Localização aérea por descritores locais de imagem baseados em inteligência artificial.

Bruno Nardi de Carvalho Dantas[1], Elcio Hideiti Shiguemori[2] e Lamartine Nogueira Frutuoso Guimarães[2]
[1]Instituto Tecnológico de Aeronáutica , São José dos campos/SP - Brasil
[2]Instituto de Estudos Avançados, São José dos campos/SP - Brasil
dantasbncd@ita.br

***Resumo*— Visando uma alternativa robusta ao sistema global de navegação por satélite (GNSS), foram propostas técnicas baseadas em redes neurais convolucionais para a tarefa de localização por imagens através da comparação entre cenas aéreas e satelitais. Porém, a maioria dessas abordagens exige treinamento da rede antes de um novo voo. Assim, este trabalho comparou os descritores de imagens *LOFTr*, *SuperGlue* e *D2NET*, baseados em redes neurais, treinadas somente em mapas de profundidade, com os descritores: *SIFT*, *ORB* e *AKAZE*. Foram utilizadas imagens de câmeras acopladas a aeronaves com voo a 80 metros de altura e os descritores geraram pontos correspondentes ao mesmo local em imagens satelitais. Após esse passo, foi estimada a homografia e selecionados os pontos comuns (homólogos) utilizando o algoritmo *RANSAC*. Os descritores baseados em rede neural e mecanismo de atenção apresentaram um número de pontos homólogos muito superior aos descritores tradicionais, além de obter uma estimativa com erro médio inferior a um metro em relação ao método manual.**

***Palavras-Chave*— Redes Neurais, Descritores locais de imagem, Imagens Aéreas.**

I. Introdução

Na busca de uma navegação aérea mais segura, a visão computacional pode compensar as desvantagens apresentadas por sistemas de posicionamento global e sistemas inerciais. O Sistema Global de Navegação por Satélite (GNSS) pode ter o sinal interrompido por fatores técnicos, naturais ou interferências humanas intencionais[1], através de técnicas como o *spoofing* e o *jamming*. Os sistemas inerciais presentes em sistemas aéreos de navegação podem apresentar acúmulo de erro ao longo da trajetória, pois trabalham apenas com a diferença entre medições ao longo do tempo, ou seja, não possuem um referencial absoluto de cálculo. Além de agir como sistema alternativo de localização, a visão computacional pode fornecer informações importantes para a tomada de decisão em ambientes dinâmicos onde operam os veículos autônomos, como visto nos carros elétricos [2]. Os sistemas de localização baseados em imagens podem ser absolutos ou relativos [1]. Ao medir as diferenças entre quadros para estimar o deslocamento e posição do veículo, os sistemas relativos estimam a odometria por meio visual, mas assim como o sistema inercial, podem acumular erros ao longo do tempo. Em contrapartida, os sistemas de localização absolutos possuem bancos de imagens do local em que o sistema de navegação está operando. Desta forma, podemos extrair a posição da aeronave ao relacionar as imagens aéreas captadas por câmeras com imagens satelitais do mesmo local [3]. Os estudos recentes na área de localização absoluta usam três principais abordagens: casamento de padrões de forma direta (*template matching*), uso de descritores locais de imagens (*feature point matching*) e uso de redes de aprendizado profundo (*deep learning*)[1]. Podemos destacar as redes neurais convolucionais que obtêm a segmentação semântica da cena aérea para o casamento das regiões semânticas idênticas em busca da localização da aeronave [4], [5]. Um problema com esse método é a necessidade do uso de grandes bancos de dados para treinamento supervisionado para classificação a nível de pixel para a segmentação semântica. Adicionalmente podemos citar o uso da rede *Autoencoder* [6] para criar uma representação em um espaço reduzido (espaço latente) de um conjunto de imagens georreferenciadas para localização futura da aeronave. Essa abordagem requer um treinamento prévio da rede *Autoencoder* antes de cada voo. Por último, a localização aérea pode ser estimada a partir de imagens satelitais com auxílio dos novos descritores locais de imagens que utilizam as redes neurais convolucionais treinadas previamente com as imagens do local sobrevoado.

A. Lacuna de pesquisa

Esse trabalho busca analisar o uso dos descritores locais automáticos de imagem que utilizam camadas convolucionais para extrair características da imagem e mecanismos de atenção para estimar a relação geométrica entre imagens aéreas e satelitais. Não foram encontrados muitos estudos na área indicando se esses novos descritores apresentam resultados satisfatórios, sem o treinamento prévio da rede neural na região de voo, ao serem utilizados em imagens aéreas e satelitais de um mesmo local.

B. Trabalhos relacionados

Durante a captação de imagens por uma aeronave, seu movimento pode causar rotações, mudanças na escala e no ângulo de visada entre a imagem aérea e a imagem satelital de referência. Comparando diversos descritores locais de imagens aplicados em imagens distorcidas e/ou com transformações geométricas, como rotações e mudança de escala, Karami *et. al.*[7] destaca a criação do algoritmo *SIFT* [8] por Lowe em 2004, o qual revolucionou as aplicações de visão computacional e robótica. Contudo, *SIFT* apresenta alta complexidade computacional. Karami destaca o algoritmo *ORB* [9] como tendo um desempenho similar ao *SIFT* e apresentando maior eficiência computacional. Esse descritor usa como detector de pontos o algoritmo FAST [10] e como descritor o algoritmo BRIEF [11]. Além disso, Ross *et. al.* [12] compararam o descritor *AKAZE* [13] com *ORB* e obtiveram resultados promissores no registro entre imagens aéreas de baixa resolução obtidas por drones para a tarefa de odometria

visual com o algoritmo *AKAZE*. Assim, para realizar uma comparação entre os descritores mais utilizados no campo da visão computacional e os novos descritores baseados em redes neurais, este trabalho elegeu os descritores *SIFT*, *ORB* e *AKAZE* para representar os descritores mais tradicionais. Com os recentes avanços na pesquisa no campo da inteligência artificial a partir do uso de camadas convolucionais nas redes neurais profundas, Daneil *et. al.* [14] desenvolveu o descritor de imagens denominado *SuperPoint* através do treino auto supervisionado das camadas convolucionais e a partir do processo intitulado adaptação por homografia, alcançando boa repetibilidade em imagens em alta resolução. Dusmanu *et. al.*[15] propuseram o algoritmo *D2-NET*, como descritor automático de imagens, que extrai mapas de caraterísticas de imagens nas camadas neurais convolucionais mais profundas e retira simultaneamente os detectores e descritores para gerar os pontos correspondentes entre as imagens, além disso, essa técnica pode ser aprimorada através do treinamento em banco de imagens específicos. Com essa nova técnica, Hou et.al. [16] utilizaram a rede D2-NET, treinada com imagens da região, para a localização de uma pequena aeronave a partir de imagens de satélites com voos a cerca de 0,5 quilômetros de altura e obteve um erro médio de apenas 15 metros em um percurso de cerca de 0,75 Km. Vaswani et al. [17], propuseram uma arquitetura inovadora de redes neurais com o uso de mecanismos de atenção, a rede *Transformers*. Apresentou-se uma nova forma de codificação da informação através de funções senoidais, formação de estruturas baseadas na teoria de banco de dados(busca, chaves e valor) e introdução das camadas de *self-attention* e *cross-attention*. Com essa nova arquitetura, Sarlin *et. al.*[18] propuseram a rede *Superglue*, que utiliza esses mecanismos de atenção para aprimorar o casamento entre pontos estimados por descritores e Sun *et. al.*[19] apresentam um novo método para encontrar pontos característicos em comum entre duas imagens baseado em mecanismos de atenção e com a seguinte estratégia: Primeiramente, o algoritmo extrai as características de pequenas áreas das imagens através de camadas convolucionais; Aplica o mecanismo de atenção para identificar a maior similaridade entre essas pequenas áreas, tanto em relação à própria imagem através do mecanismo *self-attention*, quanto em relação a imagem a ser comparada utilizando a porção chamada *cross-attention*; Finalmente, avalia novamente a correlação entre pontos característicos das imagens nas camadas convolucionais mais profundas de uma rede neural nomeada FCN[20] levando em consideração a informação obtida pelos passos anteriores para obter uma estimação de pontos mais precisa. Essa rede foi nomeada como *LoFTR* e, segundo os autores, obteve resultados superiores à rede D2-NET na tarefa de estimação de homografia em testes no banco de dados denominado *HPatches* [21] com a filtragem dos pontos obtidos pelo algoritmo *RANSAC* (*RANdom SAmple Consensus*)[22].

C. Contribuição esperada

Esse trabalho analisa três descritores empregados em visão computacional: *ORB*, *SIFT* e *AKAZE* e três métodos que utilizam inteligência artificial: D2-NET, *SuperGlue*(ao refinar a estimação obtida pelo SuperPoint) e *LOFTr*. Pretendeu-se analisar a performance desses descritores, levantando o número de pontos em comum gerados entre imagens aéreas e imagens obtidas por satélites após diversas transformações geométricas que são normalmente encontradas no cenário aéreo, como rotação e escala. Por último, foram eleitos pontos de interesse nas imagens aéreas de forma manual e foi medido a distância em metros na imagem satelital entre o resultado obtido pelo processo manual e a estimação realizada com auxílio dos descritores estudados. Como contribuição, espera-se que a pesquisa auxilie na construção de sistemas de localização por imagens em veículos aéreos.

II. Materiais e métodos

Visando-se o uso de descritores para aplicação imediata em imagens aéreas, ou seja, sem o treinamento prévio a cada novo local de voo, as redes neurais D2-NET, *SuperGlue* e *LOFTr* foram treinadas somente com o banco de dados denominado Mega Depth[23], o qual apresenta diversas imagens e seus correspondentes mapas de profundidade, assim como exemplificado na Fig.1. Todos os pesos do treinamento dessas redes foram publicamente disponibilizados pelos autores e nenhum novo treinamento foi executado nos experimentos descritos a seguir.

(a) (b)

Fig. 1. (a) Exemplo de imagem do banco de dados *MegaDepth*. (b) mapa de profundidade associado.

Para realizar a comparação entre os descritores, foi usado banco de dados contendo 30 imagens, onde 15 são oriundas de um voo a 80 metros de altura do solo e as outras 15 foram renderizadas no *Google Earth*^{TM}[24] para representar as imagens satelitais. A seleção das imagens procurou simular a complexidade do processo de busca em um banco de dados que retorna imagens com pequenas rotações em relação ao mesmo local, mudança de escala e diferentes ângulos de visada. Além disso, foi levado em consideração as características inerentes a cada fonte de imagem, ou seja, a disparidade temporal entre as imagens, a baixa resolução das imagens de satélites e a dinâmica natural de uma imagem aérea, como carros em locais diferentes e diferente iluminação ao longo do dia. Podemos observar algumas disparidades entre as imagens geradas pelos diferentes sensores na Fig. 2.

(a) (b)

Fig. 2. (a) Imagem aérea (b) Imagem satelital.

Para uma análise mais profunda entre os descritores, as imagens obtidas pelo veículo aéreo foram rotacionadas entre 0 a 170 graus e sua escala foi alterada de 30% a 100% do tamanho da imagem aérea original. Em cada par de imagens, levantou-se os pontos característicos com os descritores selecionados. Com o algoritmo de estimação robusta *RANSAC*, buscou-se o melhor modelo de homografia entre os pares de pontos obtidos em cada caso. Os pontos característicos locais que concordam com o modelo de homografia levantado são chamados de *inliers*. Foi levantada a quantidade de *inliers* obtidos por cada par de imagens e por cada descritor selecionado.

Como última etapa da metodologia, utilizou-se as imagens aéreas sem as transformações geométricas rígidas, como rotação e escala, e foram levantados pontos de interesse nas 15 imagens aéreas analisadas. Primeiramente, os pontos homólogos foram levantados manualmente entre as imagens obtidas por aeronaves(I_a) e as satelitais(I_s). Em seguida, descritores locais estudados estimaram diversos pontos homólogos correspondentes entre a imagem aérea e a imagem de referência (satelital) e através do algoritmo *RANSAC* foi obtida a homografia utilizada para o estimar a posição do ponto de interesse na imagem satelital (P_e) a partir do ponto elencado na imagem aérea (P_a). Posteriormente, foi calculada a distância euclidiana entre o ponto estimado pelos descritores (P_e) e o ponto obtido de maneira manual (P_m). Dessa distância, foi obtido o erro em pixels. A resolução espacial em cada imagem satelital foi obtida através da ferramenta de coleta de distância entre pontos disponibilizada no *Google Earth*^{TM}. Após o cálculo em pixeis, o erro foi multiplicado pela resolução espacial para a obtenção da medida em metros. A metodologia aqui aplicada foi diagramada na Fig.3 e o cálculo do erro em metros foi descrito no pseudocódigo1.

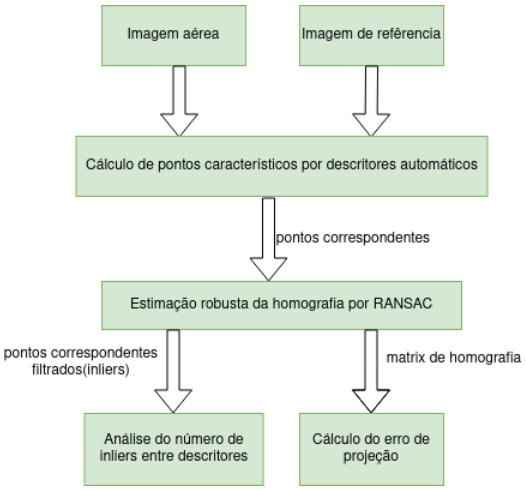


Fig. 3. Metodologia.

Algorithm 1 Cálculo do erro em metros

procedure ERRO_EM_METROS (I_a, I_s)
  $pontos\_homologos \leftarrow descritores(I_a, I_s)$
  $H \leftarrow RANSAC(pontos\_homologos)$
  $P_e \leftarrow H \cdot P_a$
  $erro\_em\_pixeis \leftarrow distancia\_euclidiana(P_e, P_m)$
  $erro\_em\_metros \leftarrow resolucao\_espacial \cdot erro\_em\_pixeis$
  return $erro\_em\_metros$
end procedure

III. Resultados

Ao comparar os descritores estudados para a geração de pontos homólogos entre as imagens aéreas e satelitais, foi observado que aqueles construídos com redes neurais profundas apresentaram um número de correspondências muito superior aos algoritmos convencionais aplicados em visão computacional, mesmo nas comparações com menor nível de complexidade, como demonstrado na Fig.4.

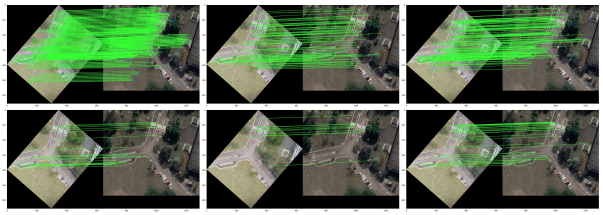


Fig. 4. Melhor caso entre descritores.

Em regiões onde a cena aérea está com pequenas translações em relação a imagem satelital ou quando há grande diferença na iluminação ou no ângulo de visada, devido a disparidade entre os sensores imageadores, os algoritmos que utilizam redes neurais convolucionais demonstraram uma robustez maior que os tradicionais. Em alguns casos, os descritores que não são baseados em inteligência artificial, falharam na estimação dos pontos homólogos, como ilustrado na Fig.5.

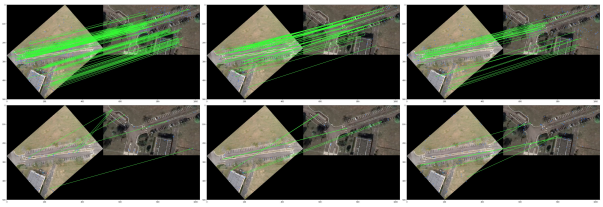
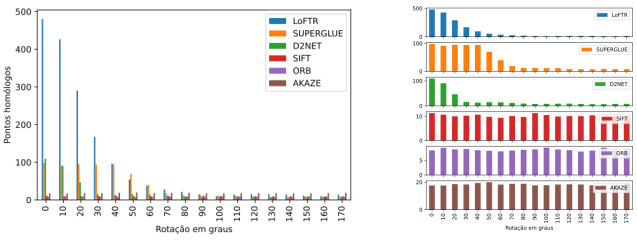


Fig. 5. Pior caso entre descritores.

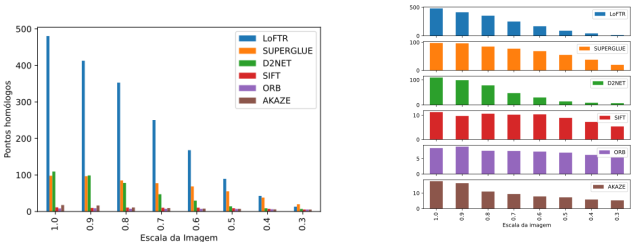
Ao analisar o impacto da rotação na imagem aérea em relação a imagem satelital, foi visto que todos os descritores baseados em inteligência artificial tiveram seu desempenho diminuído após essa transformação geométrica. Houve uma queda no número de pontos homólogos já com rotações inferiores a 90 graus. Os descritores tradicionais mantiveram relativamente o mesmo quantitativo de pontos em todas as rotações aplicadas, conforme constatado na Fig.6



(a) (b)

Fig. 6. (a) comparação entre descritores de acordo com o ângulo de rotação. (b) efeito da rotação em cada método. Observe que os métodos baseados em inteligência artificial(os três gráficos superiores) são sensíveis a rotação da imagem aérea.

Com a alteração na escala da imagem aérea em relação à imagem satelital, foi verificado que todos os descritores tiveram queda, apesar de mais suave, no número de pontos estimados, como visto na Fig7.



(a) (b)

Fig. 7. (a) comparação entre descritores em diferentes escalas (b) efeito da escala em cada método.

Nas imagens com baixa rotação e diferença de escala, o descritor *LOFTr* obteve uma quantidade superior de pontos filtrados pelos método *RANSAC* (*inliers*) em relação aos demais descritores pois é o único descritor que levanta os pontos com a combinação entre as camadas convolucionais e mecanismos de atenção, conseguindo assim relacionar pontos com baixa textura devido a relação de sua posição espacial com regiões de maior textura na mesma imagem com o advento do mecanismo de *self-attention*. A sensibilidade a rotação e escala é um problema inerente às camadas convolucionais tradicionais, como apontado por Bronstein *et. al.*[25]. Uma possível maneira de mitigar a disparidade em rotação e escala entre as imagens aéreas e as satelitais é o uso da bússola para inferir a rotação e do altímetro para o cálculo da escala. Com essas medidas, podemos extrair o máximo desempenho dos atuais descritores locais baseados em inteligência artificial. Por último, foi realizado um teste para obter a localização de pontos previamente selecionados na imagem aérea e foi calculado a distância euclidiana entre o ponto obtido manualmente e pelo estimado. Essa estimação foi obtida através da multiplicação do ponto selecionado na imagem aérea pela matriz de homografia gerada pelo algoritmo RANSAC a partir dos pontos fornecidos pelos descritores automáticos. Os Erros médios e máximos encontrados pelos métodos que utilizam inteligência artificial foram consideravelmente menores que os erros encontrados pelos descritores tradicionais, demonstrando como as redes neurais convolucionais assistidas pelo mecanismo de atenção tem a maior habilidade em lidar com cenas que apresentam grandes mudanças de perspectiva, iluminação, cor, translação, escala e rotação. O erro máximo dos métodos com inteligência artificial foi menor que 5 metros, que é o erro médio do GPS de um aparelho celular em ambiente externo com poucas nuvens [26]. O erro por descritor está demonstrado na Tabela I. Destaca-se que as redes neurais *LOFTr* e *SuperGlue* foram treinadas apenas com mapas de profundidade pelo banco de dados MegaDepth [23] e conseguiram generalizar com alta precisão na estimativa dos pontos correspondentes em uma região de voo inédita para aquelas redes, presumindo-se assim que esses métodos podem ser empregados em outros ambientes de voo.

TABELA I

ERRO POR DESCRITOR

Baseados em IA	*LOFTr*	*SuperGlue*	*D2NET*
Médio	0,88	0,98	1,00
Máximo	3,15	3,16	4,56
Tradicionais	*SIFT*	*ORB*	*AKAZE*
Médio	10,26	30,30	16,52
Máximo	56,33	194,42	163,99

IV. CONCLUSÃO

Os descritores locais de imagens baseados em inteligência artificial geraram um número elevado de pontos correspondentes entre as imagens aéreas e satelitais. Com esses pontos, foram obtidas boas estimativas de homografia e através da comparação com o método manual de localização de pontos, foi obtido um erro médio menor que um metro. Demonstrou-se também, para os dados utilizados, a superioridade dos métodos baseados em rede neural e mecanismos de atenção perante os descritores tradicionais utilizados no campo da visão computacional. Além disso, os descritores baseados em IA tiveram bons resultados sem nenhum treinamento da rede neural com imagens retiradas anteriormente do local de voo, indicando assim o poder de generalização desses algoritmos. Por último, foi verificado a perda de acurácia devido a diferença angular e de escala entre a imagem aérea e a de referência devido às características das camadas convolucionais, indicando a importância do uso de técnicas para estimar a rotação e escala, tais como, o uso de um sensor inercial ou outras técnicas de visão computacional. Desta forma, os novos descritores locais de imagem, baseados em redes neurais convolucionais e mecanismos de atenção, indicam potencial para atuar em sistemas de localização por imagens em veículos aéreos.

REFERÊNCIAS

[1] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for uav," *Robotics and Autonomous Systems*, vol. 135, p. 103666, 2021.
[2] Andrej Karpathy. Keynote in cvpr'21 wad. CVPR. [Online]. Available: https://www.youtube.com/watch?v=g6bOwQdCJrc
[3] M. Mantelli, D. Pittol, R. Neuland, A. Ribacki, R. Maffei, V. Jorge, E. Prestes, and M. Kolberg, "A novel measurement model based on abbrief for global localization of a uav over satellite images," *Robotics and Autonomous Systems*, vol. 112, pp. 304–319, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092188901830438X
[4] Y. Kim, "Aerial map-based navigation using semantic segmentation and pattern matching," *ArXiv*, vol. abs/2107.00689, 2021.
[5] A. S. Nassar, K. Amer, R. ElHakim, and M. Elhelw, "A deep cnn-based framework for enhanced aerial imagery registration with applications to uav geolocalization," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1594–159 410, 2018.
[6] M. Bianchi and T. D. Barfoot, "Uav localization using autoencoded satellite images," *IEEE Robotics and Automation Letters*, vol. 6, pp. 1761–1768, 2021.
[7] E. Karami, S. Prasad, and M. S. Shehata, "Image matching using sift, surf, brief and orb: Performance comparison for distorted images," *ArXiv*, vol. abs/1710.02726, 2017.
[8] M. C. Dorst, "Distinctive image features from scale-invariant keypoints," 2011.
[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.
[10] D. Viswanathan, "Features from accelerated segment test (fast)," 2011.
[11] M. Calonder, V. Lepetit, C. Strecha, and P. V. Fua, "Brief: Binary robust independent elementary features," in *ECCV*, 2010.
[12] D. Roos, E. H. Shiguemori, and A. C. Lorena, "Comparing orb and akaze for visual odometry of unmanned aerial vehicles," 2016.

[13] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *BMVC*, 2013.
[14] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
[15] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8084–8093, 2019.
[16] H. Hou, Q. Xu, C. Lan, W. Lu, Y. Zhang, Z. Cui, and J. Qin, "Uav pose estimation in gnss-denied environment assisted by satellite imagery deep learning features," *IEEE Access*, vol. 9, pp. 6358–6367, 2021.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[18] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.
[19] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," *CVPR*, 2021.
[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[21] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3852–3861, 2017.
[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.
[23] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050, 2018.
[24] Google. (2022) Google, map data: Google, landsat/copernicus. [Online]. Available: https://www.google.com.br/intl/pt-BR/earth/
[25] M. M. Bronstein, J. Bruna, T. Cohen, and P. Velivckovi'c, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *ArXiv*, vol. abs/2104.13478, 2021.
[26] U. government. (2022) Gps.gov: Gps accuracy. [Online]. Available: https://www.gps.gov/systems/gps/performance/accuracy/how-accurate