירועים (סוג מיוחד של מצלמה הפועל כאשר יש שינויים בתמונה) יחד עם מדדים אינרציאליים.

Real-Time Visual SLAM Using a Jointly Optimized, Compact Dense Code

1630-1830 Poster Session P2-3 (Halls C-E)

Low-level & Mid-level Vision

35. [L6] Latent RANSAC

בחירת outliers בצורה מהירה וטובה – נשמע מעניין!

שיטת להאצת RANSAC בפקטור של 4-2 עברו מספר משלימות (לא גנרי אלא צריך להתאים לכל משימה). הרעיון הוא לדגום פתרונות אפשריים minimal solutions, אך לא לבדוק אותם אלא לאgor אוותם, וליצג אותם על ידי ייצוג פשוט (Hashing) ולשמור בסוג של טבלה (hash table). ממשיכים ליצור פתרונות ולהceneיס אותם לטבלה עד שמקבלים 2 פתרונות באותו תא בטבלה – ובמקרה כזה בודקים את הפתרון. מה שיצוא זה שמייצרים הרבה פתרונות (יותר מאשר RANSAC רגיל), אך בודקים רק חלק קטן מהם (הרבה פחות מאשר RANSAC).

מבחינה אינטואיטיבית: יובילו לפתרונות דומים יחסית (תאים קרובים בטבלה) בעוד outliers outlies לא, ולכן כדאי לחכות לשתי פתרונות לפחות.

בעובדה ההז ציריך למצוא יציג יידייע (hashing function) לכל בעיה ולכן מספר הביעות בהן ניתן להשתמש בשיטה זו מוגבל. עם זאת, למידת יציג גראית כמו משימה טובה מאוד עבור למידה ולכן נראה שאפשר להכליל את השיטה באופן פשוט יחסית.

היתרון של השיטה ההז על פני שיטות אחרות מرتبط בעיות בהן זמן בדיקת הפתרון הוא משמעותי (דוגמא שחזור מצלמה CHP), ופחות בעיות בהן זמן ייצור הפתרון הוא משמעותי (דוגמא חישוב הומוגרפיה).

39. [L18] Graph-Cut RANSAC

שיטת נוספת לבחירת **Inliers**.

Object Recognition & Scene Understanding

51. [N10] ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices

רשת ישרה עבור מכשירים דל' כוח חישוב, לפי הפרסום טובה יותר מ-**MobileNet**.

56.[O3] Semantic Visual Localization

למידת דסקריפטורים לשערר מקום מצלמה יחד עם חילוץ ושימוש במידע סמנטי – משמעות!

בלונטי מואוד למקרים בהם אין ניחוש התחלתי, יש סיכון גם בקורס של היום האחרון (קישור לסייעם: - pm 3:50
4:30 pm, Keynote Talk: Andreas Geiger (MPI & University of Tübingen), Talk topic: Semantic Visual Localization).

המשימה היא לשערר מקום מצלמה בתוך שטח של מספר קמ"ר. עבור השטח יש מודל 3D עם סגמנטציה סמנטית אשר התקבל באמצעות שימוש בכלים קיימים. בעת הרצאה הקלט הוא תמונה RGB ותמונה עמוקה.

התוצאות מרשימות מואוד, עם דיקוק של מטרים בודדים (5m) עם אחוז הצלחה גבוהה מאוד – 60-90%, עבור תרחישים עם הבדלים משמעותיים: שינוי כיוון הסתכלות ב-90 מעלות, שינוי קיצוני במראה: יום \ לילה, קיץ \ חורף.

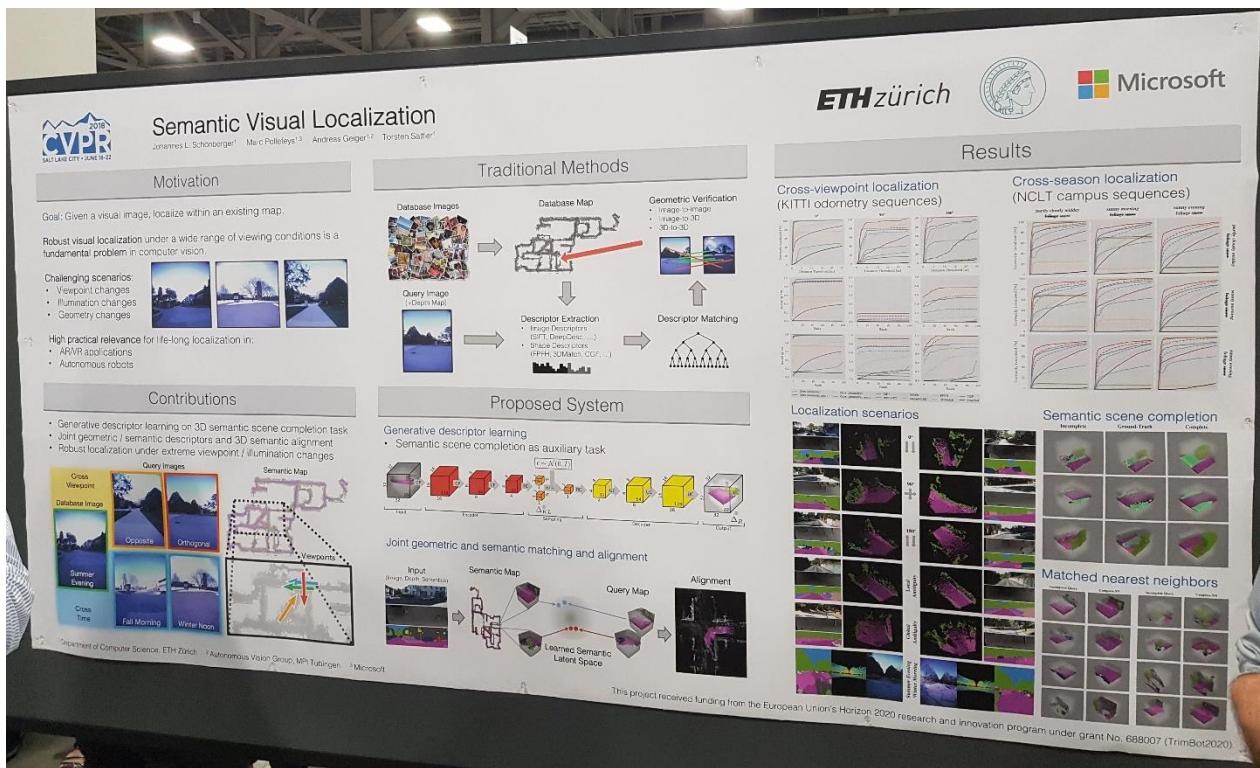
הרעilon המרכדי הוא למידה של דסקריפטור המקודד מידע גיאומטרי + סמנטי. לימוד הדסקריפטור נעשה בצורה מבריקה לדעת: עושים שימוש בAuto Encoder אשר מקודד מידע של נפח חלקו בעולם. במהלך האימון לוקחים תא נפח של 10^3 מ' ומתראים אותו על ידי 32^3 ווקסלים סמנטיים. המראה של תא הנפח תלוי בנקודות המבט, שכן יש הסתרות התלוויות במיקום המצלמה. במהלך האימון לוקחים תא נפח שצולם מנקודות מבט מסוימת ולבן מכיל מידע חלקו על איזור בעולם, ומעבירים דרך Decoder Auto Encoder עם ג'רטיבי שמטרתו לשחרר את כל המידע של תא הנפח, גם זה שלא מופיע בקלט.

לדברי המחבר מפת העומק יכולה להיות גסה יחסית. בעבודה זו התאמנו על kitti ובדקו על מישיגן, כאשר באחד מבסיסי הנתונים נתוני העומק התקבלו על ידי **lidar**, ובשני על ידי מצלמת סטראיאו. לדברי המציג ניתן לחשב עומק גם בדרךים נוספות, לדוגמה על ידי תנועה.

פרטים טכניים:

על מנת ליצור מפה סמנטית 3D צריך: 1. תמונה RGB, 2. תונות עומק (יכולות להיות גסות ולהתקבל מלידר, סטראיאו תנועה ועוד). את התמונות מעבירים דרך כלים מוכנים לsegmantציה סמנטית ולאחר מכן מייצרים מפת סמנטית 3D – המציג אמר שהשתמש ב **Octomap** + כלים לסמנטיקה, צריך לבדוק במאמר או לשלו מיל.

לפנוי תחילת האימון צריך ליצור מפה סמנטית 3D של כל השטח (מספר קמ"ר) – בוצע על ידי ספריית קוד קיימת. לאחר מכן על מנת ללמידה דסקריפטור בוחרים נפח בגודל 32^3 ווקסלים (המכילים מידע סמנטי – מקביל ל 10^3 מ'). ומסתכלים עליו מנקודות מבט מסוימת כך שחרר בו מידע. לאחר מכן מכינים **Auto Encoder** והמטרה היא לשחרר את כל המידע בנפח.



21.06.2018

0830-1010 Session 3-1A: Object Recognition & Scene Understanding IV (Ballroom)

11. [B16] CVM-Net: Cross-View Matching Network for ImageBased Ground-to-Aerial Geo-Localization

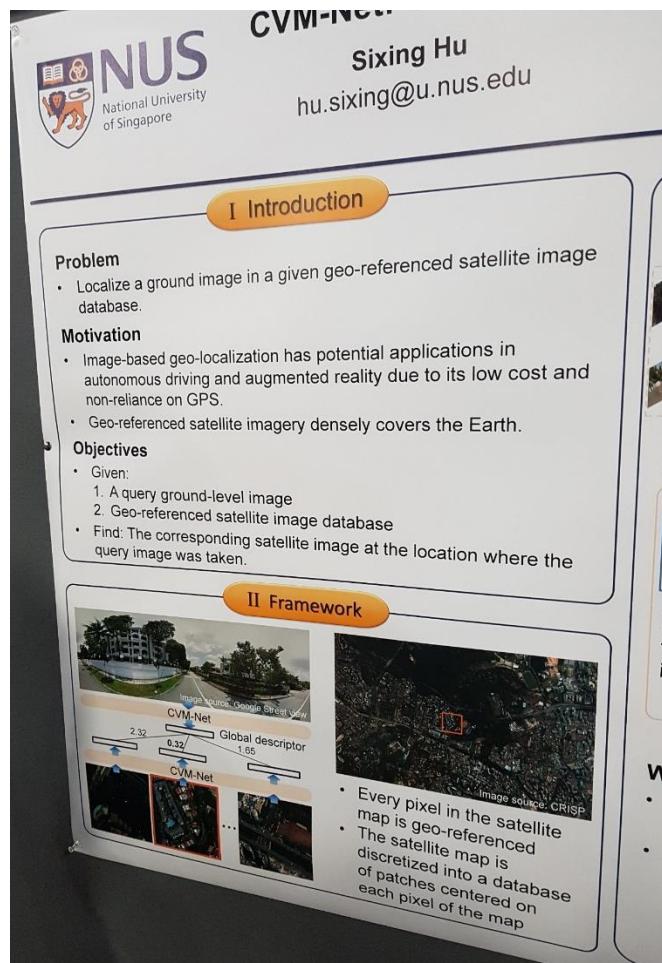
מציאת מיקום גיאוגרפי בתמונות קרקעיות תוך שימוש בתשתיית של תמונות לוויין. בלונטי למקרים בהם אין ניחוש התחלה.

בහינתן תמונות לוויין מצומדות למערכת קואורדינטות (בגודל Km 5x10 5x10 ברזולוציה של $0.5 \text{ m}/\text{pix}$) ותמונה שאילתת מנקודות מבט קרקעית (בגודל 512x512. שתי אפשרויות: 1. תמונה פנורמה $\sim 360^\circ$ מתוך google street, 2. חיתוך של תמונת הפנורמה הנ"ל), המטרה היא למקם את המצלמה של תמונת הפנורמה בתמונה הלווין.

האימון מתבצע באופן הבא: גוזרים את כל החלונות האפשריים (בגודל 512x512) מהתמונה הלווין ומכניםים אותם לרשת VGG לשם חילוץ מאפיינים. באופן דומה מחלצים מאפיינים גם מהתמונה השאילתת לרשת, כאשר כן עושים שימוש ברשת VGG אחרת (המקדים של 2 רשותות VGG אינם משותפים). רשת VGG מחלצת מספר מאפיינים לכל תמונה. בשלב הבא מייצרים דסקריפטור גלובלי ייחיד לכל תמונה המורכב מכל המאפיינים שחולצו על ידי VGG באמצעות שיטה בשם NetVLAD (גרסה נלמדת של VLAD אשר בעצמו דומה לWord Bag). את הדסקריפטור הגלובלי מכנים לפונקציה מחיר מסווג soft margin triplet loss (וואריאציה של triplet loss) אשר הבדל הוא שבגרסת הסנטדרית מגדרים את השולים margin באופן מפורש ואילו כאן השולים נלמדים באופן עקיף על ידי הרשת).

לאחר התוכנות הראשונית מבצעים כרייה של דוגמאות קשות בדומה לHardNet.

מבחינות תוצאות – נראה שיש שיפור משמעותי ביחס לשיטות קודמות, אם כי לא בטוח שהביצועים מספיק טובים לשם שימוש ביישומים אמיתיים.



-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization

Mengdan Feng fengmengdan@u.nus.edu Rang M. H. Nguyen nguyenho@comp.nus.edu.sg
National University of Singapore

III CVM-Net

Network architecture

- CVM-Net pipeline:

CVM-Net-I **CVM-Net-II**

Components:

Extract local features of images
VGG16 NetVLAD

Weighted soft-margin triplet loss

- Weighted triplet:

$$\mathcal{L}_{triplet,weighed} = \ln(1 + e^{\alpha(d_{pos} - d_{neg})})$$
- Weighted quadruplet:

$$\mathcal{L}_{quad,weighed} = \ln\left(1 + e^{\alpha(d_{pos} - d_{neg})}\right) + \ln\left(1 + e^{\alpha(d_{pos} - d_{neg}^*)}\right)$$

α is the weight
 d_{pos} is the distance of positive sample and anchor
 d_{neg} is the distance of negative sample and anchor, d_{neg}^* is another negative pair

IV Results

Dataset

- Crop: Vo and Hays' dataset [1]
- Pano: CVUSA dataset (Zhai et al.) [2]

Our results and Comparison

	Crop	Pano
CVM-Net-I	67.9%	91.4%
CVM-Net-II	66.6%	87.2%
Our variation		
CVM-Net-I (AlexNet)	65.4%	83.7%
CVM-Net-II (VGG16)	91.4%	89.9%
CVM-Net-II (AlexNet)	63.0%	83.9%
CVM-Net-II (VGG16)	87.2%	88.7%

Retrieval results

Localization results

V References

[1] N. N. Vo and J. Hays. Localizing and orienting street views using overhead imagery. ECCV, 2016.

-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization

Mengdan Feng fengmengdan@u.nus.edu Rang M. H. Nguyen nguyenho@comp.nus.edu.sg Gim Hee Lee gimhee.lee@comp.nus.edu.sg
National University of Singapore

CVPR 2018
SALT LAKE CITY • JUNE 18-22

III CVM-Net

Network architecture

- CVM-Net pipeline:

CVM-Net-I **CVM-Net-II**

Components:

Extract local features of images
VGG16 NetVLAD

Weighted soft-margin triplet loss

- Weighted triplet:

$$\mathcal{L}_{triplet,weighed} = \ln(1 + e^{\alpha(d_{pos} - d_{neg})})$$
- Weighted quadruplet:

$$\mathcal{L}_{quad,weighed} = \ln\left(1 + e^{\alpha(d_{pos} - d_{neg})}\right) + \ln\left(1 + e^{\alpha(d_{pos} - d_{neg}^*)}\right)$$

α is the weight
 d_{pos} is the distance of positive sample and anchor
 d_{neg} is the distance of negative sample and anchor, d_{neg}^* is another negative pair

IV Results

Dataset

- Crop: Vo and Hays' dataset [1]
- Pano: CVUSA dataset (Zhai et al.) [2]

Our results and Comparison

	Crop	Pano
CVM-Net-I	67.9%	91.4%
CVM-Net-II	66.6%	87.2%
Our variation		
CVM-Net-I (AlexNet)	65.4%	83.7%
CVM-Net-II (VGG16)	91.4%	89.9%
CVM-Net-II (AlexNet)	63.0%	83.9%
CVM-Net-II (VGG16)	87.2%	88.7%

Retrieval results

Localization results

V References

[1] N. N. Vo and J. Hays. Localizing and orienting street views using overhead imagery. ECCV, 2016.

0830-1010 Session 3-1B: Analyzing Humans

1. [C6] Consensus Maximization for Semantic Region Correspondences

שיטת להתאמה סמנטית בין שני מודלים תלת ממדים.

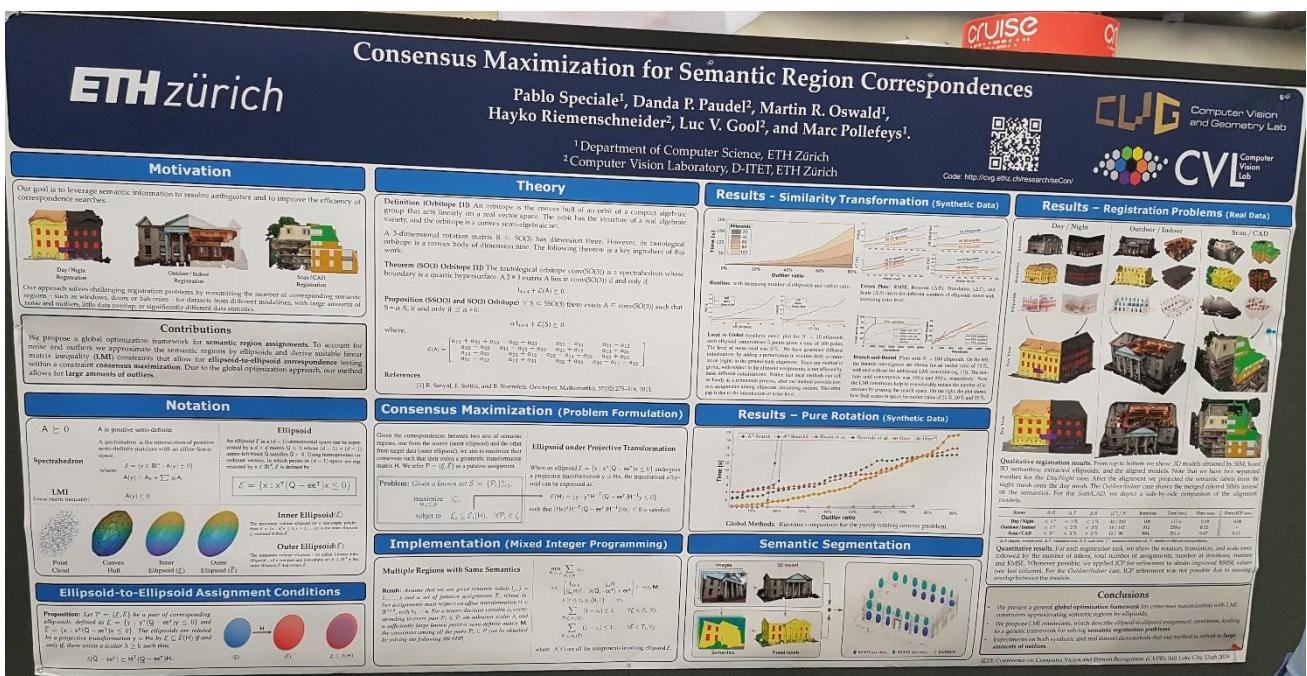
העבודה עוסקת בהתאמה בין מודלים תלת ממדים אשר יוצרים אופנים שונים (modalities) 1. מודל CAD. 2. שחזור 3D סמנטי על ידי SfM + סמנטיקה 3. תמונות יום / לילה 4. חוץ / פנים.

היעיון הוא למצוא על ידי סמנטיקה איזוריים דוגמת חלונות ודלתות, לאחר מכן, עברו כל אחד מהאיזוריים:

- לייצג כל אחד על ידי ענן נקודות 3D.
- למצוא hull convex של ענן נקודות.
- להתאים 2 convex hull אליפסואידים – פנימי (חסום על ידי hull convex) וחיצוני (חסום את hull).

כל אחד מהאליפסואידים מחושב עבור מודליות אחרות. לדוגמא חיצוני עבור מודל CAD ופנימי עבור SfM סמנטי.

לאחר מכן המטריה להתאים בין האליפסואידים של כל האיזוריים בתמונה כך שהפנימי יש בתוכו החיצוני. בעובדה זו מוצג פתרון אופטימלי לבעה זו אשר מבטיח למצוא פתרון (אם כי זמן הריצה יכול להיות ארוך).



0830-1010 Session 3-1C: Applications (Room 255)

2. [D18] Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics

שיטת לבחירת ערכי משקלות טובים לפונקציות מחיר מקרים בהם נלמדים מספר דברים במקביל (עם פונקציות מחיר שונות). במאמר מוצגת אפליקציה מעניינת מאוד: שערור של מפות עומק, סגמנטציה וסיג'ו עבור מצלמה הנמצאת במכונית נוסעת – יש סרטון יפה מאד בפייסבוק.

1010-1230 Poster Session P3-1 (Halls D-E)

Object Recognition & Scene Understanding

2. [E22] Show Me a Story: Towards Coherent Neural Story Illustration

המחשת טקסט של סיפור באמצעות תמונות.

4. [F4] Fast Spectral Ranking for Similarity Search

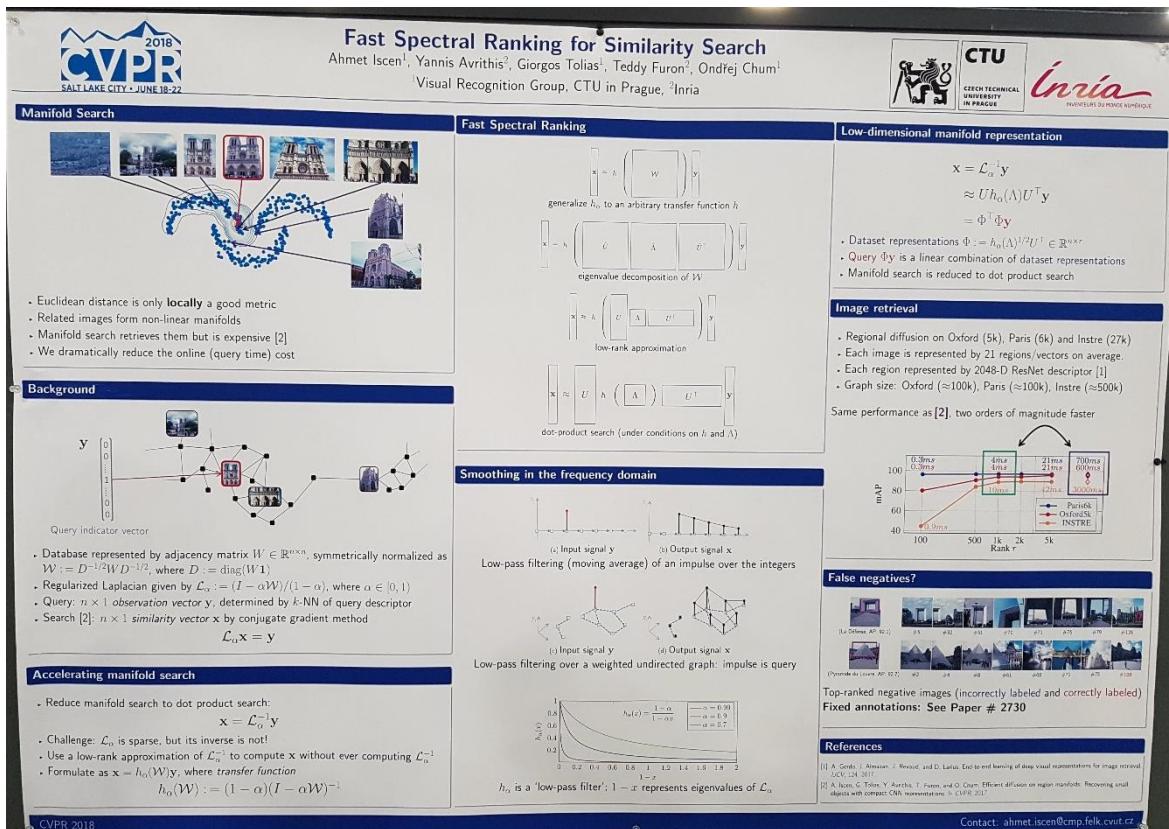
חישוב מהיר של דיסקריפטורים קרובים במרחב רב ממד'.

המשימה היא אוחזר תמונה מתוך בסיס נתונים גדול דוגמת Paris 5K, Oxford 6K. בעבודה קודמת המחברים הראו באופן אempirical שחישוב שכן קרוב במובן של מרחק אוקלידי הוא לא אופטימלי, כיון שדיסקריפטורים של התמונות של אותו העם מנוקדות מבט שונות נוטות להתקבץ על יריעה manifold, כך שעדייף לחפש על היריעה ולא על פי מרחק אוקלידי (בעבודה הקודמת הראו שחישוב כזה מוביל לשיפור של עד 20% בAPm). הבעה היא שחישוב על יריעה יקר מאוד מבחינה חישובית – בערך פי 100 מחישוב מרחק אוקלידי.

בעבודה זו מציעים שיטה מהירה משמעותית לחישוב על יריעה. מייצגים את בסיס הנתונים על ידי גרפ. מתארים את הגרף על ידי מטריצת סמיכות adjacency matrix w . בכך אפשר לעבור לティור מקביל על ידי לפלייאן L.

בහינתן לפלייאן של גרפ L אפשר לחפש על היריעה על ידי חישוב ההופכי L^{-1} , אך כיון שמדובר במטריצה ענקית זמן החישוב והזיכרון הדורשים גדולים מאוד ואיים מעשיים. בעבודה זו מציעים לפרק את A לערכים עצמיים, לקחת את 100-5K הערכים העצמיים הגדולים ביותר ולקרבב באמצעותם את w ו. לדברי המציג, בדרך זו מוגעים לביצועים דומים לפחות לא קירוב עבור כ-100 ערכים עצמיים. התהילה'ך עצמה גם בשלב ראשון של חישוב שכן קרוב במובן של מרחק אוקלידי.

זמן החישוב של התהילה'ך המוצע הוא כ-0.1-0.4. זמן חישוב של חישוב שכן קרוב במובן של מרחק אוקלידי (בסביבה קרובות) הוא כ-0.001-0.05, וכך על ידי הוספה 10% בזמן החישוב מקבלים עד 20% שיפור בAPm.



Machine Learning for Computer Vision

35. [H22] Stochastic Downsampling for Cost-Adjustable Inference and Improved Regularization in Convolutional Networks

אפשרות שליטה על סיבוכיות הרשת בזמן ח'יזי', תוך אפשרות להתאים למשאים הפנויים.

43. [I16] Unsupervised Domain Adaptation With Similarity Learning

באמצעות שמירה על מרחוקים \ דמיון דומים בין קטגוריות שונות בדומיינים השונים

51. [J10] HydraNets: Specialized Dynamic Architectures for Efficient Inference

ארכיטקטורה דינמית הבוחרת איזה חלקים להפעיל עבור כל תמונה על מנת לשפר יעילות

54. [J16] OLÉ: Orthogonal Low-Rank Embedding - A Plug and Play Geometric Loss for Deep Learning

פונקציית מחיר חדשה המוצעת כתחליף loss / triplet margin, כך שבמרחב הייצוג המרחק בין דוגמאות (inter-class variation) מואתא קטgorיה יהיה קטן, ובין דוגמאות מקטגוריות שונות (intra-class variation) יהיה גדול.

58. [K2] Fast and Robust Estimation for Unit-Norm Constrained Linear Fitting Problems

שיטת אופטימיזציה רובסטית חדשה.

1250-1430 Session 3-2B: Machine Learning for Computer Vision IV (Room 155)

Orals

1. [C1] MapNet: An Allocentric Spatial Memory for Mapping Environments

שיטת למידת רשת למיפוי (פנימי) של מבנים במערכת צירי עולם, ולא מערכת צירים פנימית כנהוג, תוך אפשרות לעדכן המודל עם מדידות חדשות. שימוש במצלמת RGBD.

Spotlights

1. [C7] Generate to Adapt: Aligning Domains Using Generative Adversarial Networks

רעיון מעניין של למידת העברת דומיין באמצעות GAN. הGAN משמש להעברת דומיין בלבד – מזכיר לי קצת את הרעיון שהוצע בעבודה Semantic Visual Localization, גם שם נעשה שימוש בdecoder גנרטיבי על מנת ללמידה ייצוג מלא יותר (אם כי שם מדובר על השלמה פרטימ חסרים וכן מדובר על העברת דומיין). לא הבנתי את כל הפרטים.

3. [C11] A PID Controller Approach for Stochastic Optimization of Deep Networks

שימוש בעקרונות של בקר PID בתהיליך האופטימיזציה של רשתות נירוניים.

קשה מאוד להבין את הדובר... מראים שקיים קשר בין בקר PID לבין רשתות קונבולוציה: מבחינה איקוית ומתמטית.

4. [C13] “Learning-Compression” Algorithms for Neural Net Pruning

המטרה: לעשות שימוש ברשתות קלות וקטנות ביישומים. הגישה היא לבצע גיזום, המבוצע על ידי אופטימיזציה עם אילוץ על גיזום, בשונה מושיטות קודמות הגדמות את הרשת לאחר האימון. האילוץ מ被执行 על ידי הגבלת מספר המקדמים ברשת (על ידי הגבלת נורמת L0 של כל מוקדי הרשת). תוצאות מראות שהשיטה שלהם טוביה יותר מאשר שיטות גיזום אחרות.

5. [C15] Large-Scale Distance Metric Learning With Uncertainty

למידת מטריקת מרחק באופן יעיל עבור בסיסי נתונים גדולים.

אתגרים בלמידה דיסקריפטוריים: 1. אותו אובייקט מופיע בנסיבות שונות, מנחים, תוארה רעש ועוד. 2. שימוש בתריפלט מוביל להרבה מאוד פרמטריזציות אפשריות (n^3). לא הבנתי את ההסבר על השיטה בדיקון. הרצוי היה למדוד ייצוג כזה שדוגמאות קרובות במרחב הקטלט יהיו קרובות במרחב הייצוג. הבעיה שגם במרחב הקטלט לא תמיד ידוע אם הדוגמאות דומות כיון שיש רעים שונים כאמור לעיל. אם אני מבין נכון הפתרון המוצע הוא למדוד יחד גם את הקרבה האמתית וגם את הייצוג, אם כי לא הבנתי כיצד בדיקון עושים זאת.

להשלים!

11. [D5] Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions

בסיס נתונים חדש ומבחן השוואתי (Benchmark) של שיטות עדכניות לשערוך מודל מצlama – נראה מעניין!

המטרה: שחזור 6DOF מתוך ענן נקודות. (אותו המחבר של Semantic Visual Localization) הצגה של מספר שיטות לבצע לוקלייזציה שכזו. המציג מדגיש שהשיטה תוכל להיות רובוטית לשינויים גדולים בנסיבות דוגמת קיץ וחורף. בעבודה זו מציגים 3 בסיסי נתונים להשוואה Benchmark (אשר יהיו זמינים בהמשך השנה).

הראשון הוא בסיס נתונים יום לילה – יש תמונות וענין נקודות והמטרה לבצע לוקלייזציה.

השני הוא שינויים עונתיים העיקריים בעיר בעיר, והשלישי הוא שינויים עונתיים העיקריים מחוץ לעיר. שימושות: שחזור תמונה, שחזור מבנה, שחזור מסלול, עוד כמה שלא הספקתי לרשום.

1250-1430 Session 3-2C: Object Recognition & Scene Understanding V (Room 255)

1. [D11] Learning Descriptor Networks for 3D Shape Synthesis and Analysis

למידת דיסקריפטורים לייצוג צורות תלת ממדיות.

1. [D17] Learning Compositional Visual Concepts With Mutual Consistency

שימוש בGAN לשם למידת מאפיינים שונים המשפיעים על התמונה: תוארה, גיאומטריה או בלבד חסר ממד המבטא את מידת ההחזרות של משטח או גוף, מתור ויקפדייה), תוך שימוש בסיסי נתונים שונים אשר כל אחד מכיל חלק מהמאפיינים, ושימוש בראשת לשם ייצור דוגמאות אימון ריאליות.

1450-1630 Session 3-3B: Image Motion & Tracking (Room 155)

6. [H1] Real-World Repetition Estimation by Div, Grad and Curl

ניתוח תנויות מחזוריות בסרטים וידאו תוך חלוקה ל-3 סוג תנוצה בסיסיים (Div, Grad, Curl) – נשמע נחמד.

1450-1630 Session 3-3C: Machine Learning for Computer Vision VI (Room 255)

1450 Orals (O3-3C)

1. [H17] Feature Space Transfer for Data Augmentation

למידת אוגמנטציה של מנה של עצמים תלת ממדים במרחב המאפיינים באמצעות פירוק (Factorization) לשני גורמים נפרדים: מראה (appearance) ומנח (pose).

בסיסי נתונים סטנדרטיים בדרך כלל מכילים עצמים במינחים 'רגלים', אבל במצבים יש מינחים רבים נוספים של מינחים שאינם מופיעים בסיסי הנתונים. בעבודה זו מוצע לבצע אוגנטציה במרחב הייצוג כך שנוכל למדוד מינחים נוספים. המחברים מתייחסים לייצוגים שונים של אותו העצם במרחב הייצוג על נקודות שמרכיבותן עוקם על יריעה במרחב

היצוג, והמטרה היא לבצע אוגמטציה כך שהייצוג נע לנוקודות שונות על העקום. הדרך לעשות זאת היא על ידי הפרדה בין שני מרחבים \ "יצוגים: מראה Appearance ומונח pose. בעבודה מסוימת נקודה במרחב הביצוג, להטיל אותה למרחב המונח, לבצע הזזה למרחב המונח, ולאחר כך להטיל חזרה לעקום תוך חישוב הערך המתאים של המראת. במהלך האימון צריך לוודא שהמנוח אליו מזוזים אכן אפשרי. לשם גם משתמשים בשני פונקציות מחיר – אחת למידת מונח והשנייה למידת קטגוריה (לדוגמא אם העצם הוא כסא אז רוצים שהייצוג החדש יהיה גם הוא של כסא).

3. [H21] Detail-Preserving Pooling in Deep Networks

שיטת pooling חדשה אשר אמורה לשמור פרטים בתמונה טוב יותר מאשר שיטות קיימות: stride, max/avg pooling. הבעיה שבדרכן זו מפסדים פרטים קטנים. דרך אפשרית אחרת היא לכל אפשרות ללקחת את הפיקסל הבולט ביותר ביחס לשביבה, במקרה זה הפרטים אומנם נשמרים אך התוצאה הסופית לא חלקה ויפה. עם זאת, באופן הרובה פעמים עושים לבדוק את הפעולה הזו ברשותות עמוקות – כאשר משתמשים pooling. בעובדה זו מנוסים למצוא שיטה חדשה של pooling אשר משמרת את הפרטים הקטנים. בעובדה אחרת הגדרו פרטים על ידי צבע שונה מהסבירה שלהם. השיטה המוצעת היא עבר פיקסל מסוים בתמונה, דוגמאות הסביבה שלו ולאחר כך ממשיכו כמו: ממצאים מצד אחד סינון, ומצד שני לא עושים כלום – ולאחר כך ממחיתים אחד משני, ובדרך זו מקבלים את תമונת הഫשים המורכבת מהפרטים שהלכו לאיבוד בדרך בעקבות הסינון (וכנראה גם downsampling) – המציג מתיחת לשיטה זו בשם Inverse bilateral filtering. השיטה האחרונה היא שיטה 'קלאסית', ובעובדה זו מוציאים גרסה נלמדת שלה בשם DDP. יש עוד אבל לא סימטרי. יש מהו עם פרמטר בקרה בשם λ אשר ערכיהם שונים שלו. מוביילים להתחגויות שונות, לדוגמה ערך אינסופי מוביל להתנהגות דומה (אך לא זהה) לזה של pooling max. מבחינה ביצועים מוצג שיפור שיחס לשיטות קודמות (נראה לי שבמשימת סיווג CIFAR10 אך לא בטוח), כאשר בוצעה השוואת מספר רשותות | ארכיטקטורות מתקדמות.

ביצועים במשימה נוספת: סיווג ImageNet, גם כאן הציג שיפור ביצועים ביחס לרשותות קודמות, גם כאשר משתמשים ברשותות פשוטות יותר מבחינת מספר מקדים.

בסיכום המציג אמר שהשיטה המוצעת מציעה שיפור לא כל כך גדול, אך כן עיקבי, ביחס למספר ארכיטקטורות קודמות.

1534 Spotlights (S3-3C)

2. [I3] Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions

הצעת מודול תחליף לשכבות קונבולוציה: הזזה + קונבולוציה חד מימדית. אמור להיותיעיל שימושית ולאפשר trade-off בין ביצועים ליעילות חישובית. מעניין – לבדוק!

המטרה: להקטין את הסיבוכיות החישובית של רשותות קונבולוציה, כך שתיאימו גם למערכות מושבצות. קונבולוציה מרוחבית יקרה כיון שעלות החישובית שלה היא ריבועית ביחס לגודל המנסון. בעובדה ההזזה מוציאים לביצוע שימוש בפעולת הזזה shift. לא הבנתי בדיק כיצד מבוצעת הפעולה. ערכים שונים מזוזים לכיוונים שונים. הפעולה ההזו דורשת אפס FLOPS ואפס פרמטרים ולכך היא חינמית. עשו שימוש בפעולה זו על מנת ליצור מספר רשותות והשוו לרשותות קודמות, לדוגמה Alexnet על Imagenet – שיפור של פי 77 בגודל המודול ייחד עם 1.5% שיפור ביצועים. דוגמה נוספת הוספה שלא הספקתי לרשום. דוגמא אחרתה בנושא style transfer, שם השיגו שיפור של פי 6 בגודל המודול עבור ביצועים דומים.

9. [I17] NISP: Pruning Networks Using Neuron Importance Score Propagation

שיטת לגזימה (Pruning) של רשותות קונבולוציה תוך ביצוע אופטימיזציה גלובלית על כל מקדמי הרשת כך שהמאפיינים בשכבה الأخيرة יובילו לתוצאות הטובות ביותר. מוצג שיפור שימושית ביעילות לצד ירידת קטנה ביצועים.

12. [J1] 3D Semantic Segmentation With Submanifold Sparse Convolutional Networks

רשות קונבולוציה המתאימה למידע דליל (sparse) דוגמת ענן נקודות 3D (دليل לעומת תМОנות), עם הדגמה של סגמטציה סמנטית של ענן נקודות 3D.

1630-1830 Poster Session P3-2 (Halls D-E)

Biomedical Image

2. [J7] An Unsupervised Learning Model for Deformable Medical Image Registration
5. [J13] CNN Driven Sparse Multi-Level B-Spline Image Registration
7. [J17] 3D Registration of Curves and Surfaces Using Local Differential Information

רגיסטרציה תלת ממדית בין עקומות למשטחים.

Machine Learning for Computer Vision

16. [K13] Spatially-Adaptive Filter Units for Deep Neural Networks

פילטרים בהם מקום המקדמים נלמד, ולא נגם מרשת (grid).

17. [K15] SO-Net: Self-Organizing Network for Point Cloud Analysis

רשת הלומדת ייצוג (descriptor) של ענן נקודות 3D. במאמר מוצגות מספר אפליקציות תוך שימוש בייצוג הנלמד, דוגמת שחזור, זיהוי וסגמטציה.

- 20.[K21] Explicit Loss-Error-Aware Quantization for Low-Bit Deep Neural Networks

שיטת לקואנטייזציה של מקדמי הרשות לשם שיפוריעילות (זיכרון + חישוב) ללא פגיעה משמעותית בBITS.

Applications

23. [L5] ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing

המשתמש ב spatial transformerGAN על מנת לשלב אובייקט בתמונה רکע באופן ריאלי. יכול לעזור לאלאן \ כפיר.

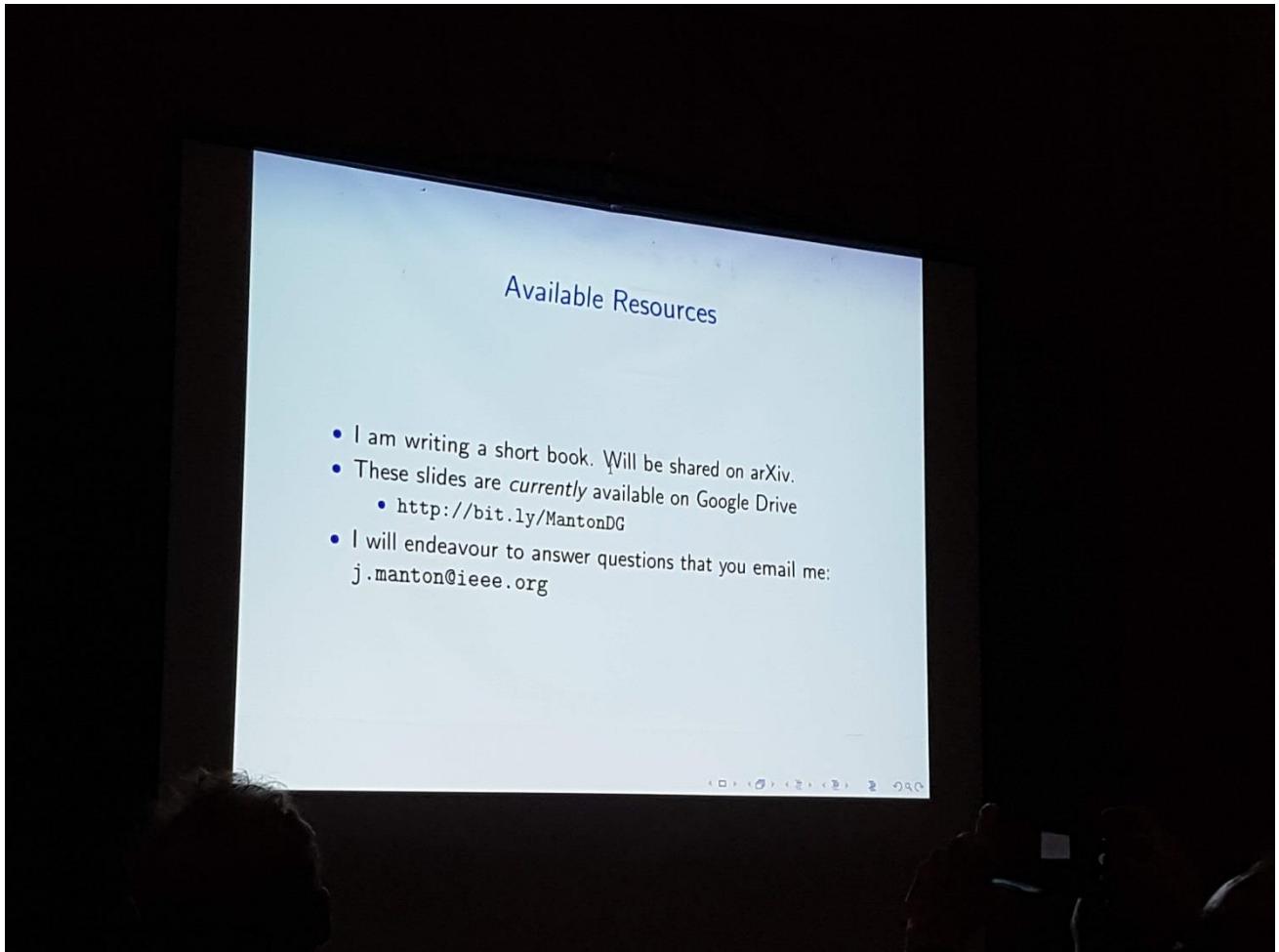
22.06.2018

0830-1230 Tutorial: Differential Geometry for Engineers (Ballroom H)

מדריך בנושא גיאומטריה דיפרנציאלית.



Differential
Geometry for Enginee



0840-1710 Workshop: Visual Odometry and Computer Vision Applications Based on Location Clues

1:30 pm - 2:10 pm, Keynote Talk: Ruigang Yang (Baidu Research & University of Kentucky)

השיעור עוסקת בפלטפורמה פתוחה עבור נהיגה אוטונומית בשם **SalloApp** המפותחת על ידי חברת **Baidu**. המערכת מורכבת מחומרה, תוכנת קוד פתוח ו咎ת תמייה בחישובי ענן (לא פתוח). הפרויקט התקדם באופן הדרמטי וכעת נהגים בככבים מהירים. המערכת כוללת לידר, מצלמה וגם רדאר, ויש מערכת זולה יותר (גראה לי ללא לידר או רדאר, לא זוכר אם אחד מהם או שניהם).

אפלו הוא סט של כלים פתוחים לביצוע מחקר בנושא נהיגה אוטונומית.

לפרויקט יש רכב לאיסוף מידע המגיע עם הרבה מאוד סנסורים: מצלמות רבות עם פוקוסים שונים, לידר זמן אמיתי, רדאר, מצלמות סטריאו, וכו'.

חלק מהמערכת שחררו גם בסיס נתונים עם נתונים רבים: תמונות מקור, סגמנטציה סמנטית, מפת עומק 3D. התמונות צולמו בסין ומכללות מגוון תרחישים הכוללים גם אנשים רבים ומספר מכוניות בסצנה.

מערכת לתיאוג 3D/2D:

תחליה אוספים מספר ענייני נקודות של איזוריים ומבצעים להם רגיסטרציה. לאחר מכן מבצעים קלאסטרינג, כאשר המטרה בתיאוג הוא לבצע תיאוג עבור כל קלאסטר. לאחר מכן אפשר להטיל נקודות 3D לתמונות 2D וכן מקבלים תיאוג 2D. הוצג GUI לתיאוג בסיס נתונים 3D וגם 2D. בנוסף לתיאוג הסמנטי יש גם Lane markings (אני מבין שהכוונה לסייעו דרך על הכביש: קו הפרדה, מעבר ח齊ה וכו') – אמר שהבינו זהה חשוב.

אתגר עתיד להתקיים ב-ECCV 2018: 3 משימות:

1. סגמנטציה עדינה (fine grained) של סימוני דרך על הכביש (מעבר ח齊ה, קו הפרדה וכו').

2. שערוך מיקום עצמי *on the fly* : נראה רלוונטי מאוד למקומות בהם אין ניחוש ההתחלתי ([קישור](#))
המשימה היא לשערוך מצלמה (6DOF) מתוך וידאו של רכב נושא, בזמן אמיתי ולא מדדים נוספים.
שאלתי אם משתמשים בנתוני IMU/GPS – אמרו שלא

שאלתי האם נתונים מפת תלת ממדית של העיר – אמרו שלא כיון שהיא גדולה מדי, אבל כן נתונים מנה בכל נקודה כך שלמעשה אני צריך לחשב \ לבנות עצמי את המפה תוך כדי התנועה (לא ברור מה בדיקות נתונים – לא יכול להיות שנ נתונים מיקום כיון שהוא צריך למצוא, אולי נתונים רק זווית?).
בדיקה באתר נראה שנותנים ענן נקודות 3D סמנטי של האיזור.

3. הבנת סצנה 3D: המשימה היא שערוך מודל \ צורה 3D של רכבים מתוך תמונה (לא ברור לי אם תמונה בודדת או וידאו).

פתרונות – מאמראים

מאמר ראשון – מיקום וסגמנטציה סמנטית בעזרת מפה סמנטית 3D. שהוא כמו DelUSac
היתוך מודיע מספר סנסורים חיוני במיקום עצמי אוטומטי. בדרך כלל ל모צרים צרכנים יש רוש ולן צריך לטיב את הפתרון.

מנחים שמקבלים:

תמונות עם שייר לנוטוני תלמטריה GPS וIMU
משהו נוסף.

המטרה: ליצור שערוך מנה מחזיק יותר.

שלב ראשון: ייצור מפה סמנטית.

שלב שני: שימוש ברשות עמוקה על מנת למצוא מנה בתהיל'ר איטרטיבי'

שלב שלישי: ייצור מפה סמנטית מתוקנת

שלב רביעי: שימוש ברשות על מנת לייצר מפה סמנטית מתוקנת (?? נראה לי' אחת בP2 והשנייה בP3).

תוצאות כמותיות: הוסיף רעש גאוס' למוח האמייתי ובדקו תוצאות שחזור. התוצאות הטובות ביותר התקבלו עבור שימוש בNNR. ביצעו את אותו הניסוי עבור 2 בסיסי נתונים.

הציגו תמונות סגמנטציה D2 מתוך מצלמת רכב הנושא בעיר – נראה יפה.

שאלן לגבי עלות התulos: ענו שלא יכולים לתת מספרים, אבל מדובר בעלות גדולה, אך עם זאת הם כל הזמן משפרים את הכלים האוטומטיים כך שצריך פחות עבודה של מתיגים אנושיים.

2:10 pm - 2:50 pm, Keynote Talk: Anelia Angelova (Google Brain), Talk topic: Unsupervised Learning of Depth and Ego-motion using 3D Geometric Constraints

המטרה: לשחזר מפת עומק ומיקום מצלמה motion Ego (אולי הכוונה לתנועה עצמית,(Clomer יודיעם את מיקום המצלמה באופן ייחסי לתמונות קודמות אבל לא ביחס לעולם) מתוך תמונות בלבד של רכב נושא בגישה supervised. אמרה שחייבי עומק לא כל כך אמינים ולכן עדיף להסתדר בלבדיהם.

מה עושים אם אין GPS? משתמשים בראיה ממוחשבת! לדוגמה: ניוט במאדים – מבוסס ראייה ממוחשבת visual odometry.

מה עושים אם אין חיישני עומק? לומדים עומק מתוך תמונה! המצלמה (רכב) נעה, והמטרה היא לשחזר את העומק (밀ולית SfM מבנה מתוך תנועה). אני מבין שעושים זאת באמצעות למידה עמוקה, כאשר גם העומק וגם המצלמה נלמדים.

כיצד מאמנים? מקבלים מספר תמונות (אמרה אחת או שתים), מכניסים לרשות ומנסים לשערך את העומק והמצלמה. נניח תמונה ומספר עומק בזמן t ותמונה $t+1$, אם מעריכים את התנועה של המצלמה אפשר להעיר כיצד תראה תמונה העומק הבאה. מצד שני אפשר להשווות את השערך לתמונה שהתקבלה – וכך לקבל אילוץ (?). יש עבודה צזו של Zhou מ2017: העבודה הראשונה שלמדה באופן supervised את המשימה הנוכחיית.

אנו מניחים שהמצלמה היא גוף קשיח שיש לו סיבוב והזזה, ומשתמשים באילוץ זהה במהלך האימון באמצעות פונקציית מחיר. אני מבין שיש פונקציית מחיר נוספת הדורשת דמיון (?) פוטומטרי – וכאשר עושים שימוש ביותר מפונקציית מחיר אחת מקבלים תוצאות טובות יותר.

כיצד לשלב את התוצאות בעולם 3D? מעבירים את פונקציית המחיר ל3D! זה החידוש בעבודה זו ביחס לקודמות. משתמשים באילוץ גיאומטרי 3D.

גישה:

מתחלים עם התמונה הנוכחיית ונראה לי גם עומק. מעריכים ענן נקודות. ועכשו נראה לי מקבלים תמונה אחרת, מעריכים ענן אחר, וצריכה להיות התאמה בין העננים השונים. מתחשבים גם בתנועה המשוערת של המצלמות בין התמונות.

ICP: גישה ידועה לחישוב טרנספורמציה בין עיקומות \ משטחים וcad. מוצעים בצורה איטרטיבית. בעבודה זו מימוש פועלה מקבילה באמצעות TensorFlow. בדקו את התוצאות עם \ ללא ICP הנ"ל, ונראה שיש שיפור קל בביטויים כאשר משתמשים בו.

פונקציית המחיר הכוללת מרכיבת מספר איברים – לא הספקתי לרשום, שהוא על איבר החלקה ואיברים גיאומטריים. מהهو על שימוש במסכות (אולי בשביל לבחור מאייזה איזוריים להתעלם ?)

תוצאות מראות שימוש במחיר גיאומטרי 3D מוביל לתוצאות טובות יותר מאשר בעבר.

הראתה מספר תוצאות השוואתיות לאחרים במספר מקרים. אם הבנתי נכון את התוצאות היה לך השוואת לעובדת שמשתמשת במצלמת סטריאו (כאן מצלמה יחידה), ובעוד שימוש בסטריאו היתה טובת יותר, העבודה זו התקרצה מבחינת הביצועים.

בחינה של יכולת הכללה על ידי שימוש בסיס נתונים שונים מעט שונה – ידאו שצולם מתוך נסעה על אופניים, ובדיקה על PDTK ללא טיפול הרשות: התקבלו תוצאות טובות ודי קרובות לאלו שלaimon על Cityscapes (אאולי גם טיפול על Kitty). ובדיקה על Kitty.

לאחר סיום האימון – צריך רק תמונה אחת בשבייל לבצע חיזוי عمוק (אמרה שצריך גם אורך מוקד), ורק שתי תמונות בשבייל שעורק ego motion.

איפה האלגוריתם לא עובד? לא הספקתי לרשום – יש שקייפ במצגת. נראה לי שיש בעיה כאשר יש עצמים נעים בסצנה.

כיוונים עתידיים: למצוא עצמים נעים על מנת לשפר דיקוגרף גם ...

סיכום: שעורק מנה ומקיים מצלמה מתוך ידאו באופן supervised. שימוש באילוצים גיאומטריים 3D שיפור ביצועים. שאלת: כיצד התוצאות ביחס ל slam orb?

תשובה: יש להם שתי גרסאות של slam orb: אחת שעושה שימוש ב 5 תמונות ואחרת גלובאלית שצוברת שגיאה, בתגובה לשאלת איך התוצאות של הגישה הגלובאלית אמרה שלא בטוחה זהה מוביל לתוצאות היכי טובות היום על בסיס נתונים Kitty.

שאלת: כיצד מחשבים scale?

תשובה: בעיות מסווג זה לא מחשבים scale, פתרים בכך שהוא scale invariant.

2:50 pm - 3:10 pm, Oral: A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration With Applications to UAV Geolocalization, Ahmed Nassar (IRISA Institute), Mohamed ElHelw (Nile University), Karim Amer (Nile University), Reda ElHakim (Nile University)

בעבודה זו ניסו לבצע רגיסטרציה בין תמונות אויריות שצולמו באמצעות חיישנים שונים. הציג מספר שיטות סטנדרטיות עבור שימושים כלליים.

הבעיה: בהינתן תמונה שצולמה מ UAV, צריך למקם אותם בתשתיות תמונות אחרות. נעזרים גם בסגנטציה סמנטית על מנת לעזור לפטור את הבעיה.

מקבלים 2 תמונות אחת מהרחפן והשנייה מגוגל מפות (תצלום לוויין). מקבלים את הנץ של 4 הפינות של תמונת הייחוס.

לא הבנתי בדיק – אמר שעושים התאמת עם SIFT, ואחכ אמר שעושים גם עם ORB (אולי orb רק בשבייל לעקב אחר שינויים בין תמונות).

סגנטציה באמצעות UNet, עבור 2 התמונות. לאחר מכן מבצעים ניתוח צורות ומקבלים מספר פרמטרים לכל עצם דוגמאות שטח רדיוס ועוד, ולאחר מכן משתמשים במאפיינים הללו keypoints על מנת לבצע התאמת טובת יותר.

תוצאות

2 ניסויים בשטחים שונים: אחד בגרמניה והשני ביון ??

מתוך כל ידאו חילצו את נתוני האמת של מסלול הטיסה, ולאחר בדקנו טיב שחזור של מסלול.

תוצאות תוך שימוש בסIFT: שגיאה של 50-55 מ' ללא SIFT (נראה לי רק עם orb), ושהגאה של 10-15 מ' עם SIFT.

תוצאות של סגנטציה סמנטית: מנסים לסוג בניינים ודריכים.

תוצאות עם שימוש במאפיינים שתוארו לעיל (שטח רדיוס ...) : השגאה ירדה ל 5-3 מ'.

3:50 pm - 4:30 pm, Keynote Talk: Andreas Geiger (MPI & University of Tübingen), Talk topic: Semantic Visual Localization

הוציג גם פוסטר בכנס המרכזית (קישור לסייע: Semantic Visual Localization (56.[O3])

המטרה: למצוא מקום מצלמה בעולם.

שיטות למציאת מקום:

GPS: בעיתי, הראה יידאו של מישחו מארן שמראה את הבעיה הקשה של שימוש בGPS בסביבה אורובאנית עם בניינים גבוהים, שם מגעים לשגיאה של 100 מ'.

מבוססת מפות: בהינתן תמונה ומפה סכמתית המטריה היא למקם את המצלמה בתוך המפה. השיטה שהציג מבוססת על visual odometry, עבר כל פריים מחשבים סיבוב והזזה. בתחילת לוקחים חלק של המפה ומניחים שהצלמה יכולה להימצא בכל מקום בפיג'ו אחד, ובמהלך הרצצת האלגוריתם המערכת מתכנסת לאט לאט. חיסרונו: התוכנות אורכת זמן, דורש מפות סכמטיות.

בעבודה זו: יש אוסף תמונות, מהם בונים מפה 3D על ידי שיטות דוגמת SfM, אחר כך מקבלים תמונה שאליה ציריך מקום אותו במפה.

העבודה מנסה להתמודד עם שינויים גדולים מאוד בתמונות, לדוגמא נקודת מבט שונה מאוד על הסצנה (עד 180 מעלות), עונה שונה (קיז' \ חורף עם שלג), יומן \ לילה – התמונות שהציגו נראות שונות מאוד וגם לבן אדם קשה להבין האם מדובר באותו המקום.

השיטה עשויה שימוש גם בגיאומטריה וגם בסמנטיקה.

לדברי המציג, שיטות קיימות אשר עושות שימוש בסקריפטורים לוקליים בעיתית במקרים כאלה כיוון שהמראות שונות משמעותית. גישה אחרת היא גישה סמנטית, בה רוצחים להבין מה על מה מסתכלים.

בעבודה זו לוקחים תמונה וDEPTH深深, עושים סגנטציה סמנטית באמצעות רשת קיימת, ומטילים את הסמנטיקה על מפת העומק כך שיש למפה סמנטית 3D.

בעיה - הסתירות: כאשר משנים את נקודת המבט אזי עצמים מסוימים בסצנה מסתירים עצמים אחרים, כאשר דפואו ההסתירה משתנה בהתאם לנקודת המבט, וכך שהICTURE תמורה בפועל של אותו המקום בעולם עשויה להיות שונה מאוד.

מה ניתן לעשות? לנוסות לקבל החלטה level high של הסצנה, לדוגמה שיש עץ ליד ספסל ומכוונית וכו'.

בעבודה זו לומדים דסקריפטור בשבייל לבצע השלה סמנטית של ציניה תלת ממדית. לוקחים את ענן הנקודות הסמנטי המתאים איזור בעולם מנוקדות מבט מסוימת, ומכניםים אותו לומד דסקריפטור המתאים את הסצנה, ולאחר מכן מנסה לשחזר את כל המידע התלת ממד' בסצנה – גם זה שמוצג בנקודת המבט הזה.

הראה מספר תמונות של שחזור ענן סמנט – נראה יפה מאוד.

האיזור בעולם שמתיחסים אליו הוא קופסה בגודל 10^3 מ'.

תוצאות:

Kitti חדש שעדין לא שוחרר (נראה לי שהמציג הוא מהאחים על kitti): מספר גרסאות של השיטה שלהם, סמנטיקה \ גיאומטריה, עם \ ללא צבירה של 5 פריים ?? . השוואה לשיטות קיימות בינהן גם posenet. התוצאות הטובות ביותר התקבלו על ידי שימוש בסמנטיקה ייחד עם צבירה. השיטה המתחרה היא שיטה קלאסית ללא למידה בשם 3DMatch. בתוצאות שהראה הגיעו לדיקוק של 5 מ' ב-95% מהמקרים עבור סיבוב של 90 מעלות, 5 מ' בקרוב 100% עבור סיבוב של 180 מעלות. תוצאה קצר מוזרה – הסביר אפשר שהוצע הוא שתמונות של צמתים עם 180 מעלות הם יותר קלות ??

דיקוק של 5 מ' – השאלה ביחס למה – מה גודל המפה בהם הם מוחפשים?

תוצאות נוספת: מידת עומק מלידר, אבל הדסקריפטורים אומנו על עומק מסטרeo.

המהירות היא בערך 15 קמ"ש, וצבריה על 5 תמונות – אפשר להעריך את הדינמיקה בין התמונות. קלט בבדיקה: 2 אפשרויות: עם \ ללא צבריה. לא הזכיר אך אני מבין שהיבת להיות גם מפה סמנטית S3. ללא צבריה: מקבלים 2 תמונות ממצלת סטריאו, באמצעות אחד מהם מחשבים תמונה סמנטית, ובאמצעות שתיהן מחשבים מפת עומק. מתחשבים במפת העומק עד לטווח של כ-30 מ' כיון שלאחר מכן רועש מדי ומפריע למערכת. שאלתי מה רצולצת העומק הדרישה, ענה שלא בדקנו אבל הרצולצת של Kitty היא 57 ס"מ ? לפיקסל. עם צבריה: לא הסביר, נראה 5 זוגות של תמונות סטריאו.

שאלתי עד כמה scalable למפות גדולות יותר, לדוגמה בגודל של עיר. ענה שלא בדקנו כיון שאין להם בסיס נתונים גדול יותר. עם זאת מדובר בשיטה מבוססת דסקריפטורים (אם אני מבין נכון: הקלט הוא קופסה בגודל 32 בשלישית, מתוכו מוצאים כמה אלפי דסקריפטורים ולאחר כך משתמשים RANSAC כאשר הפתרון המינימלי הדרוש הוא שיור אחד ולכן השיטה מהירה) וכן אמרה להיות scalable באופן דומה לשיטות מבוססות דסקריפטורים. המציג אמר שככל שמאגדים את המפה אזי יש יותר ויתר איזוריים דומים ובסופו של דבר זה קשה מדי לפתור (אני מבין שגם מבחינת זה שיש הרבה איזוריים דומים וגם מבחינת זה שיש זמן ריצה ארוך).

לאחר ההרצאה שאלתי שוב את המרצה על אופן הפעולה לאחר חישוב הדסקריפטורים. הוא הסביר שמדובר RANSAC כזה שבכל פעם דוגמים דוגמא בודדת. זאת כיון שמדובר בודדת אפשר למצוא גם מיקום וגם סיבוב: המיקום נמצא פשוט על ידי חישוב שcn קרוב ביותר. הסיבוב נמצא באופן הבא: במהלך אימון הרשת 'צילמו' כל מקום 16 מבטים, כך שעבור כל מקום יש 16 סיבובים אפשריים שנמצאים כבר בבטן של המודל (אני מבין שמייקום הוא קופסה בעולם בגודל 32^3 מ', כאשר את הקופסה זו סובבו כך שmbטאים עליה 16 כיונים שונים ובכל כיון יש הסתרות אחרות). את הדסקריפטור של השאלה משווים לכל אחד מ-16 המבטאים האפשריים, והסיבוב המשוער הוא זה הדומה ביותר.