ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DEEP LEARNING-BASED KEYPOINTS DRIVEN VISUAL INERTIAL ODOMETRY FOR GNSS-DENIED FLIGHT

M.Sc. THESIS

Arslan ARTYKOV

Department of Aeronautics and Astronautics Engineering

Aeronautics and Astronautics Engineering Programme

JULY 2023

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DEEP LEARNING-BASED KEYPOINTS DRIVEN VISUAL INERTIAL ODOMETRY FOR GNSS-DENIED FLIGHT

M.Sc. THESIS

Arslan ARTYKOV
(511201110)

Department of Aeronautics and Astronautics Engineering

Aeronautics and Astronautics Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Emre KOYUNCU

JULY 2023

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YAPAY SİNİR AĞLARI TABANLI NOKTA ÇIKARICILI GÖRSEL-ATALETSEL ODOMETRİ İLE GPS'SİZ ORTAMDA UÇUŞ

YÜKSEK LİSANS TEZİ

Arslan ARTYKOV
(511201110)

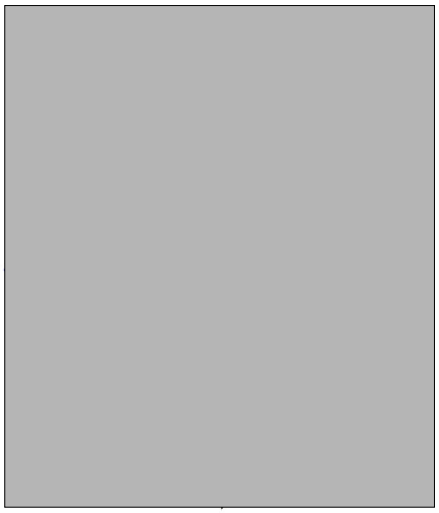
Uçak ve Uzay Mühendisliği Anabilim Dalı

Uçak ve Uzay Mühendisliği Programı

Tez Danışmanı: Assoc. Prof. Dr. Emre KOYUNCU

TEMMUZ 2023

Arslan ARTYKOV, a M.Sc. student of ITU Graduate School student ID 511201110 successfully defended the thesis entitled "DEEP LEARNING-BASED KEYPOINTS DRIVEN VISUAL INERTIAL ODOMETRY FOR GNSS-DENIED FLIGHT", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.



Thesis Advisor :	**Assoc. Prof. Dr. Emre KOYUNCU** Istanbul Technical University
Jury Members :	**Asst. Prof. Dr. Ramazan YENİÇERİ** Istanbul Technical University
	Asst. Prof. Dr. Hakan AKÇA Ege University

Date of Submission : **26 May 2023**
Date of Defense : **03 July 2023**

To my mother and father,

FOREWORD

I would like to thank myself, my family, and my advisor Assoc. Prof. Emre KOYUNCU for providing me with their support during the journey. Although it was a little bit long journey, I learned a lot of new things, developed a new perspective, and, most essentially, had a chance to try out different fields. Looking forward to switching to the next stage of my academic life.

JULY 2023

Arslan ARTYKOV
Aeronautical Engineer

TABLE OF CONTENTS

ABBREVIATIONS

AI	**:** Artificial Intelligence
ANN	**:** Artificial Neural Network
App	**:** Appendix
CNN	**:** Convolutional Neural Network
IMU	**:** Inertial Measurement Unit
RANSAC	**:** Random Sample Consensus
RNN	**:** Recurrent Neural Network
VINS	**:** Visual-Inertial Navigation System
VIO	**:** Visual-Inertial Odometry
VO	**:** Visual Odometry

SYMBOLS

p_t*	:	A warped point via Homography matrix
m	:	A distance between dissimilar descriptor vectors
s_i	:	Score
r	:	A probability of a keypoint being inlier or outlier
α	:	Alpha
β	:	Beta
λ	:	Lambda

LIST OF TABLES

LIST OF FIGURES

Page

DEEP LEARNING-BASED KEYPOINTS DRIVEN VISUAL INERTIAL ODOMETRY FOR GNSS-DENIED FLIGHT

SUMMARY

In this thesis, our primary objective was to comprehensively investigate and enhance the visual-inertial navigation capabilities of autonomous 3D quadrotor flight, which presents a multitude of challenges when operating in outdoor environments. The successful execution of autonomous flight in such settings necessitates the implementation of a navigation algorithm that is not only highly precise but also exceptionally robust, enabling the quadrotor to safely navigate through a variety of static and dynamic obstacles encountered along its flight path. A critical requirement for addressing this complex problem is the accurate estimation of the agent's ego-motion. While global positioning tools like GPS can offer valuable information for ego-motion estimation, their effectiveness is often limited in challenging scenarios where their signals may be distorted or unavailable. In light of this, we pursued an alternative approach by exploring the fusion of visual and inertial equipment to achieve more reliable and accurate ego-motion estimation.

Conventional visual-inertial odometry (VIO) techniques commonly suffer from performance degradation or even complete failure when confronted with the adverse conditions typically encountered in outdoor environments. Factors such as high dynamic range, textureless scenes, motion blur, and rapid motion can pose significant challenges to the accuracy and reliability of traditional VIO methods. To overcome these limitations and enhance the state-of-the-art in visual-inertial navigation systems (VINS), our research focused on the development of innovative techniques that go beyond traditional human-engineered salient point detectors. Instead, we leveraged the power of deep learning-based feature extractors to create a novel hybrid visual-inertial odometry method that exhibits remarkable resilience in the face of the aforementioned harsh conditions.

By adopting this approach, we not only achieved a significant improvement in the speed of the conventional algorithm but also witnessed a substantial enhancement in its accuracy and robustness. The integration of deep learning-based feature extractors allowed for more effective identification and tracking of relevant visual features, enabling the system to better handle challenging environmental conditions and dynamic flight maneuvers. Consequently, our proposed method not only offers superior performance in terms of ego-motion estimation but also paves the way for safer navigation in GPS-denied environments where traditional techniques may falter or become unreliable.

The outcomes of our research make a substantial contribution to the advancement of autonomous flight systems, providing a more reliable and effective solution for

visual-inertial navigation. By successfully addressing the limitations of conventional VIO techniques and improving the overall performance and resilience of visual-inertial navigation systems, we have opened up new possibilities for the deployment of autonomous quadrotors in real-world scenarios that were previously deemed too challenging or risky. The insights gained from our findings have the potential to shape the future development and implementation of autonomous flight technologies, enabling their broader adoption and integration into various industries and applications where precise and robust navigation capabilities are essential.

YAPAY SİNİR AĞLARI TABANLI NOKTA ÇIKARICILI GÖRSEL-ATALETSEL ODOMETRİ İLE GPS'SİZ ORTAMDA UÇUŞ

ÖZET

Bu tezde, temel amacımız, dış mekan ortamlarında otononom uçuşlarda karşılaşılan zorluklara karşı insansız hava aracının görüntüsel-ivmesel navigasyon yeteneklerini kapsamlı bir şekilde araştırmak ve geliştirmektir. Otonom uçuşun başarılı bir şekilde gerçekleştirilmesi, sadece son derece hassas değil aynı zamanda olağanüstü derecede sağlam bir navigasyon algoritmasının uygulanmasını gerektirir. Bu sayede insansız hava aracı, uçuş rotası boyunca karşılaşılan çeşitli statik ve dinamik engeller arasında güvenli bir şekilde seyahat edebilir. Bu karmaşık sorunu çözmek için kritik bir gereklilik, aracın hareket durum değişkenlerinin doğru bir şekilde tahmin edilmesidir.

Görsel navigasyon sistemleri, açık ve karmaşık ortamlarda aracın konumlandırmasını ve navigasyonunu sağlayarak otonom uçuşu mümkün kılar. Bu sistemler, GPS sinyalinin bozuk veya sınırlı uydu görünürlüğü gibi dış etkenlerden dolayı GPS sinyallerinin güvenilmez veya yanlış olduğu durumlarda önemlidir. GPS'in kullanılamadığı senaryolara örnek olarak iç mekanda uçuş, uzay uçuşu veya yüksek binaların olduğu yoğun kentsel ortamlarda manevra yapma gibi durumlar örnek olarak verilebilir.

GPS'in güvenli şekilde kullanılamadığı durumlarda otonom araçlar konumlarını ve yönlerini doğru bir şekilde belirlemek için alternatif yöntemlere başvurmak zorundadır. Geniş çapta tanınan ve yoğun bir şekilde araştırılan bir yaklaşım görsel odometri olarak bilinir ve hayvanlar ve insanlardaki navigasyon sistemlerinden ilham alır. Görsel odometri, bir veya birkaç kamera kullanarak aracın konumunu aşamalı olarak tahmin etmek için kullanılır. Kamera kullanımı sadece maliyet açısından avantajlı değil, aynı zamanda güvenilir konumlandırma ve navigasyon için temel olan zengin mekansal bilgileri de sağlar.

Tipik bir görsel odometri işlem hattı dört ana aşamadan oluşur. İlk olarak, her kamera görüntüsünden faydalı noktalar çıkarılır. Bu, ayırt edici noktaların belirlenmesini ve bunların görsel görünümlerini karakterize eden benzersiz tanımlayıcıların oluşturulmasını içerir. Bu noktalar, sonraki aşamalarda ayırt edici işaret noktaları olarak kullanılır.

Bir sonraki adımda, çıkarılan noktalar tanımlayıcı vektörlerin arasındaki mesafeler hesaplanarak ard arda gelen görüntüler arasında eşleştirilir. Bu eşleştirme süreci, görsel noktalar arasında bağlantılar kurarak, görüntülerin zaman içinde hareketlerinin takibini sağlar ve aracın hareketini tahmin etmek için önemli bilgiler sağlar. Nokta eşleştirmeleri sağlandıktan sonra, ardışık görüntüler arasındaki göreli konumu hesaplamak için geometrik kurallar kullanılır. Eşleşen noktaların gözlemlenen yer değiştirmelerinden yararlanarak ve sahnenin altında yatan geometriyi dahil ederek,

algoritma, aracın hareketini ve konum değişikliklerini zaman içinde tahmin eder ve doğru bir hareket rotasının oluşturulmasına katkıda bulunur.

Tahmin edilen konumların doğruluğunu daha da artırmak için, önceden tahmin edilen tüm konumlar optimize edilerek bir konum iyileştirme aşaması gerçekleştirilir. Bu optimizasyon süreci, aşamalı konum tahmininde ortaya çıkabilecek hataları ve tutarsızlıkları en aza indirmeyi amaçlar. Böylece daha hassas ve güvenilir bir konum tahmini elde edilir.

Görsel navigasyon, statik iç mekan ortamlarında tatmin edici tahminler sağlayabilirken, gerçek dünya koşulları birçok ele alınması gereken zorluklar sunar. Bu zorluklar arasında hareket bulanıklığı, yüksek dinamik aralık (HDR), dinamik dış ortam ve görüntü bozulması yer alır. Bu engelleri aşmak ve konum tahmininin güvenilirliğini artırmak için sistemimize ataletsel ölçüm birimi (IMU) sensörü dahil ediyoruz.

Kameraların aksine, IMU sensörü HDR gibi dış etkenlerden veya hareketli nesnelerin varlığından etkilenmez. IMU, kameralar tarafından yakalanan görsel bilgilere tamamlayıcı bir sensör olarak hizmet eder. Hem kameraların hem de IMU'nun entegrasyonu genellikle görsel-ataletsel odometri (VIO) olarak adlandırılır. Bu sensörlerin güçlerini birleştirerek, her birinin kısıtlamalarıyla başa çıkabilir ve daha sağlam ve doğru bir hareket tahmini hattı oluşturabiliriz.

VIO yaklaşımında, IMU sensörü kameralardan daha yüksek hızda çalışarak ajanın lineer ivmelerini ve açısal hızlarını hassas bir şekilde ölçer. Bu ataletsel ölçümler daha sonra iki ardışık kamera görüntüsü arasındaki hareket tahmini sürecine entegre edilir ve dahil edilir. Kameralardan gelen görsel bilgileri IMU'nun ataletsel ölçümleriyle birleştirerek, zorlu senaryoların bile varlığında hareket tahmini sürecinin genel doğruluğunu ve sağlamlığını artırabiliriz.

IMU'nun görsel odometri hattına dahil edilmesi birkaç avantaj sağlar. IMU ölçümleri hareket bulanıklığından etkilenmez, bu nedenle yüksek hızlı hareketin olduğu senaryolarda büyük avantaj sağlar. Ayrıca, kameralar tarafından tek başına doğrudan gözlemlenemeyen aaracın hareketi hakkında kritik bilgiler sağlar. Kameralar tarafından yakalanan görsel ipuçlarını IMU ölçümleriyle birleştirerek, VIO bize dinamik sahneleri, değişen aydınlatma koşullarını ve görsel navigasyon sistemlerinin doğruluğunu etkileyebilen diğer zorlu gerçek dünya faktörlerini ele alma imkanı sağlar.

Görsel ve ataletsel sensörlerin birbirini tamamlayan güçlerinden faydalanarak, görsel-ataletsel odometri yaklaşımı çeşitli ve dinamik gerçek dünya ortamlarında görsel navigasyonun kısıtlamalarını aşar. Bu sensör verilerinin birleştirilmesi, daha güvenilir ve doğru bir hareket tahmini hattı sağlar ve otonom sistemlerin geniş bir yelpazede zorlu senaryolarda artan hassasiyet ve güvenilirlikle navigasyon yapmasını ve çalışmasını mümkün kılar.

Klasik görsel-ataletsel navigasyon sistemleri (VINS), belirli senaryolarda ileri düzeyde doğruluk ve sağlamlık sergiler. Ancak, yukarıda bahsedildiği gibi zor durumlarla karşılaştıklarında, bu sistemlerin performansı olumsuz etkilenir. Dahası, görsel-ataletsel odometri (VIO) algoritmaları, titiz bir şekilde ayarlanmaz ve uygun şekilde kalibre edilmezse, yanlış konum hesaplamaları üretebilme eğilimindedir. Bir ataletsel ölçüm biriminin (IMU) görsel navigasyon hattına entegrasyonu, içsel ve dışsal

kalibrasyon prosedürlerini gerektirir. Doğru kalibrasyonun elde edilememesi veya dış zorlukların ve tasarım hatalarının ele alınmaması, yanlış tahminlere veya sistem hatalarına yol açabilir.

Görsel-ataletsel navigasyon sistemlerinin güvenilir ve doğru performansını sağlamak için kalibrasyon sürecine dikkat edilmelidir. İçsel kalibrasyon, kameranın iç parametrelerini, odak uzaklığı, merkez noktası ve lens distorsiyon katsayıları gibi parametreleri karakterize etmeyi içerir. Bu parametrelerin doğru şekilde belirlenmesi, kamera görüntüsünden hassas görsel bilgi çıkarmak için hayati öneme sahiptir.

Öte yandan, dışsal parametrelerin kalibrasyonu, kamera ve IMU arasındaki geometrik ilişkiyi belirlemeye odaklanır. Bu kalibrasyon prosedürü, iki sensörün koordinat sistemlerini hizalayan dönüşüm matrisini belirler. Doğru dışsal kalibrasyon, görsel ve ataletsel ölçümlerin doğru bir şekilde birleştirilmesi ve hareket tahmin sürecinin bütünlüğünün sağlanması açısından önemlidir. Doğru kalibrasyon elde edilemezse, konum tahmininde önemli hatalar ve navigasyon sorunları ortaya çıkabilir. Yanlış içsel kalibrasyon, görsel verilerde distorsiyon veya hizalanma hatalarına neden olabilir. Bu da yanlış özellik takibine ve konum tahminine yol açabilir. Benzer şekilde, dışsal kalibrasyon hataları, görsel ve ataletsel ölçümler arasında hizalanma hatası oluşturarak, tahmin edilen hareketin doğruluğunu etkileyebilir.

Yapay sinir ağlarının ortaya çıkmasıyla birlikte, araştırmacılar önceden geleneksel yöntemlerle modellenmesi zor olan sorunları yapay zeka ile çözme potansiyelini keşfetmiştir. Derin sinir ağlarının büyük etkisi olan bir alan da bilgisayar ile görüdür. Fotoğraf gibi görsel verilerin karmaşık ve yüksek boyutlu doğası, derin öğrenme tekniklerinin gücünden yararlanmak için uygun sebeplerdir.

Bu alandaki ilk çalışmalardan biri, konum tahmini için tamamen derin sinir ağlarının kullanmayı önermiştir. Bu fikir, derin öğrenme modellerinin çeşitli bilgisayar ile görü uygulamarında etkileyici yetenekler sergilemesi nedeniyle çekici görünmüştür. Bununla birlikte, konum tahmini, geometri ve dinamiğin derinlemesine anlaşılmasını gerektiren son derece doğrusal olmayan ve karmaşık bir problemdir. Bu içsel karmaşıklık, doğru ve güvenilir konum tahmini için yalnızca veriye dayalı öğrenmeye dayanmanın pratik olmadığını göstermektedir.

Araştırmacılar, derin sinir ağlarının gücünü klasik yöntemlerden gelen iyi kurulmuş prensiplerle birleştirerek, tamamen uçtan uca yöntemlerin sınırlamalarını aşan, güvenilir performans sunan ve aşırı durumlarla etkin bir şekilde başa çıkan hibrit yaklaşımlar geliştirmeyi başardılar. Bu sürekli keşif ve hibrit metodolojilerin geliştirilmesi, görsel-ataletsel odometri ve konum tahmini alanında daha pratik ve güvenilir çözümlerin yolunu açmaktadır.

Bu fikri daha da geliştirmek için, var olan VINS yöntemlerini derin öğrenme teknikleriyle zenginleştirerek, her iki yaklaşımın sağladığı avantajlardan faydalanmayı öneriyoruz. Yaklaşımımızın önemli bir yönü, geleneksel navigasyon yöntemlerinde kullanılan insan tarafından tasarlanmış nokta tespiti algoritmasını derin öğrenmeye dayalı tamamen uçtan uca bir nokta çıkarma tekniği ile değiştirmektir. Bu değişim, derin öğrenme modellerinin önemli noktaları doğrudan sensör verilerinden çıkarabilme yeteneğini kullanmamızı sağlar.

Araştırmamızda, önerdiğimiz yaklaşımın etkinliğini değerlendirmek için kapsamlı deneyler ve değerlendirmeler yaptık. Sonuçlar, derin öğrenmeye dayalı nokta çıkarma yöntemini VINS çerçevesine entegre ederek, sistem hızında önemli bir iyileşme elde ettiğimizi ve aynı seviyede doğruluk sağladığımızı gösterdi. Bu gelişme, hesaplama verimliliğinin önemli olduğu gerçek zamanlı uygulamalar için kritik bir öneme sahiptir.

1. INTRODUCTION

Visual navigation systems play a pivotal role in enabling autonomous 3D flight by providing essential capabilities for localizing and navigating in outdoor, complex environments. These systems are particularly vital in situations where GPS signals are unreliable or inaccurate due to external factors, such as signal interference or limited satellite visibility. Examples of such GPS-denied scenarios include indoor flight, space navigation, or maneuvering through dense urban environments with towering skyscrapers.

In GPS-denied cases, autonomous agents must rely on alternative methods to accurately determine their position and orientation. One widely recognized and extensively studied approach is visual odometry, which draws inspiration from the navigation systems observed in animals and humans. Visual odometry leverages one or two consumer-grade cameras to estimate the pose of an agent in an incremental manner. The utilization of cameras is not only cost-effective but also provides rich spatial information that is essential for robust localization and navigation.

A typical visual odometry pipeline comprises four main stages. Firstly, useful features are extracted from each frame, encompassing the identification of distinctive points and generating corresponding unique descriptors that characterize their visual appearance. These features serve as distinctive landmarks for subsequent pose estimation.

In the next step, the extracted keypoints are matched across consecutive frames based on the distances between their descriptors. This matching process establishes correspondences between the visual features, enabling the tracking of their movements over time and providing crucial information for estimating the agent's motion.

Once the feature correspondences are established, geometric rules are utilized to calculate the relative pose between consecutive frames. By leveraging the observed displacements of the matching points and incorporating the underlying geometry of

the scene, the algorithm estimates the agent's movement and position changes over time, contributing to the construction of an accurate trajectory.

To further enhance the accuracy of the estimated poses, a pose refinement stage is performed by optimizing all the previously estimated poses. This optimization process aims to minimize errors and inconsistencies that may have arisen during the incremental pose estimation, resulting in a more precise and reliable trajectory estimation.

While visual navigation can provide satisfactory estimations in static indoor environments, the real world presents numerous challenges that need to be addressed. Some of these challenges include motion blur, high dynamic range (HDR), the presence of dynamic agents, and visual degradation. To overcome these obstacles and enhance the reliability of pose estimation, we incorporate an inertial measurement unit (IMU) into the system.

Unlike cameras, the IMU is unaffected by external factors such as HDR or the presence of moving objects. It serves as a complementary sensor to the visual information captured by cameras. The integration of both sensors, cameras and IMU, is commonly referred to as visual-inertial odometry (VIO). By combining the strengths of these sensors, we can tackle the limitations associated with each and create a more robust and accurate motion estimation pipeline.

In the VIO approach, the IMU performs at higher sampling rates than cameras, providing precise measurements of the agent's linear accelerations and angular velocities. These inertial measurements are then integrated and incorporated into the motion estimation process between two consecutive camera frames. By fusing the visual information from cameras with the inertial measurements from the IMU, we can improve the overall accuracy and robustness of the motion estimation process, even in the presence of challenging scenarios.

The incorporation of the IMU into the visual odometry pipeline offers several advantages. The IMU measurements are immune to motion blur, making them particularly valuable in scenarios involving high-speed motion. Furthermore, they provide critical information about the agent's motion that is not directly observable

by cameras alone. By combining the visual cues captured by cameras with the inertial measurements, VIO enables us to handle dynamic scenes, varying lighting conditions, and other challenging real-world factors that can affect the accuracy of visual navigation systems.

By leveraging the complementary strengths of both visual and inertial sensors, the visual-inertial odometry approach overcomes the limitations of visual navigation in diverse and dynamic real-world environments. This fusion of sensor data ensures a more robust and accurate motion estimation pipeline, enabling autonomous systems to navigate and operate with increased precision and reliability across a wide range of challenging scenarios.

Classical visual-inertial navigation systems (VINS) demonstrate commendable accuracy and robustness in specific scenarios. However, when confronted with extreme cases, as previously mentioned, these systems encounter challenges that can impede their performance. Moreover, visual-inertial odometry (VIO) algorithms are prone to generating inaccurate pose calculations if not meticulously fine-tuned and appropriately calibrated. The integration of an inertial measurement unit (IMU) into the visual navigation pipeline necessitates precise intrinsic and extrinsic calibration procedures. Failure to achieve accurate calibration or address external challenges and design errors can lead to erroneous estimations or system failures.

In order to ensure reliable and accurate performance of visual-inertial navigation systems, attention must be given to the calibration process. Intrinsic calibration involves characterizing the intrinsic parameters of the camera, such as focal length, principal point, and lens distortion coefficients. Accurate determination of these parameters is vital for extracting precise visual information from the camera frames.

Extrinsic calibration, on the other hand, focuses on establishing the geometric relationship between the camera and the IMU. This calibration procedure determines the transformation matrix that aligns the coordinate systems of the two sensors. Accurate extrinsic calibration is crucial for accurately fusing the visual and inertial measurements and ensuring the integrity of the motion estimation process.

Failure to achieve accurate calibration can lead to significant errors in pose estimation and subsequent navigation. Inaccurate intrinsic calibration can introduce distortions or misalignment in the visual data, leading to erroneous feature tracking and pose estimation. Similarly, errors in the extrinsic calibration can result in misalignment between the visual and inertial measurements, compromising the accuracy of the estimated motion.

Moreover, external challenges such as variations in environmental conditions, sensor noise, or sensor drift can further impact the performance of the visual-inertial navigation system. Robust algorithms and appropriate sensor fusion techniques are required to handle these challenges effectively. Additionally, design errors, such as inadequate sensor selection, suboptimal algorithms, or improper sensor integration, can undermine the overall performance of the system.

Addressing these calibration, external challenges, and design error issues is crucial to achieving accurate and reliable pose estimations in visual-inertial navigation systems. Through meticulous calibration procedures, robust algorithms, and rigorous system design, the potential pitfalls and limitations of VINS and VIO algorithms can be mitigated, leading to improved accuracy and robustness in a wide range of challenging scenarios.

The universal approximation theorem [8] states that neural networks have the remarkable capability to approximate any continuous function to an arbitrary degree of accuracy. This fundamental theorem has played a significant role in shaping the field of deep learning.

With the emergence of deep neural networks, researchers have eagerly explored their potential in tackling problems that were previously challenging to model using traditional approaches. One domain that has been profoundly impacted by deep neural networks is computer vision. The intricate and high-dimensional nature of visual data, such as images, makes them well-suited for leveraging the power of deep learning techniques.

In recent years, researchers have come to recognize that neural networks excel in handling extreme scenarios that pose significant challenges for classical methods.

These scenarios include featureless environments, motion blur, and rapid changes in intensity. Neural networks can effortlessly adapt to such situations, exhibiting robust performance and surpassing the limitations of traditional approaches.

Among the early studies in this field, there was a suggestion to employ neural networks for learning pose estimation entirely. This concept appeared enticing, as deep learning models have demonstrated impressive capabilities in various computer vision tasks. However, pose estimation is a highly non-linear and complex problem that requires a thorough understanding of geometry and dynamics. This inherent complexity makes it impractical to rely solely on data-driven learning for accurate and reliable pose estimation.

Despite attempts to learn visual-inertial odometry (VIO) solely from data using fully end-to-end methods, practical and reliable solutions have proven elusive. The intricate interplay between visual information and inertial measurements, coupled with the need for accurate calibration and robust sensor fusion, poses significant challenges for purely data-driven approaches.

While neural networks have revolutionized computer vision and shown promise in addressing challenging scenarios, the nature of pose estimation and VIO necessitates a careful balance between data-driven learning and incorporating domain-specific knowledge and models. Hybrid methods that combine the strengths of deep learning with classical techniques have emerged as a practical approach to achieve accurate and robust pose estimation in real-world applications.

By harnessing the power of neural networks alongside well-established principles from classical methods, researchers can develop hybrid approaches that effectively handle extreme cases, offer reliable performance, and overcome the limitations of fully end-to-end methods. This ongoing exploration and refinement of hybrid methodologies pave the way for more practical and dependable solutions in the field of visual-inertial odometry and pose estimation.

1.1 Purpose of Thesis

The advancements made in visual-inertial navigation research have provided us with robust mathematical tools and efficient algorithms for accurate state estimation. These classical methods are widely recognized for their speed and accuracy, yet they often encounter challenges when operating in demanding outdoor environments. However, deep learning-based techniques have shown promise in their ability to adapt to adverse conditions, especially when trained on diverse datasets. By leveraging the strengths of both classical methods and deep learning, we can enhance the state-of-the-art visual-inertial navigation systems (VINS) and make them faster and more robust, particularly in challenging situations.

To achieve this goal, we propose augmenting existing VINS methods with deep learning techniques, thereby capitalizing on the advantages offered by both approaches. One key aspect of our approach involves replacing the conventional human-engineered feature detection algorithm used in traditional VINS methods with a fully end-to-end keypoint extraction technique based on deep learning. This substitution allows us to harness the capabilities of deep learning models to extract informative keypoints directly from the sensor data.

By integrating deep learning into the VINS pipeline, we can achieve several benefits, Figure 1.1. Firstly, deep learning models have the potential to adapt and generalize to a wide range of challenging outdoor conditions, including featureless environments, motion blur, and rapid changes in illumination. This adaptability arises from the ability of deep learning models to learn complex representations and capture intricate patterns present in the data.

Furthermore, the inclusion of deep learning techniques enables us to train the system on diverse datasets, capturing variations in environmental conditions and sensor characteristics. This data-driven learning approach allows the system to better handle situations that were previously challenging for classical methods, improving robustness and accuracy.

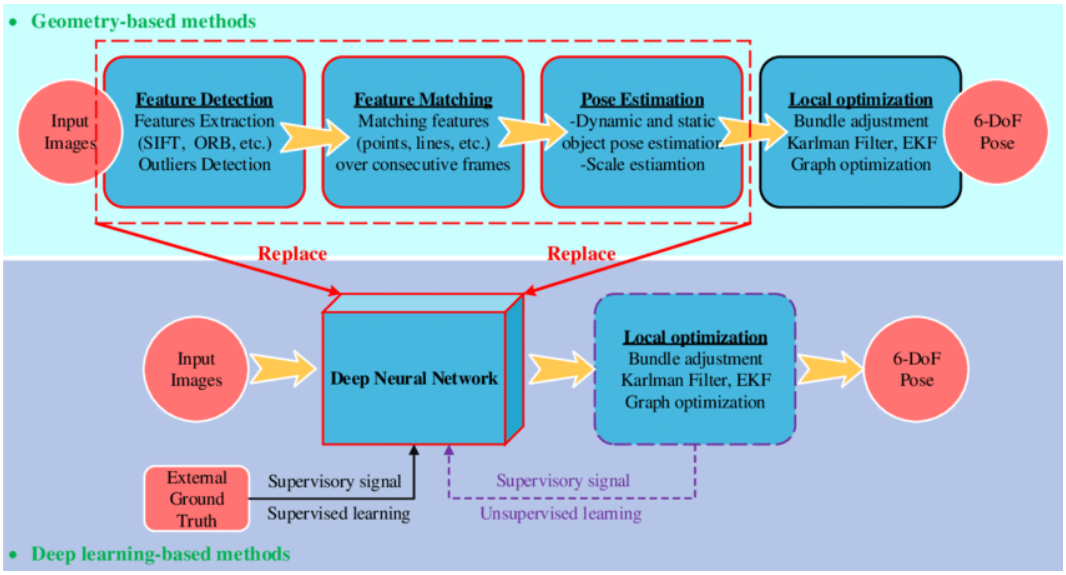


Figure 1.1 : Classical vs. deep learning-based visual localization pipeline.

In our research, we conducted extensive experiments and evaluations to assess the effectiveness of our proposed approach. The results demonstrated that by incorporating deep learning-based keypoint extraction into the VINS framework, we achieved a significant improvement in the speed of the system while maintaining a similar level of accuracy compared to the original method. This enhancement is crucial for real-time applications, where computational efficiency is paramount.

Overall, by fusing the strengths of classical VINS methods and deep learning techniques, we have developed a novel approach that enhances the performance of visual-inertial navigation systems in challenging outdoor conditions. Our method capitalizes on the adaptability and robustness of deep learning models, while preserving the speed and accuracy characteristics of classical methods. This advancement brings us closer to realizing more efficient and reliable autonomous systems capable of navigating and localizing themselves accurately in complex and dynamic environments.

2. BACKGROUND

2.1 Feature Extraction

SLAM, or Simultaneous Localization and Mapping, is composed of two parts: the front-end and the back-end. The front-end, also known as visual odometry, estimates the relative pose transformation between camera frames, while the back-end performs an optimization on the generated poses to improve their accuracy and build a detailed environment map [9]. Visual odometry can be achieved through feature-based or direct methods. Feature-based methods extract salient points from frames, while direct methods utilize photometric consistency between frames [9]. In this thesis, our main focus will be on feature-based methods. Feature extraction is a critical component of any feature-based visual self-localization algorithm pipeline, and finding appropriate features is crucial for achieving accurate self-localization.

To compute the relative motion between two consecutive frames, it is necessary to identify the previously detected points in the current frame [9]. To accomplish this, the detected salient points must meet certain criteria, such as locality, distinctiveness, generality, efficiency, and quantity. Features are distinguishable points in an image, such as corners, edges, and distinctive points. Corners, which are the simplest salient points in any image, lack rotation invariance and are therefore ineffective when the camera rotates [9]. Feature calculation also involves descriptor computation, which produces 1D vectors summarizing the information around a feature point. Descriptors are used to match the extracted points between frames. To be referred to as robust features, keypoints must satisfy several criteria [9]:

1) **Repeatability:** The extracted feature should exist in other frames.
2) **Distinctiveness:** The features should posses distinguishable characteristics from their peers.
3) **Efficiency:** The feature quantity should not exceed the pixel quantity in the same

image.

4) **Locality:** The feature should characterize only a small portion of an image.

There are various hand-crafted algorithms that satisfy some or all of these criteria. The well-known of them are SIFT [10], SURF [2], ORB [11], Shi-Tomasi [12]. Certain feature extraction methods are suitable for SLAM applications because they can be executed in real-time on a CPU. However, since the generation of accurate features requires a substantial amount of computational resources, algorithms attempt to strike a balance between precision and speed. For instance, while SIFT (Scale-Invariant Feature Transform) can provide highly accurate keypoints, the computational demands of the method are quite high [10]. On the other hand, ORB (Oriented FAST and Rotated BRIEF) features, which use FAST [13] keypoints and binary BRIEF descriptors, can be calculated significantly faster [11]. Therefore, ORB features are commonly preferred for real-time SLAM systems.

2.1.1 FAST (Features from accelerated segment test) feature detector

The FAST corner detector is renowned for its high speed, making it a popular choice for real-time applications such as simultaneous localization and mapping (SLAM) and visual odometry. In fact, many state-of-the-art visual-inertial self-localization algorithms rely on FAST for corner detection. The principle behind FAST is elegantly simple: it selects a pixel from an image and evaluates 16 surrounding pixels to determine whether they have higher or lower intensity values. If n pixels out of the 16 are all brighter or darker than the selected pixel, it is classified as a corner. Given its remarkable speed and accuracy, FAST has become an indispensable tool for feature extraction in various computer vision tasks [13].

2.1.2 SIFT (Scale invariant feature transform)

SIFT consists of several involved steps. First, it identifies candidate interest points where salient points might occur. This process is called scale-space peak selection. Then it localizes keypoints and calculates orientations for each of them. Finally, high dimensional vectors, known as descriptors, are computed for keypoints that uniquely describe each point. Because of the complex calculations SIFT can provide highly

accurate features, which possess the qualities such as distinctiveness, locality, quantity, and extensibility. On the other hand, the main drawback of the algorithm is its slow speed [10].

2.1.3 SURF (Speeded-up robust features)

SURF is relatively faster than SIFT, Figure 2.1. It can be used for real-time applications, like object detection, and tracking. The algorithm uses Hessian matrix to detect local and robust features. Therefore, it can demonstrate good performance [2].

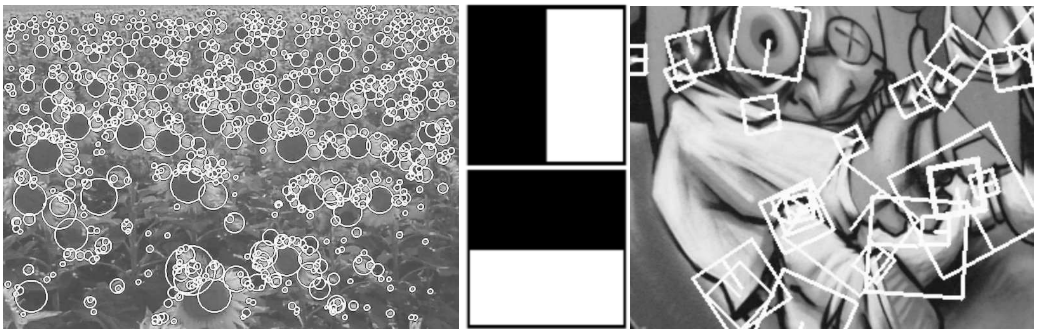


Figure 2.1 : Detected SURF features for sunflower field and graffiti scene [2].

2.1.4 BRIEF (Binary robust independent elementary features)

Although SIFT and SURF provide accurate features for computer vision tasks, they are still slow to be used in high-speed real-time applications. Therefore, BRIEF occurred as a solution to this problem. The main idea of BRIEF is to extract image patches around keypoints and define them as binary feature vectors. These feature vectors represent keypoint descriptors. A main drawback of the algorithm is the need for smoothing operations before calculating the descriptors since it works in pixel space where a lot of noise is present. Therefore, a Gaussian kernel is applied to smooth the patches [14], Figure 2.2.

2.1.5 ORB (Oriented FAST and rotated BRIEF)

ORB was invented by OpenCV in 2011. It is a cheap and efficient algorithm, that makes it a good candidate for real-time appications. ORB uses FAST keypoints with BRIEF descriptors. As it was stated in the previous section, FAST evaluates the

Figure 2.2 : A sample of the smoothed picture.

candidate pixel by comparing it with the surrounding points. If the pixel is decided to be keypoint, then an associated binary descriptor vector is calculated. It is called BRIEF descriptor. BRIEF descriptor is a vector that consists of 1s and 0s and uniquely describes the keypoint [11]. The famous ORB-SLAM [15] takes its name from this algorithm.

2.2 Feature Matching

Upon detecting salient points and computing unique descriptors, the next stage involves matching features between different frames or between frames and a reconstructed map. However, this phase is time-consuming and entails high energy consumption, and is particularly vulnerable to errors. Such errors are typically due to feature mismatch, which can be attributed to the locality of features and textureless areas in the scene that can cause the descriptors to appear similar to each other. The complexity of this matching process further compounds this challenge, requiring significant computational resources and specialized techniques to overcome these limitations [9].

In the context of feature matching, a commonly employed but simplistic method is brute-force matching. This technique involves computing the distance between the feature descriptors of two images and subsequently ranking the closest matches. The selection of distance measurement techniques is contingent on the nature of the descriptors. For instance, if descriptors are in the floating-point type, Euclidean

distance can serve as the most suitable metric, while the Hamming distance metric is preferable for binary descriptors. Despite its ease of implementation, this approach is computationally expensive and may lead to suboptimal performance when dealing with large datasets or complex features. Therefore, there is a need to explore more sophisticated matching algorithms that can enhance the accuracy and efficiency of this process [9].

2.3 Visual-Inertial Self-Localization

Navigating a robot in GNSS-denied environments is an extremely challenging task. To accomplish this, a visual-inertial system is employed, which must be highly efficient and lightweight, given the limited computation and memory resources available. In the aerospace industry, the use of lightweight and efficient hardware is just as important as the design of lightweight software. To achieve optimal performance, a combination of two complementary sensors - a camera and an Inertial Measurement Unit (IMU) - are utilized for visual-inertial navigation. Visual-inertial odometry (VIO) is a complex process composed of several components, Figure 2.3, that work together seamlessly [9]. The system computes agent location starting from a reference point and incrementally adding to the previous pose. It is especially useful in those cases when we do not have access to an accurate map of the environment.

The pipeline takes two inputs, namely camera frames and IMU measurements, to produce 3D pose estimates. The first step involves feature extraction and matching, where we detect and track useful features across consecutive frames [16]. Meanwhile, the IMU measurements, comprising radial velocity and linear acceleration, are pre-integrated to generate a single measurement for the duration between two camera frames. These measurements are then used to calculate the relative motion between the camera frames and obtain initial pose estimates. To refine these estimates further, a local bundle adjustment optimization process is performed, utilizing information from both visual and inertial measurements. Overall, the pipeline combines information from both camera and IMU sources to produce highly accurate 3D pose estimates. [17].

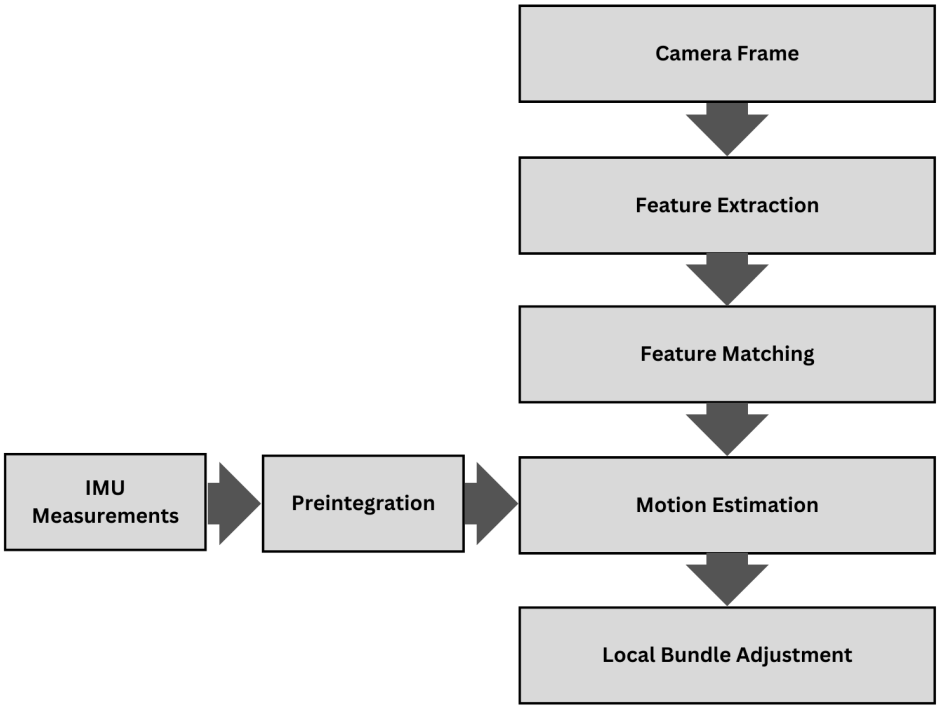


Figure 2.3 : VIO Pipeline.

2.4 Deep Learning

2.4.1 Feed-forward neural networks

Feed-forward neural networks are artificial neural networks that consist of layers of nodes, Figure 2.4. The aim of the whole network is to approximate a function $y = f(x;\theta)$, where x represents input data and θ represents network parameters. Each node in the layers is responsible for linear and non-linear computations. The linear calculation is simple matrix production, which is followed by non-linear function, such as ReLU, or Sigmoid. The stack of these layers is also called multi-layer perceptrons (MLP). We can build very large networks by stacking these layers and approximate highly non-linear functions.

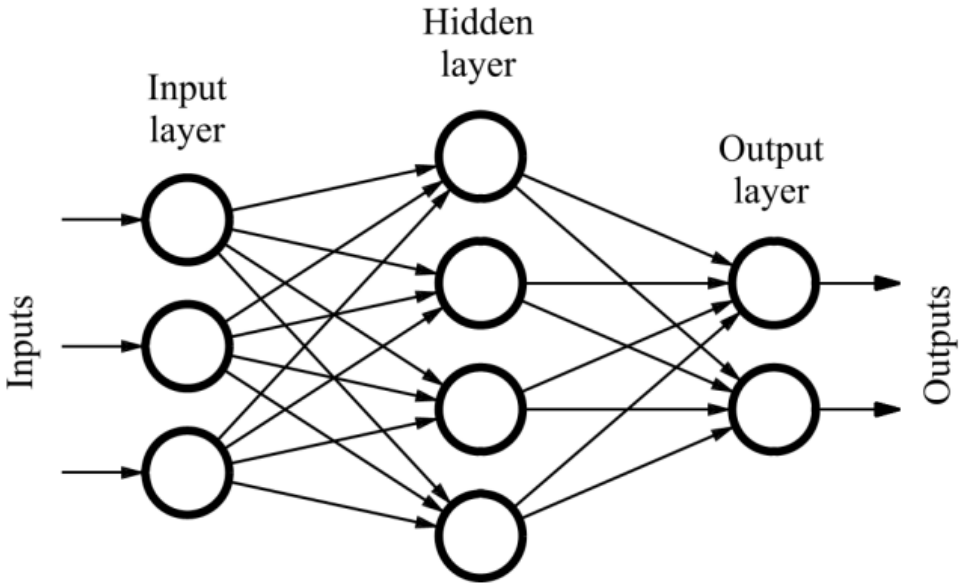


Figure 2.4 : A sample of feed-forward neural network [3]

2.4.2 Convolutional neural networks

Convolutional neural network (CNN) is a type of feed-forward networks, Figure 2.5. These networks apply convolution operation on 2D data in order to extract the spatial relationship between the patches. Each convolution layer is followed by non-linear layer, so that non-linearity is handled in the input data. The most preferred non-linear function for CNNs is ReLU. The kernels of CNNs share information through the layers. Therefore, they are effective in processing and extracting spatial information from 2D data such as images [4].

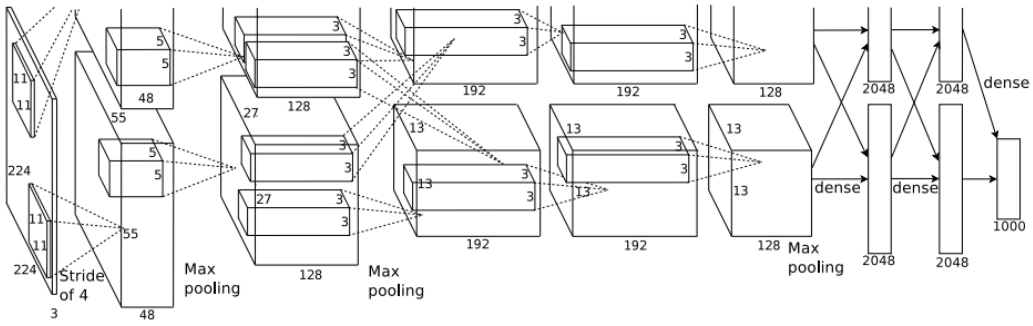


Figure 2.5 : A sample of convolutional neural network [4]

3. DEEP LEARNING-BASED KEYPOINTS DRIVEN VISUAL INERTIAL ODOMETRY FOR GNSS-DENIED FLIGHT

3.1 Related Works

3.1.1 Classical visual odometry methods

Visual odometry has been a subject of extensive research in the past, with numerous methods relying on well-established mathematical and physical principles. The literature on visual odometry is typically categorized into two main groups: feature-based techniques and direct methods. Feature-based techniques involve detecting sparse and salient points from consecutive camera frames and calculating corresponding descriptor vectors. The keypoints are then matched across frames by comparing the descriptors, and geometric constraints are constructed from the matched point pairs to estimate the relative motion between frames. On the other hand, direct methods follow a distinct approach and utilize the photometric consistency assumption to obtain relative pose. Direct methods estimate the relative pose by minimizing the intensity difference between image patches. Both feature-based techniques and direct methods offer unique advantages and disadvantages, and choosing the appropriate method is dependent on the specific application and computational resources available [16]. Some of the famous feature-based visual odometry methods are MonoSLAM [18] and PTAM [19]. Some of the famous direct methods are DTAM [20], LSD-SLAM [21], and DSO [22]. It was proven that direct methods perform better in textureless environments than feature-based methods.

3.1.2 Classical visual-inertial odometry methods

In real-world scenarios, we are likely to encounter various extreme conditions that can significantly degrade the quality of the images, making it difficult to extract useful features. Consequently, visual odometry can become error-prone, particularly when dealing with cases such as shadows, high dynamic range Figure 3.1, occlusion,

textureless scenes, and highly dynamic environments [23]. One potential solution to improve visual odometry's accuracy is to integrate an inertial measurement unit (IMU) into the pipeline. The use of an IMU can be advantageous due to its lightweight design and low energy consumption, making it an excellent candidate for the task at hand. We utilize inertial measurements in two ways, to calculate translation and velocity by integrating linear acceleration and to compute rotation by integrating angular velocity. However, high measurement noise makes the integration of IMU extremely challenging.

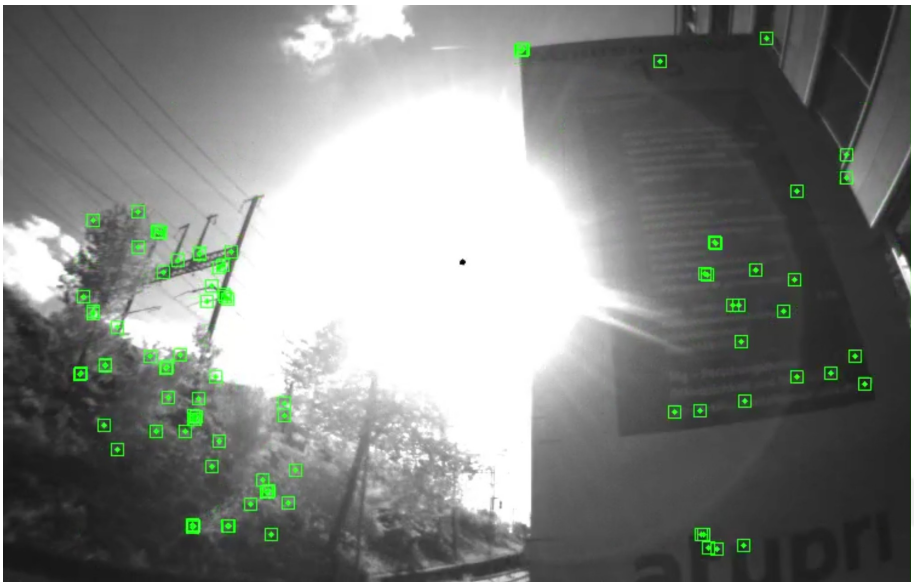


Figure 3.1 : Sample high dynamic range scene

Extensive research has been conducted in the past to effectively integrate these two sensors. The existing visual-inertial navigation (VINS) literature typically divides into two main approaches: loosely-coupled and tightly-coupled methods. Loosely-coupled methods process visual and inertial measurements separately and fuse them at the end of the pipeline. They usually adopt EKF for sensor fusion. However, tightly-coupled techniques such as OKVIS [24], VINS-Mono [6], ORB-SLAM [15], and MSCKF [25] integrate inertial measurements into the pipeline at the feature level. These methods also decrease scale drift caused by monocular view by leveraging metric space of inertial measurements.

3.1.3 Deep learning-based methods

Following the advent of convolutional and recurrent neural networks (RNN), there has been a proliferation of successful deep learning applications in areas such as object detection, semantic segmentation, object tracking, and scene flow estimation. Nevertheless, comparatively little work had been accomplished in the realm of learning-based state estimation until the arrival of the DeepVO [5] method. The DeepVO architecture is composed of CNN and RNN layers stacked one after another, Figure 3.2. Consecutive camera frames are sent to convolutional layers to extract useful features. Then the feature vectors are flattened and fed to a recurrent neural network, where sequential relationships between the frames are extracted and relative pose is regressed. DeepVO is the first method that demonstrated the fully end-to-end pose estimation from camera frames. Another method, which is known as ESP-VO [26], builds on top of the previous one by estimating pose uncertainty along with relative camera pose. The algorithm for the first time demonstrated that extreme cases such as high dynamic range, rapid motion, and textureless scenes can be handled by deep neural networks.

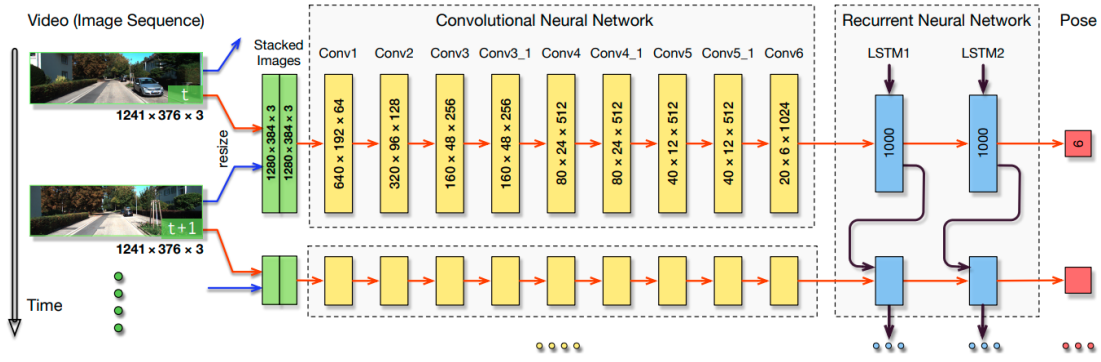


Figure 3.2 : The DeepVO architecture [5]

Another interesting approach was to model visual-inertial odometry with fully end-to-end neural networks. This method is called VINet [27], where authors adopted LSTM layers to process IMU measurements and CNNs to transform images into

feature vectors. The inertial and visual features are then concatenated and the relative pose is estimated at the end of the pipeline.

3.1.4 Hybrid methods

As we've noted, conventional methods often struggle with extreme cases in outdoor navigation. However, relying solely on fully end-to-end algorithms is also not an ideal solution for the odometry problem. The challenge is that this problem is highly nonlinear, making it difficult for neural networks to accurately interpret raw camera frames and generate precise pose information. To overcome this limitation, researchers have explored combining traditional and deep learning algorithms to leverage the advantages of both approaches. LS-Net [28] is one of the first works that used neural networks and hand-crafted feature extraction methods in the same pipeline. They proposed to employ LSTM layers to estimate the incremental update of the nonlinear least squares optimization problem. PF-Net [29] estimates robot pose by mimicking a particle filter algorithm with a neural network. [30] estimates the relative pose and the corresponding uncertainty from consecutive frames by employing DeepVO [5] and fuses with the inertial measurements in a loosely-coupled manner. They also incorporate EKF, where prediction is made from IMU data and update is performed with DeepVO estimations. [31] incorporates self-supervised learning pipeline to estimate relative pose, depth, and pixel uncertainty. Depth maps and poses are first used in front-end tracking. Then the estimated information is further utilized in back-end optimization. Overall, they utilize deep learning in the front-end and employ traditional non-linear optimization in the back-end. SuperPointVO [32] replaces human-engineered features with learned features by CNNs. They adopt a self-supervised framework, called SuperPoint [33], for interest point extraction. The network provides both keypoints and the corresponding descriptors at the same time. A more recent algorithm, named LIFT-SLAM [34], replaces the front-end of the state-of-the-art ORB-SLAM [15] with LIFT [35] based detector.

In our work, we adopt a hybrid approach for the real-time visual-inertial navigation of the quadrotor. By this, we aim to make state-of-the-art VINS more robust against

challenging extreme cases, that are impossible to be handled by human-engineered solutions.

3.2 Method

Accurate state estimation plays a pivotal role in the success of visual-inertial navigation systems (VINS). While this task is relatively simpler in indoor environments, outdoor scenarios present a multitude of challenges that can significantly impact the performance of VINS, especially in the context of drone flight. Outdoor phenomena often exert a negative influence on visual interest point detectors, which are commonly employed in VINS. Real-world tests have demonstrated that hand-engineered algorithms struggle to cope with these situations, rendering them ineffective. However, neural networks possess a remarkable capability to approximate any function and extract intricate patterns from data. Leveraging the adaptability of learning-based detectors across diverse scenarios and the efficiency of state-of-the-art conventional VINS, we made the decision to merge these two domains. Our research demonstrates that this hybrid approach yields robust and accurate real-time performance, which is vital for enabling autonomous 3D flight. By combining the strengths of both worlds, we overcome the limitations posed by outdoor challenges and achieve superior performance in state estimation, thereby enhancing the capabilities of visual-inertial navigation systems for drone navigation.

We build our system on top of the well-known traditional VIO, named VINS-Mono [6]. VINS-Mono is a monocular VIO, that employs a camera and IMU to estimate the 3D pose of a robot.

The algorithm comprises several key components, each playing a vital role in achieving accurate state estimation, Figure 3.3. It initiates by gathering visual and inertial measurements at different rates, harnessing the rich information provided by both sensor modalities. Within the visual domain, the algorithm employs feature detection and tracking techniques to identify and trace salient features across consecutive camera frames. Simultaneously, the IMU measurements are pre-integrated over the time interval between two consecutive frames, consolidating them into a single measurement that encapsulates the accumulated inertial information.

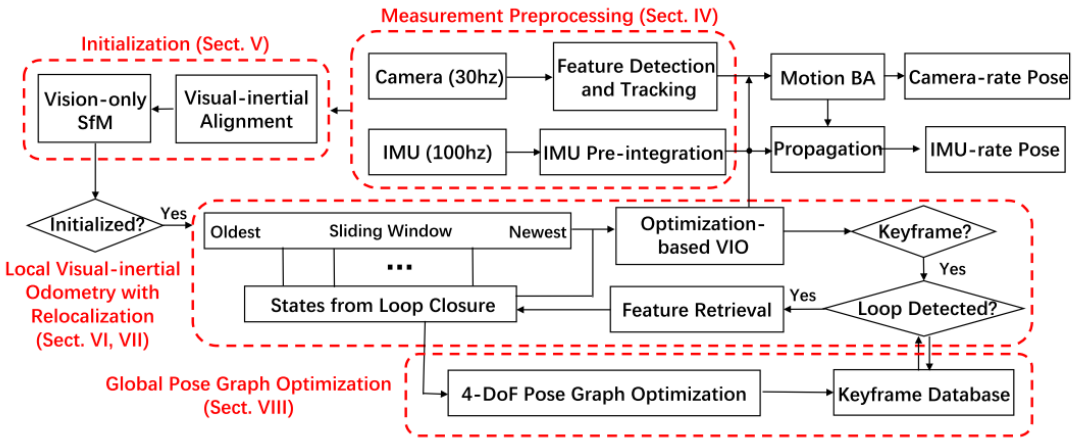


Figure 3.3 : The block diagram of VINS-Mono [6]

Following the collection of measurements, the algorithm proceeds to estimate the initial local poses based solely on the visual features. This step establishes an initial approximation of the drone's position and orientation within the local coordinate system. Subsequently, these initial estimates are aligned and refined using complementary inertial measurements, which provide additional cues for accurate pose estimation.

Once the initialization stage is complete, the algorithm enters an optimization phase aimed at further enhancing the accuracy of the estimated poses and the reconstructed point cloud. This optimization process, known as Bundle Adjustment, leverages mathematical techniques to iteratively refine the estimated poses and the 3D representation of the environment. By jointly optimizing these variables, the algorithm achieves a more precise alignment of the poses and the point cloud, ultimately improving the overall accuracy of the state estimation.

In addition to the aforementioned components, the algorithm incorporates a dedicated loop-closure block. This crucial module is responsible for detecting loops, which occur when the drone revisits a previously visited location, and leveraging this information to refine the global poses. Loop detection mechanisms analyze the accumulated visual and inertial cues to identify such revisits, and the obtained loop-closure information is utilized to correct and enhance the accuracy of the global pose estimates.

Through the integration of these various components, the algorithm encompasses a comprehensive pipeline for accurate state estimation in visual-inertial navigation systems. By leveraging both visual and inertial measurements, incorporating feature tracking, pre-integration, initialization, optimization via Bundle Adjustment, and loop-closure mechanisms, the algorithm achieves robust and precise real-time performance, facilitating reliable and autonomous 3D flight in dynamic outdoor environments.

In the VINS-Mono system, the Shi-Tomasi [12] corner extractor is employed to detect and extract relevant features from the visual input. While this algorithm exhibits fast processing capabilities, it is susceptible to failure in challenging scenarios such as textureless environments, rapid motion, and various outdoor phenomena. Recognizing the need for a more robust feature extraction approach without compromising real-time performance, we endeavored to enhance the pipeline.

In order to meet the desired specifications of improved robustness and real-time processing, we incorporated a Convolutional Neural Network (CNN)-based keypoint extractor known as KP2D [7]. Unlike the Shi-Tomasi algorithm, KP2D leverages the power of deep learning to detect and extract discriminative keypoints from the visual data. By employing a CNN architecture specifically designed for this task, KP2D can effectively handle challenging scenarios, including textureless environments, rapid motion, and other outdoor phenomena that often lead to the failure of traditional feature extraction algorithms.

3.2.1 Neural outlier rejection for self-supervised keypoint learning (KP2D)

In our pursuit of extracting robust features from the given images, we embraced the innovative KeypointNet framework, extracted from the remarkable KP2D research endeavor. The KP2D architecture encompasses two distinct networks, Figure 3.4: KeypointNet and IO-Net, each playing a crucial role in the feature computation process.

First and foremost, KeypointNet assumes the responsibility of detecting interest points within the images, while simultaneously calculating their corresponding descriptors

and scores. This pivotal network operates with precision, enabling the identification of salient keypoints that serve as reliable landmarks for subsequent analysis.

On the other hand, IO-Net undertakes the crucial task of segregating the estimated keypoints into two distinct sets: inliers and outliers. This separation mechanism renders the need for additional outlier rejection algorithms, such as RANSAC [36], unnecessary. By eliminating the reliance on supplementary algorithms, the method achieves a streamlined and efficient feature extraction process, capable of detecting false matchings without compromise.

Furthermore, the algorithm incorporates a source image, which is derived by transforming the target image using a known homography matrix. This inclusion facilitates the training of the entire pipeline in a self-supervised manner. By leveraging the knowledge of the homography transformation, the method capitalizes on a valuable source of supervision, enabling comprehensive and effective training of the feature extraction pipeline [7].

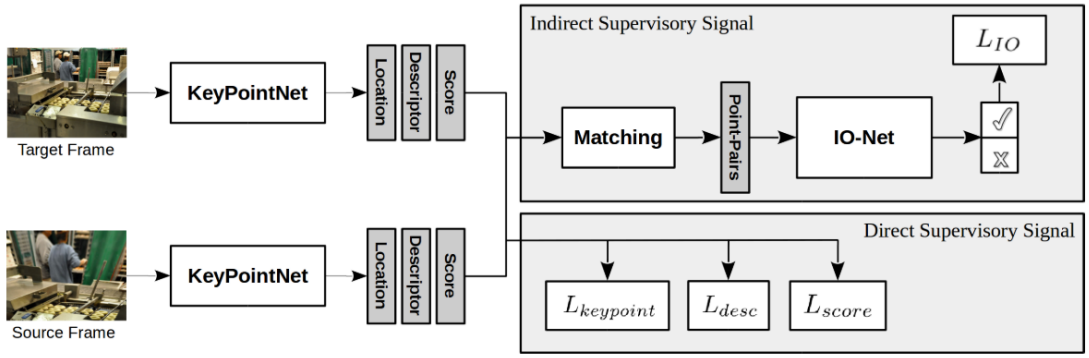


Figure 3.4 : The block diagram of KP2D [7]

The KeypointNet was meticulously crafted as a sophisticated encoder-decoder network, ingeniously engineered to tackle the challenges of feature extraction. Its architecture is meticulously designed, featuring an encoder section and a decoder section, each contributing uniquely to the network's remarkable capabilities.

The encoder segment of KeypointNet is composed of four VGG-type blocks, Figure 3.5. These blocks effectively process the input image, systematically reducing its spatial dimensions by a factor of eight. Through a series of convolutional operations,

pooling layers, and non-linear activations, the encoder adeptly compresses the image representation while preserving salient features necessary for subsequent analysis.

Conversely, the decoder portion of KeypointNet incorporates three distinct heads, each dedicated to producing specific outputs: keypoints, descriptors, and scores. The decoder branches operate in parallel, harnessing the information encoded in the compressed representation obtained from the encoder. This design choice enables the network to simultaneously extract keypoint locations, compute corresponding descriptors, and assign scores to the detected keypoints.

Of particular note is the network's remarkable capacity to generate keypoints for images of varying sizes. For an input image with dimensions (H, W), the KeypointNet network is capable of producing a total of (HxW)/64 keypoints. This flexibility ensures that the network can effectively handle images of diverse resolutions, adapting to the specific requirements of each scenario [7].

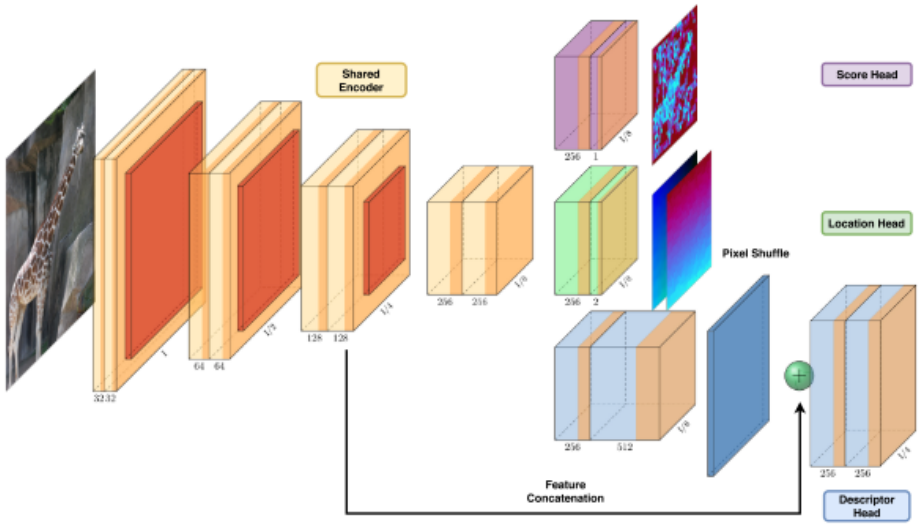


Figure 3.5 : The block diagram of KeypointNet [7]

3.2.2 Integration of the KeypointNet to VINS

In our proposed enhancement to the VINS-Mono pipeline, we introduce a significant modification by replacing the conventional feature detection component, specifically the Shi-Tomasi corner detector, with the state-of-the-art KeypointNet, Figure 3.6. This substitution aims to leverage the capabilities of deep learning models in

extracting informative keypoints directly from the sensor data, eliminating the need for handcrafted feature detection algorithms.

Given that our visual-inertial odometry (VIO) system operates at relatively high speeds while the flight speed is relatively low, we have adopted an optical flow-based approach for feature matching instead of relying on computationally expensive descriptor matching techniques. This strategic decision allows us to efficiently track and match features across consecutive frames, mitigating the computational burden associated with descriptor-based matching.

Furthermore, since our primary focus is on the accurate estimation of the relative pose, we do not utilize the descriptor vectors outputted by the KeypointNet. Instead, we leverage the well-established Kanade-Lucas-Tomasi (KLT) [37] algorithm as a reliable feature tracker throughout the frames. The KLT algorithm has a proven track record in visual tracking tasks and serves as an effective tool in our enhanced VINS-Mono pipeline.

By integrating KeypointNet for feature extraction and employing optical flow-based matching alongside the KLT tracker, we aim to achieve a more robust and efficient VIO system. This modification allows us to leverage the strengths of deep learning in extracting informative keypoints while capitalizing on established feature tracking techniques to maintain accurate and reliable pose estimation.

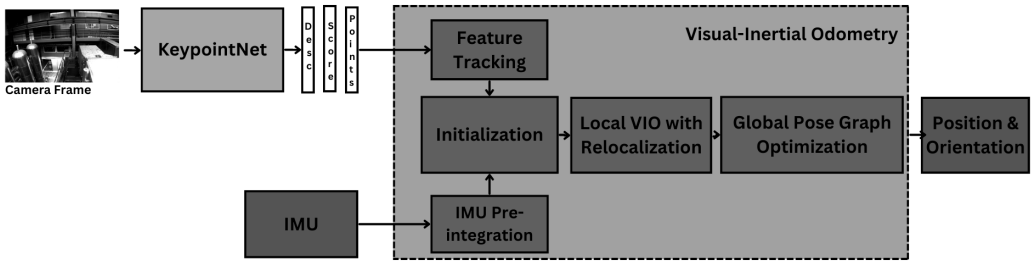


Figure 3.6 : The overview of the proposed method.

3.2.3 Training

We adopted the same loss functions used in the original paper. There are three different loss functions to effectively learn keypoint detection, descriptor calculation, and score computation. We also employ an additional loss function to learn to separate the extracted points into inlier and outlier sets.

Detector learning is conducted via keypoint location consistency loss. For this, we transform estimated salient points from the source image to the target image via a known homography matrix. Then we calculate the Euclidean distance between the transformed point location and the corresponding keypoints in the target frame. The difference should be less than the predefined threshold value. This self-supervised loss function enables the network to learn to extract rich useful points [7].

$$L_{loc} = \sum_{i} ||p_t^* - \hat{p}_t||_2 \tag{3.1}$$

$p_t^* = [u_i^*, v_i^*]$ is the warped point location via known Homography matrix and $\hat{p}_t = [\hat{u}_t, \hat{v}_t]$ is the associated keypoint in the target frame [7].

The KeypointNet also learns to compute unique descriptors for extracted keypoints. In order to catch finer details in the given image, the network employs a fast upsampling step before outputting the descriptor vectors. A triplet loss function is designed for descriptor learning [7].

$$L_{desc} = \sum_{i} max(0, ||f_i, f_{i,+}^*||_2 - ||f_i, f_{i,-}^*||_2 + m) \tag{3.2}$$

$f_{i,+}^*$ and $f_{i,-}^*$ represent positive and negtive samples. The goal is to minimize the distance between f_i and positive sample and maximize the distance between f_i and negative sample. m represents the distance that dissimilar descriptors should be pushed away from each other [7].

The third head of the KeypointNet network estimates score values for each calculated keypoint. The score value indicates how qualitative the extracted keypoints are. The

most reliable points should get high score values, while unreliable ones should get lower score values. To enable the network to learn to score, the following loss function was designed.

$$L_{score} = \sum_i \left[\frac{(s_i + \hat{s}_i)}{2} . (d(p_i, \hat{p}_i) - \overline{d}) + (s_i - \hat{s}_i)^2 \right] \tag{3.3}$$

Here, s_i and $\hat{s}_i$ represent scores in the source and target images, respectively. $\overline{d}$ corresponds to the average reprojection error of points in the current frame [7].

As I mentioned above, the KeypointNet already provides an inlier set of keypoints. Therefore, there is no need to apply outlier detection algorithms, like RANSAC after the feature-matching process. This is achieved by introducing a network, called IO-Net, and a new loss function, which will also supervise the descriptor learning phase. The IO-Net accepts keypoint pairs and the corresponding Euclidean distance between their descriptor vectors as input and outputs a probability of the points being inliers.

$$L_{IO} = \sum_i \frac{1}{2} (r_i - sign(||p_i^* - \hat{p}_i||_2 - \varepsilon_{uv}))^2 \tag{3.4}$$

Here, r is the probability of the keypoints being inliers or outliers calculated by the IO-Net. The IO-Net is only used to supervise the outlier rejection task during the training and it is ignored on the inference time [7].

Finally, we obtain an overall loss function by taking the weighted sum of the individual loss functions as follows,

$$\mathscr{L} = \alpha L_{loc} + \beta L_{desc} + \lambda L_{score} + L_{IO} \tag{3.5}$$

After some experimentation, we decided the weighting values to be as follows,

Table 3.1 : Loss weighting values.

α	β	λ
1.0	2.0	1.0

We adopted the weights of the model pre-trained with COCO dataset [38] and finetune it with a dataset from our domain. The original training was done with ADAM [39] optimizer. The initial learning rate was chosen as 1e-3 and it was halved after 40 epochs of training. In total, 50 epochs of training were conducted with a batch size of 8. Finally, we finetuned the network with the EuRoC MAV dataset [40] for about 50 epochs with a batch size of 8 and a learning rate of 1e-4. The other training hyperparameters were chosen as it was shown in the paper, Table 3.1.

3.3 Results

In our research, we conducted thorough testing and evaluation using the EuRoC MAV dataset [40] to validate the effectiveness of our proposed method, Figures 3.7 and 3.8. Since the dataset itself is relatively small for training a deep learning model, we adopted a specific approach. We chose to train our algorithm with the majority of the dataset and reserved one sequence for testing purposes. This strategy allowed us to assess the generalization and performance of our method effectively.

The primary focus of this thesis is to enhance the speed of state-of-the-art Visual-Inertial Navigation System (VINS) methods. In this section, we present compelling evidence that highlights the superior speed performance of our proposed approach. To conduct a comprehensive comparison, we benchmarked the execution times of our method, which incorporates KeypointNet, against other existing methods. For our evaluations, we utilized the EuRoC MAV dataset, employing a frame size of (480, 640) and extracting 300 keypoints.

Our method demonstrated exceptional efficiency in the extraction of salient points, achieving an impressive rate of 184 frames per second (fps) as it is shown in Table 3.2. This exceptional speed result surpassed all other methods that we evaluated, solidifying the superiority of our proposed approach in terms of speed within the context of VINS. This notable advantage highlights the potential impact of our method on real-time applications, where fast and efficient processing is crucial for autonomous systems and robotics.

By achieving such a remarkable speed, our proposed method addresses a significant limitation of existing VINS algorithms and provides a practical solution for real-time applications. The enhanced speed performance not only improves the overall efficiency of the VINS system but also opens up possibilities for time-critical scenarios and applications that demand rapid and accurate navigation and localization.

The comprehensive comparison conducted with the EuRoC MAV dataset, along with the impressive speed performance exhibited by our method, underscores the significance and effectiveness of our proposed approach. The results obtained in our research provide strong evidence of the advantages offered by integrating KeypointNet into the VINS pipeline, further reinforcing the potential impact of our method in advancing the field of visual-inertial navigation systems.

Table 3.2 : Execution times of KeypointNet and other detectors for 300 keypoints in frames per second(FPS). * Numbers from [1]

Detectors	FPS @ 480x640
ORB*	33
SIFT*	12
SURF*	10
AKAZE*	18
BRISK*	5
SuperPoint*	67
LF-Net*	25
UnsuperPoint*	65
Shi-Tomasi (Original)	160
KeypointNet (KP2D)	**184**

In addition to assessing the speed performance of our proposed method, we conducted a thorough comparison of its accuracy against existing approaches. To evaluate the accuracy, we examined the estimated trajectories in a 3D flight scenario using the EuRoC MAV dataset, a widely recognized benchmark for quadrotor navigation.

The trajectory plots Figure 3.9, Figure 3.13, and Figure 3.14 generated from our method clearly illustrate its capability to achieve accurate results in the challenging flight conditions of the EuRoC MAV dataset. We meticulously analyzed the estimated trajectories, paying particular attention to key parameters such as altitude and y-axis errors.

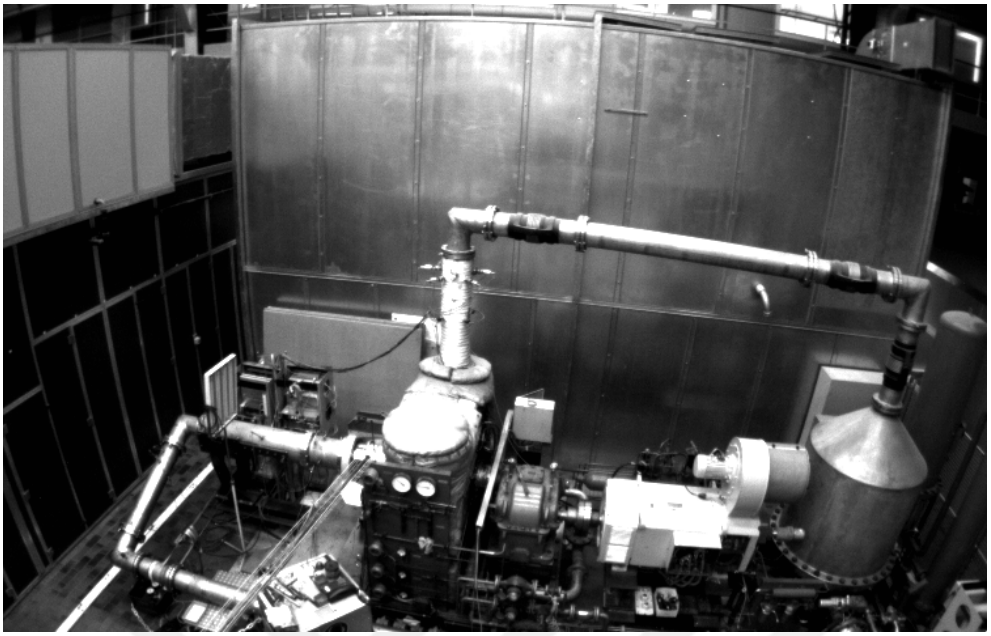


Figure 3.7 : Example scene from EuRoC MAV dataset

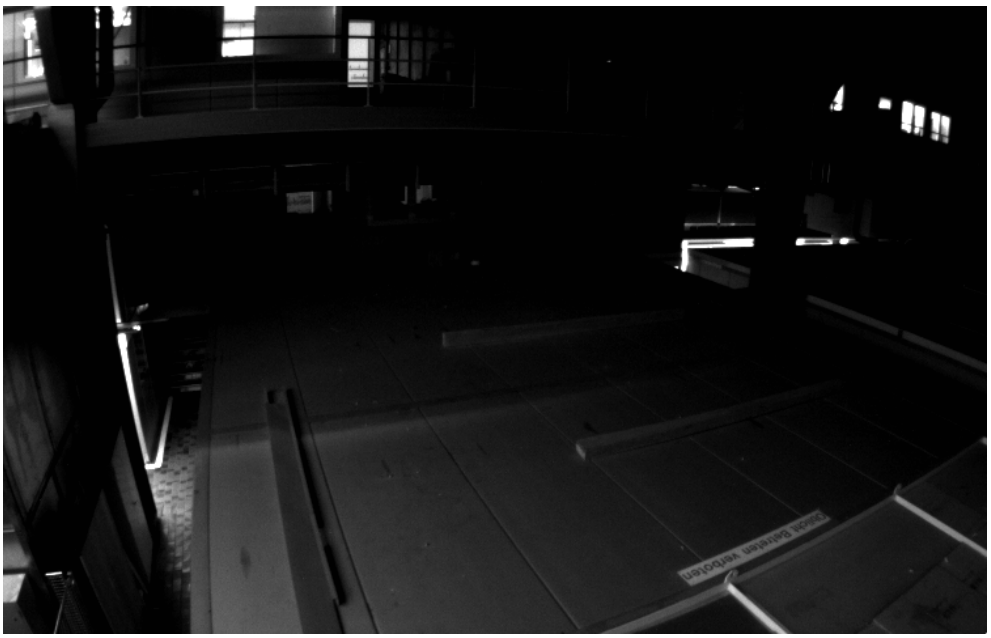


Figure 3.8 : Example scene from EuRoC MAV dataset

Our method exhibited highly satisfactory results, as the altitude and y-axis errors consistently remained within the range of ±20cm. This outcome represents a notable achievement, as it demonstrates the ability of our proposed method to achieve precise and reliable pose estimation in complex flight scenarios. You can see the results in Figure 3.10 and Figure 3.11.

The accurate trajectory estimation achieved by our method is of significant importance in real-world applications where reliable navigation and localization are essential. The acceptable error range observed in the altitude and y-axis provides confidence in the robustness and accuracy of our proposed approach.

These findings validate the effectiveness of our enhanced VINS method, as it successfully addresses the challenges presented by the EuRoC MAV dataset and delivers accurate pose estimation. By achieving such reliable results, our method offers a solid foundation for various applications in autonomous flight, robotics, and augmented reality, where precise trajectory estimation is crucial for safe and efficient operation.

Based on the presented plots Figure 3.12, it is evident that the proposed method demonstrates comparable performance to the original VINS-Mono algorithm. In fact, in certain instances, our method exhibits superior performance. This noteworthy outcome substantiates the successful validation of our approach through rigorous testing, particularly on the most demanding sequence of the challenging quadrotor benchmark. The comprehensive assessment of our method's performance, as depicted on the plots, reinforces its effectiveness and positions it as a viable alternative to the established VINS-Mono algorithm. By achieving comparable or even superior results, our method showcases its potential for advancing the field and addressing the complexities inherent in the quadrotor benchmark.

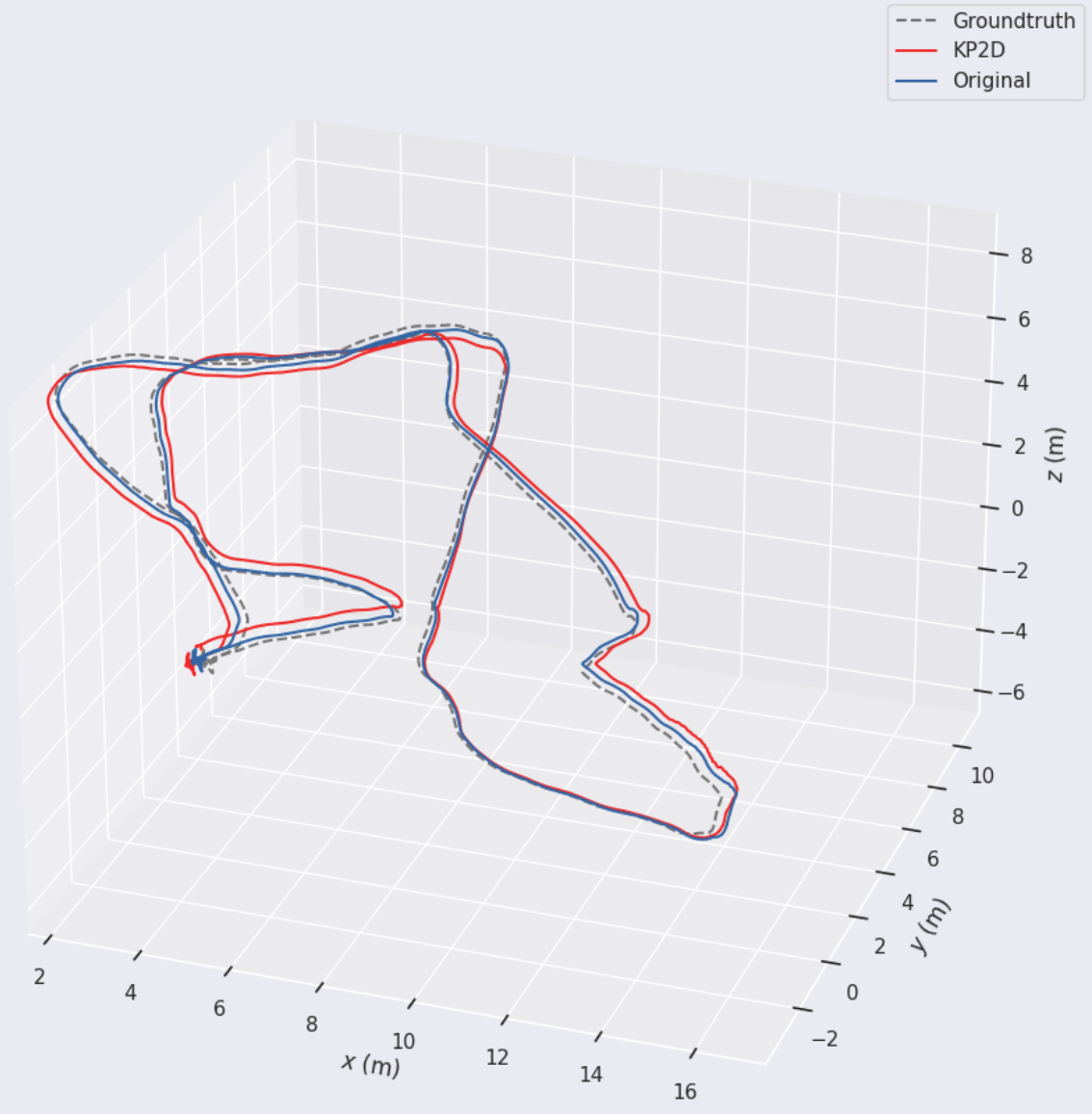


Figure 3.9 : 3D Plot of the trajectories obtained with KP2D and original point detector of VINS-Mono on the Euroc MAV MH 05 difficult dataset.

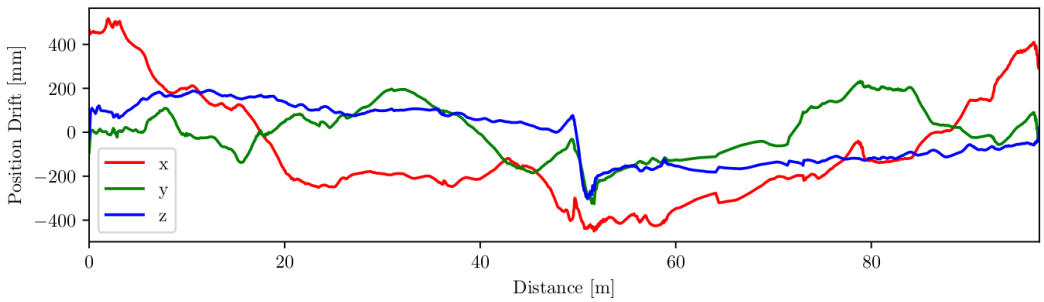


Figure 3.10 : Translation error of KeypointNet augmented VIO on the Euroc MAV MH 05 difficult dataset.

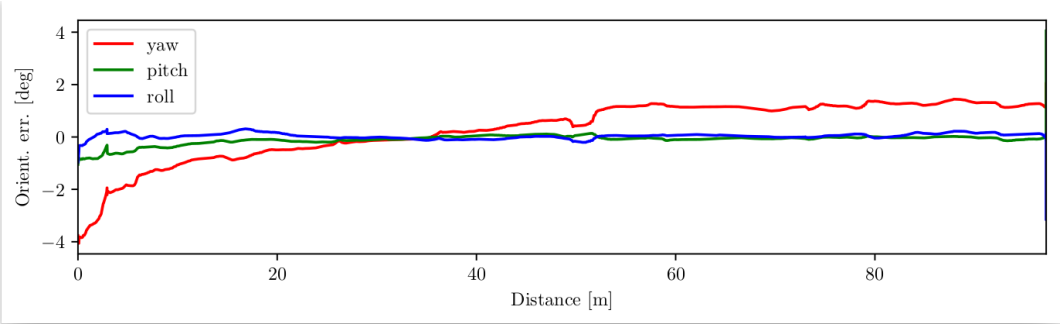


Figure 3.11 : Rotation error of KeypointNet augmented VIO on the Euroc MAV MH 05 difficult dataset.

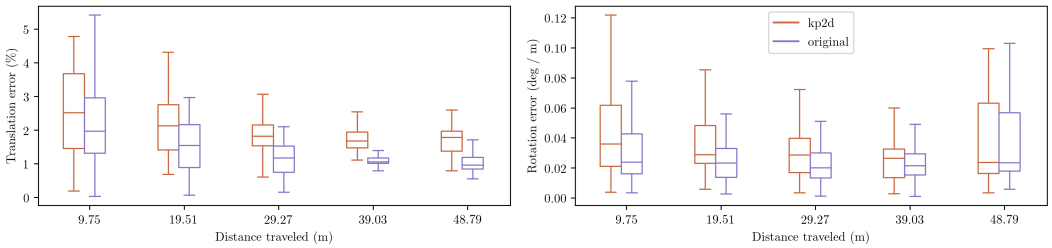


Figure 3.12 : Comparison of the translation and rotation error of the proposed and original methods on the Euroc MAV MH 05 difficult dataset.

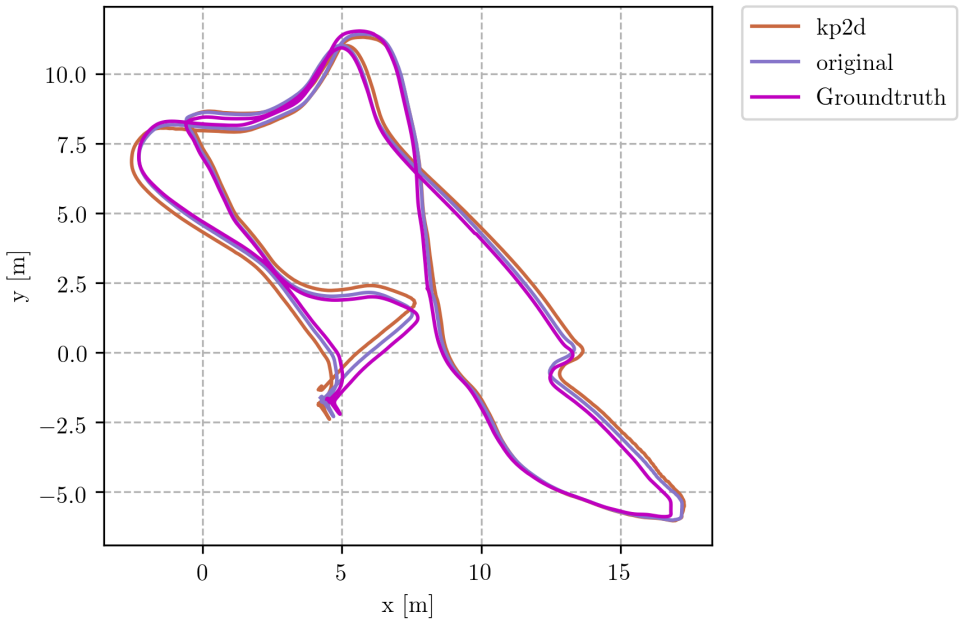


Figure 3.13 : Top view of the trajectories obtained with KP2D and original point detector of VINS-Mono on the Euroc MAV MH 05 difficult dataset.

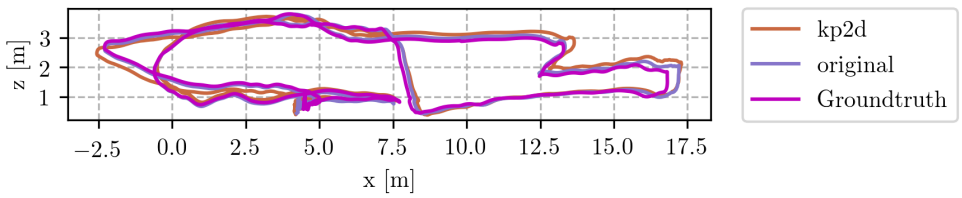


Figure 3.14 : Side view of the trajectories obtained with KP2D and original point detector of VINS-Mono on the Euroc MAV MH 05 difficult dataset.

4. CONCLUSIONS

Self-localization in GPS-denied environments is an exceptionally complex task that poses significant challenges. This difficulty is further amplified when it involves three-dimensional (3D) flight in outdoor settings where high-resolution maps are not readily available. Effectively addressing these constraints necessitates the development of a rapid, lightweight, highly accurate, and robust algorithm capable of estimating the agent's states by efficiently leveraging visual and inertial information from onboard sensors.

While classical Visual-Inertial Odometry (VIO) methods can partially fulfill some of these requirements and function adequately in indoor conditions, their performance tends to deteriorate or become error-prone when confronted with the demanding conditions of outdoor environments. Given the paramount importance of safety during 3D outdoor flight, it becomes imperative for our techniques to yield exceptionally precise calculations in order to avoid any potential accidents resulting from erroneous state estimations.

Despite the suitability of classical methods for indoor environments, they encounter significant challenges when confronted with high dynamic range, featureless scenes, and motion blur. To overcome these limitations, fully end-to-end deep learning-based approaches have emerged as a promising solution. These algorithms leverage raw visual and inertial measurements as input and generate precise 6-DoF (six degrees of freedom) relative pose estimations as output.

In these deep learning-based methods, features are extracted using convolutional neural networks (CNNs), which excel at handling complex image-based tasks. By employing CNNs for feature regression, the algorithms can effectively address the intricacies associated with processing visual information. However, when it comes to incorporating inertial measurements, the use of LSTM-like components is unnecessary. This is due to the fact that linear acceleration and axial velocity, which constitute

inertial measurements, are low-dimensional data that can be efficiently handled using well-established physical equations.

The visual-inertial ego-motion estimation field has recently witnessed a notable trend in combining the strengths of both classical VIO methods and deep learning approaches. This integration aims to leverage the impressive function regression capabilities of deep learning, irrespective of complexity, while benefiting from the lightweight and robust nature of classical state estimation techniques. By merging these two paradigms, a lightweight, accurate, and robust estimation algorithm can be achieved.

In this thesis, we embark on demonstrating the superiority of hybrid methods over the traditional state-of-the-art VIO techniques in terms of speed, while maintaining a comparable level of accuracy. Our approach entails the utilization of a convolutional network to identify the most salient points within a given frame. These salient points are then fed into the motion estimation block of the renowned classical VIO algorithm. Through this fusion, we capitalize on the deep learning model's ability to efficiently extract relevant visual features, while relying on the robustness of the classical VIO algorithm for motion estimation.

To validate the effectiveness of our method, we conduct extensive experiments using the EuRoC MAV dataset, which serves as a benchmark in the field. We meticulously compare the results obtained from our hybrid approach with those derived from other classical and hybrid methods. By doing so, we aim to establish the superiority of our proposed method, highlighting its advantages in terms of both speed and accuracy.

4.1 Practical Application of This Study

The field of autonomous flight or driving is rapidly expanding, and it is poised to become increasingly crucial in the coming years, Figure 4.1. Autonomous systems heavily rely on accurate and real-time ego-motion estimation to operate effectively. However, performing ego-motion estimation in outdoor environments poses significant challenges. Therefore, it becomes imperative to develop robust navigation systems that can fulfill these requirements.

In light of these challenges, this study holds particular significance in the context of 3D autonomous flight and 2D autonomous driving within GPS-denied environments. By addressing the intricacies of ego-motion estimation in such challenging conditions, the study aims to provide valuable insights and practical solutions that can enhance the performance and reliability of autonomous systems.

The successful development of a reliable navigation system capable of accurately estimating ego-motion in real-time is crucial for ensuring the safe and efficient operation of autonomous vehicles and drones. By overcoming the limitations imposed by the absence of GPS signals in challenging outdoor environments, this study contributes to the advancement of autonomous technologies, paving the way for their widespread adoption in various industries and applications.

Figure 4.1 : Autonomous driving and flight.

REFERENCES

[1] **Christiansen, P.H.**, **Kragh, M.F.**, **Brodskiy, Y. and Karstoft, H.** (2019). UnsuperPoint: End-to-end Unsupervised Interest Point Detector and Descriptor, http://arxiv.org/abs/1907.04011.

[2] **Leonardis, A.**, **Bischof, H.**, **Pinz, A.**, **Bay, H.**, **Tuytelaars, T. and Van Gool, L.** (2006). Computer Vision – ECCV 2006 SURF: Speeded Up Robust Features, *Computer Vision – ECCV 2006*, *3951*, 404–417, http://www.springerlink.com/content/e580h2k58434p02k/.

[3] **Davim, J.P.** (2011). *Machining of hard materials*, January.

[4] **Gonzalez, T.F.** (2007). Handbook of approximation algorithms and metaheuristics, *Handbook of Approximation Algorithms and Metaheuristics*, 1–1432.

[5] **Wang, S.**, **Clark, R.**, **Wen, H. and Trigoni, N.** (2017). DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks, *Proceedings - IEEE International Conference on Robotics and Automation*, 2043–2050.

[6] **Qin, T.**, **Li, P. and Shen, S.** (2018). VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator, *IEEE Transactions on Robotics*, *34*(4), 1004–1020.

[7] **Tang, J.**, **Kim, H.**, **Guizilini, V.**, **Pillai, S. and Ambrus, R.** (2019). Neural Outlier Rejection for Self-Supervised Keypoint Learning, (2019), 1–14, http://arxiv.org/abs/1912.10615.

[8] **Masters, T.** (1993). Multilayer Feedforward Networks, *Practical Neural Network Recipies in C++*, 77–116.

[9] **Cleary, J.D.** (1997). *Introduction to visual observations*, volume 31.

[10] **Low, D.G.** (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, 91–110.

[11] **Rublee, E.**, **Rabaud, V.**, **Konolige, K. and Bradski, G.** (2011). ORB: An efficient alternative to SIFT or SURF, *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571.

[12] **Shi, J. and Tomasi, C.** (1994). Good Features, *Image (Rochester, N.Y.)*, 593–600.

[13] **Trajković, M. and Hedley, M.** (1998). Fast corner detection, *Image and Vision Computing*, *16*(2), 75–87.

[14] **Calonder, M.**, **Lepetit, V.**, **Strecha, C. and Fua, P.** (2010). BRIEF: Binary robust independent elementary features, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6314 LNCS*(PART 4), 778–792.

[15] **Mur-Artal, R.**, **Montiel, J.M. and Tardos, J.D.** (2015). ORB-SLAM: A Versatile and Accurate Monocular SLAM System, *IEEE Transactions on Robotics*, *31*(5), 1147–1163.

[16] **Scaramuzza, D. and Fraundorfer, F.** (2011). Visual Odometry [Tutorial], (December).

[17] **Fraundorfer, F. and Scaramuzza, D.** (2012). Visual odometry: Part II: Matching, robustness, optimization, and applications, *IEEE Robotics and Automation Magazine*, *19*(2), 78–90.

[18] **Davison, A.J.**, **Reid, I.D.**, **Molton, N.D. and Stasse, O.** (2007). MonoSLAM: Real-time single camera SLAM, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1052–1067.

[19] **Klein, G. and Murray, D.** (2007). Parallel tracking and mapping for small AR workspaces, *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR*, 225–234.

[20] **Newcombe, R.A.**, **Lovegrove, S.J. and Davison, A.J.** (2011). DTAM: Dense tracking and mapping in real-time, *Proceedings of the IEEE International Conference on Computer Vision*, 2320–2327.

[21] **Engel, J.**, **Sturm, J. and Cremers, D.** (2013). LSD-SLAM: Large-Scale Direct Monocular SLAM, *Proceedings of the IEEE International Conference on Computer Vision*, 1449–1456.

[22] **Engel, J.**, **Koltun, V. and Cremers, D.** (2016). DSO Journal: Direct Sparse Odometry.

[23] **Takahashi, T.** (2007). 2D localization of outdoor mobile robots using 3D laser range data, *Robotics*, (May).

[24] **Marchal, A.** (2006). Keyframe-Based Visual-Inertial SLAM Using Nonlinear Optimization, (september), 1–30.

[25] **Mourikis, A.I. and Roumeliotis, S.I.** (2007). A multi-state constraint Kalman filter for vision-aided inertial navigation, *Proceedings - IEEE International Conference on Robotics and Automation*, 3565–3572.

[26] **Xue, F.**, **Wang, X.**, **Li, S.**, **Wang, Q.**, **Wang, J. and Zha, H.** (2019). Beyond tracking: Selecting memory and refining poses for deep visual odometry, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 8567–8575.

[27] **Clark, R.**, **Wang, S.**, **Wen, H.**, **Markham, A. and Trigoni, N.** (2017). VINet: Visual-inertial odometry as a sequence-to-sequence learning problem, *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 3995–4001.

[28] **Clark, R.**, **Bloesch, M.**, **Czarnowski, J.**, **Leutenegger, S. and Davison, A.J.** (2018). Learning to solve nonlinear least squares for monocular stereo, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11212 LNCS*(1), 291–306.

[29] **Karkus, P.**, **Hsu, D. and Lee, W.S.** (2018). Particle Filter Networks with Application to Visual Localization, (CoRL), 1–10, `http://arxiv.org/abs/1805.08975`.

[30] **Li, C. and Waslander, S.L.** (2020). Towards End-to-end Learning of Visual Inertial Odometry with an EKF, *Proceedings - 2020 17th Conference on Computer and Robot Vision, CRV 2020*, 190–197.

[31] **Yang, N.**, **Von Stumberg, L.**, **Wang, R. and Cremers, D.** (2020). D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1278–1289.

[32] **Han, X.**, **Tao, Y.**, **Li, Z.**, **Cen, R. and Xue, F.** (2020). SuperPointVO: A Lightweight Visual Odometry based on CNN Feature Extraction, *Proceedings - 5th International Conference on Automation, Control and Robotics Engineering, CACRE 2020*, 685–691.

[33] **Detone, D.**, **Malisiewicz, T. and Rabinovich, A.** (2018). SuperPoint: Self-supervised interest point detection and description, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2018-June*, 337–349.

[34] **Bruno, H.M.S. and Colombini, E.L.** (2021). LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method, *Neurocomputing*, *455*, 97–110.

[35] **Yi, K.M.**, **Trulls, E.**, **Lepetit, V. and Fua, P.** LIFT : Learned Invariant Feature Transform.

[36] **Fischler, M.A. and Bolles, R.C.** (1981). Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Communications of the ACM*, *24*(6), 381–395.

[37] **Zhao, M.**, **Zhang, J. and Figueiredo, R.** (2004). *Distributed file system support for Virtual Machines in Grid computing*.

[38] **Lin, T.Y.**, **Maire, M.**, **Belongie, S.**, **Hays, J.**, **Perona, P.**, **Ramanan, D.**, **Dollár, P. and Zitnick, C.L.** (2014). Microsoft COCO: Common objects in context, *Lecture Notes in Computer Science (including subseries Lecture*

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *8693 LNCS*(PART 5), 740–755.

[39] **Kingma, D.P. and Ba, J.L.** (2015). Adam: A method for stochastic optimization, *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

[40] **Burri, M.**, **Nikolic, J.**, **Gohl, P.**, **Schneider, T.**, **Rehder, J.**, **Omari, S.**, **Achtelik, M.W. and Siegwart, R.** (2016). The EuRoC micro aerial vehicle datasets, *International Journal of Robotics Research*, *35*(10), 1157–1163.

APPENDICES

CURRICULUM VITAE

Arslan ARTYKOV:

EDUCATION:

- **B.Sc.:** 2020, Istanbul Technical University, Aeronautics and Astronautics Engineering Faculty, Aeronautical Engineering Department
- **M.Sc.:** 2023, Istanbul Technical University, Aeronautics and Astronautics Engineering Faculty, Aeronautics and Astronautics Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2021-2022 İstanbul Technical University at the Aerospace Research Center.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Artykov, A.** and Koyuncu, E. (2023). Deep Learning-Based Keypoints Driven Visual-Inertial Odometry For GNSS-Denied Flight, *12th Ankara International Aerospace Conference*.