Comparing U-Net based architectures in monocular depth estimation

Antônio Carlos Durães da Silva[1], Kelly Assis de Souza Gazolli[2]
Instituto Federal do Espírito Santo, Serra, ES, Brasil
antoniocds1996@gmail.com[1], kasgazolli@gmail.com[2]

***Abstract*—Monocular depth estimation is a computer vision problem which has diverse applications ranging from augmented reality to surgical procedures. Given the similarity between the segmentation and monocular depth estimation tasks, in addition to the good performance of the U-net network and its variations in the segmentation task, this study aims to compare the performance of variations of U-Net and UNet++ architectures, each one adopting a different network as encoder, and the TransUnet architecture in monocular depth estimation. The results achieved on the NYU Depth V2 dataset shows that U-Net using Mix Transformer (MiT-B2) as encoder outperforms all other evaluated approaches.**

***Index Terms*—Monocular depth estimation, U-Net, UNet++, Transunet.**

I. Introdução

A tarefa de estimativa de profundidade (EP) consiste em calcular um valor numérico que representa a distância entre um pixel de uma imagem e o seu observador [1]. Essa atividade desempenha um importante papel em diversas aplicações, tais como, realidade aumentada [2], identificação de proximidade com elementos em guiagem autônoma [3] e até mesmo em procedimentos cirúrgicos auxiliados por computador [4, 5].

Existem dispositivos capazes de realizar a estimativa de profundidade, um dos mais utilizados é o sensor LIDAR (do inglês, *Laser Imaging Detection and Ranging*), juntamente com as câmeras estéreo [6]. No entanto, tais dispositivos são propensos a problemas mecânicos e complicações relacionadas à natureza das superfícies do ambiente, além do alto custo de aquisição [6]. Paralelamente às questões que envolvem o uso de dispositivos dedicados, há a oportunidade de construir aplicações de realidade aumentada para dispositivos móveis, sendo essa uma das principais motivações para estimativa de profundidade usando imagens obtidas por câmeras de baixo custo, que capturam fotografias monoculares [7].

Nesse contexto, foram desenvolvidas diversas abordagens que realizam a estimativa de profundidade monocular. De acordo com Ming et al. [8], inicialmente, essa tarefa era executada por meio de técnicas primitivas que se baseavam em pistas de profundidade (como foco, sombra e formas) [9, 10], posteriormente por modelos probabilísticos baseados em aprendizado de máquina [11, 12] e por fim, por métodos baseados em aprendizado profundo, como redes neurais artificiais.

Assim como na tarefa de estimativa de profundidade, a segmentação semântica vem sendo aprimorada por meio do emprego de redes neurais artificiais [13], sendo a arquitetura U-Net e suas variantes frequentemente aplicadas [14, 15, 16]. Inicialmente projetada com foco na segmentação de imagens médicas [17], a rede U-Net, que é do tipo codificador-decodificador, recebeu diversas variações [18] e passou a ser utilizada na solução de outros problemas de visão computacional [19, 20].

Segundo Ming et al. [8], tanto a segmentação semântica quanto a estimativa de profundidade são classificações no nível de pixel, o que possibilita o compartilhamento das características extraídas da imagem entre ambas tarefas, sendo necessários apenas dois módulos de saída distintos, um para cada finalidade.

Dada a semelhança entre as tarefas de segmentação e de estimativa de profundidade, este trabalho tem como principal objetivo verificar o desempenho de combinações de arquiteturas baseadas em U-Net aplicadas à estimativa de profundidade de imagens monoculares. Foram selecionados codificadores bem estabelecidos na literatura para extração de características de imagens, são eles: MiT-B2, Inception ResNet V2, VGG-19 e Xception. Por fim, além da U-Net original, foram selecionadas duas de suas variantes: TrasUnet e UNet++.

O aprendizado auto-supervisionado têm sido frequentemente aplicado na estimativa de profundidade, alcançado bom desempenho. No entanto, os métodos existentes utilizam redes de arquiteturas complexas, que dependem de módulos para reconstrução e discriminação da entrada ou para estimativa de pose [21, 22, 23]. Assim, nos experimentos realizados foi adotado o aprendizado supervisionado, utilizando-se a base de dados NYU Depth V2 [24], mais especificamente um subconjunto de 50 mil imagens, conforme proposto por Alhashim and Wonka [25].

Os resultados obtidos apontam que a combinação da U-Net com o codificador Mix Transformer (MiT-B2) apresenta desempenho superior dentre as abordagens avaliadas, inclusive as mais complexas, com maior número de camadas e conexões aninhadas.

Este trabalho está organizado da seguinte forma: Na Seção II é feita uma revisão da literatura. Na Seção III, é apresentada a metodologia utilizada, detalhando-se os modelos empregados. Na Seção IV, são apresentados os experimentos e resultados alcançados. Por fim, na Seção V, são apresentados a conclusão e os trabalhos futuros.

II. Revisão da Literatura

As redes neurais convolucionais (CNN - *Convolutional Neural Network*) são arquiteturas que usam de operações de convolução para extrair características da entrada e camadas totalmente conectadas para construir sua saída [8]. Essa categoria de rede neural se popularizou após resultados promissores na área de classificação de imagens com arquiteturas como Google LeNet [26], ImageNet [27] e Residual Net (ResNet) [28].

Devido a sua arquitetura flexível e facilmente modificável, a U-Net tem sido uma arquitetura de CNN comumente adotada como base para a construção de novas redes na tarefa de estimativa de profundidade [29, 30, 31]. Saxena et al. [31] combinaram a U-Net, utilizando a EfficientNet como codificador, com a reconstrução de mapas de profundidades ruidosos para melhorar significativamente a estimativa de profundidade. Jan and Seo [32] substituíram camadas simples de convolução 2D da U-Net por camadas de convoluções residuais para aprimorar a extração de características mais expressivas, além de adicionarem um mecanismo de atenção antes de cada camada do decodificador, permitindo que a rede também seja capaz de identificar pequenas características e de construir mapas de profundidade refinados. Yang et al. [29] propuseram uma arquitetura supervisionada baseada na UNet++ (versão da U-Net que incluiu conexões aninhadas entre codificador e decodificador [14]) e adotaram pirâmides de convoluções dilatadas para reduzir o custo computacional e ampliar a capacidade da rede para capturar características da imagem em diversas escalas. Com o intuito de desenvolver uma solução mais leve, Guzzo and Gazolli [33] propuseram uma abordagem para estimativa de profundidade que utiliza a arquitetura UNet++ empregando como codificador a rede neural MobileNetV2 pré-treinada.

III. Metodologia

A. Arquiteturas

1) U-Net: Ronneberger et al. [17] propuseram a rede U-Net, uma arquitetura do tipo codificador-decodificador, para a segmentação semântica de imagens biomédicas. Uma das principais contribuições dos autores é a proposta de unir uma sequência de contração (usada para extrair informações contextuais) com uma sequência de expansão (utilizada para capturar informações de localização) por meio de conexões salto (*skip connections*).

Na Figura 1 está a representação da arquitetura U-Net. O processo se inicia com o codificador, que recebe a imagem de entrada e extrai os mapas de características (representadas pelos blocos azuis) com diferentes dimensões e número de canais.

A concatenação dos dois mapas de características com maior número de canais (F_4 e F_5) passa por duas sequências de convolução 2D, função de ativação e normalização em lote, formando assim o primeiro bloco do decodificador (representado na cor verde).

Com exceção do primeiro bloco do decodificador, que é alimentado por duas saídas do codificador, todos os outros blocos subsequentes são alimentados pela concatenação da saída do bloco decodificador anterior com o mapa de características de um nível acima. À medida que o número de canais do decodificador é reduzido pela metade, a resolução dos mapas de características é dobrada, garantindo que, ao final, a saída da rede tenha a mesma dimensão da entrada.

As seguintes arquiteturas foram utilizadas em conjunto com a U-Net no papel de codificador: Mix Transformer (MiT-B2)[34]; Inception ResNet-V2 [35]; Xception [36], e VGG-19 (com normalização em lote) [37].

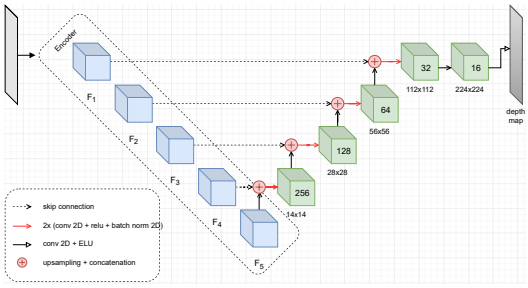

Figura 1: Arquitetura U-Net utilizada

2) UNet++: O trabalho de Zhou et al. [14] redesenhou as conexões de salto da U-Net, incluindo blocos densos de convolução a fim de melhorar o compartilhamento de informações entre codificador e decodificador por meio da concatenação de características de dimensões distintas. Além das mudanças na arquitetura, os autores propuseram o uso de um mecanismo de supervisão profunda (*deep supervision*) para podar camadas da rede treinada durante a inferência de dados e reduzir o tempo necessário para realizar essa tarefa. Essa nova abordagem foi chamada de UNet++.

Nos experimentos realizados neste trabalho, as seguintes arquiteturas foram utilizadas como codificador da rede UNet++: Inception ResNet-V2 [35]; Xception [36], e VGG-19 (com normalização em lote) [37].

3) TransUnet: Considerando o uso promissor de modelos *transformers*, originalmente empregados em tarefas de processamento de linguagem natural [38], na área de visão computacional [39, 40, 41], Chen et al. [42] propuseram uma solução híbrida que combina a rede CNN com camadas *transformers*. Essa arquitetura foi desenvolvida com base na estrutura da U-Net e chamada de TransUnet.

A Figura 2 apresenta o fluxo de dados pela rede e a comunicação entre o módulo CNN, módulo *transformers* e o decodificador. O codificador da arquitetura é composto pelo módulo CNN (que faz uso de uma ResNet-50) e de uma sequência de 12 camadas *transformers*. O módulo CNN recebe a imagem de entrada, extrai suas características e reduz pela metade a dimensão das informações a cada bloco convolucional por uma sequência de 3 blocos, entregando características com 3 dimensões distintas. A saída do módulo CNN passa por um processo de *embedding* e seu resultado é processado pela sequência de camadas *transformers*.

Seu decodificador é inicializado com a concatenação de características extraídas pela sequência de camadas *transformers*

e as características de menor dimensão obtidas pelo módulo CNN. Com exceção da primeira camada do decodificador, as próximas são resultados da concatenação da saída de sua respectiva camada do codificador com a camada anterior do decodificador após uma operação de *upsampling* para dobrar sua resolução, assim como na U-Net.

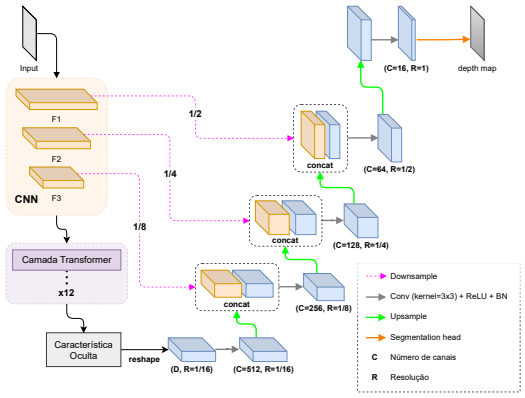

Figura 2: Arquitetura TransUnet

B. Função de perda

Para avaliar o aprendizado das redes analisadas, foi utilizada uma abordagem que envolve a soma ponderada de outras três funções de perdas, conforme proposto por Alhashim and Wonka [25]: erro médio absoluto (L_{depth}), perda L1 sobre os gradientes dos mapas de profundidade (L_{grad}) e medida do índice de similaridade estrutural (L_{SIM}), apresentadas nas Equações 1,2, 3, onde y representa o mapa de estimativa de profundidade verdadeiro e $\hat{y}$, o mapa de estimativa de profundidade predito.

$$L_{\text{depth}}\left(y, \hat{y}\right) = \frac{1}{n}\sum_{n}^{n}\left|y_p - \hat{y}_p\right|. \tag{1}$$

$$L_{\text{grad}}\left(y, \hat{y}\right) = \frac{1}{n}\sum_{p}^{n}\left|\boldsymbol{g}_{\mathbf{x}}\left(y_p, \hat{y}_p\right)\right| + \left|\boldsymbol{g}_{\mathbf{y}}\left(y_p, \hat{y}_p\right)\right| \tag{2}$$

$$L_{\text{SIM}}(y, \hat{y}) = \frac{1 - \text{SSIM}(y, \hat{y})}{2}. \tag{3}$$

A função final, resultante da soma das três funções de perda, é apresentada na Equação 4:

$$L(y, \hat{y}) = \lambda L_{\text{depth}}\left(y, \hat{y}\right) + L_{\text{grad}}\left(y, \hat{y}\right) + L_{SSIM}(y, \hat{y}) \tag{4}$$

IV. Experimentos e resultados

A. Base de dados

Os experimentos foram executados utilizando-se a base de imagens coloridas e *indoor* NYU V2 Depth [24]. Embora o conjunto original seja composto por mais de 100 mil imagens, cada uma com uma resolução de 640x480, optou-se por empregar o subconjunto de 50 mil amostras, conforme estabelecido por Alhashim and Wonka [25], para a fase de treinamento das redes. Para a avaliação das arquiteturas foi utilizado o subconjunto oficial de imagens e mapas de profundidade NYU V2 dedicados à etapa de teste, com 654 pares de imagens e mapas de profundidade.

Ainda na etapa de treinamento, a resolução das imagens foi reduzida para 224x224, utilizando a técnica de interpolação bilinear com o propósito de mitigar distorções significativas do processo de redimensionamento. Com intuito de desconsiderar bordas vazias das imagens, durante a etapa de teste, tanto as imagens coloridas quanto os mapas de profundidade foram recortados utilizando os valores propostos por Eigen et al. [43].

Visando aprimorar a capacidade de aprendizado da rede ao lidar com novos dados (generalização) e mitigar o problema de sobre-ajuste, foram incluídas as duas transformações (50% de chance de espelhar horizontalmente a imagem e 25% de chance de permutar seus canais de cores) utilizadas por Alhashim and Wonka [25] para manipular os dados do conjunto de treinamento. Ao carregar cada par de imagem e mapa de profundidade, a estratégia de aumento de dados (*Data Augmentation*) pode ser aplicada ao par, atuando sobre o conjunto de treino sem acrescentar novos dados, apenas transformando os dados existentes antes de processá-los.

B. Detalhes de implementação

Todos os experimentos foram executados em máquinas que seguem a configuração: Placa gráfica NVIDIA Tesla P100 (16GB), CPU Intel Xeon 2.00GHz, 16GB RAM e Debian 8.10 como sistema operacional. Em todas as arquiteturas, foram utilizadas as versões pré-treinadas no conjunto de dados para classificação de imagens ImageNet de seus codificadores. Durante a etapa de treinamento, não foi realizado o congelamento de qualquer camada de nenhum dos codificadores.

C. Avaliação

Para verificar o desempenho dos modelos, foram utilizadas as seguintes métricas de estimativa de profundidade, propostas por Eigen et al. [43] e apresentadas nas Equações 5,6, 7, 8:

- *Threshold* (δ_j):

$$\delta_j : \%\text{ of } \hat{i} \text{ s.t. } \max\left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}\right) < 1.25^j, j \in \{1, 2, 3\} \tag{5}$$

- *Root Mean Square Error* (RMSE):

$$RMSE : \sqrt{\frac{1}{P}\sum_{i=1}^{P}(y_i - \hat{y}_i)^2} \tag{6}$$

- *Scale invariant error* ($\log_{10}$):

$$log10 = \frac{1}{P}\sum_{i=1}^{P}\left|\log_{10}(y_i) - \log_{10}(\hat{y}_i)\right| \tag{7}$$

- *Absolute Relative Difference* (rel):

$$rel = \frac{1}{P}\sum_{i=1}^{P}\frac{|y_i - \hat{y}_i|}{y_i} \tag{8}$$

Tabela I: Comparação quantitativa considerando as arquiteturas propostas e trabalhos relacionados aplicados ao conjunto de dados NYU Depth V2

Method	Encoder	RMSE ↓	rel ↓	log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Eigen et al. [43]	-	0,641	0,158	-	0,769	0,950	0,988
Fu et al. [44]	-	0,509	**0,115**	**0,051**	0,828	0,965	**0,992**
He et al. [45]	VGG	0,572	0,151	0,064	0,789	0,948	0,986
Qi et al. [46]	ResNet-50	0,569	0,128	0,057	0,834	0,960	0,990
He et al. [47]	Xception + ASPP	0,514	0,145	0,062	0,805	0,962	**0,992**
Zhang et al. [48]	ResNet-50	0,501	0,144	-	0,815	0,962	**0,992**
TransUnet	ResNet-50	0,478	0,134	0,055	0,844	0,965	0,988
U-Net adaptada	MiT (B2)	**0,461**	0,129	0,054	**0,851**	**0,97**	**0,992**
	Inception ResNet V2	0,479	0,128	0,055	0,846	0,967	0,991
	VGG-19 (BN)	0,481	0,129	0,055	0,843	0,967	**0,992**
	Xception	0,501	0,133	0,057	0,838	0,963	0,988
U-Net++ adaptada	Inception ResNet V2	0,489	0,13	0,055	0,844	0,964	0,989
	VGG-19 (BN)	0,499	0,14	0,059	0,821	0,964	0,991
	Xception	0,492	0,132	0,058	0,843	0,965	0,988

D. Resultados

Na Tabela I são apresentados os resultados quantitativos obtidos pelos modelos avaliados: arquiteturas U-Net e UNet++ com diferentes codificadores, e TransUnet, bem como por outras arquiteturas reconhecidas por sua aplicação na estimativa de profundidade monocular. Nela podemos observar que o modelo obtido por meio da utilização da rede MiT-B2 como codificador da U-Net apresenta valores superiores para todos os limiares (δ) e RMSE, quando comparado a outras abordagens, inclusive aquelas com arquiteturas mais complexas, como as propostas por He et al. [47] e Zhang et al. [48], que combinam informações das tarefas de segmentação e estimativa de profundidade para melhorar a acurácia em ambas. Vale dizer que a rede MiT-B2 é um codificador *transformer* hierárquico projetado e otimizado para a segmentação semântica. A TransUnet, outra abordagem proposta para segmentação por Chen et al. [42], também apresenta resultados relevantes para as mesmas métricas, com exceção do δ_3, em que a rede é superada por quase todas soluções. Vale dizer que todas as abordagens avaliadas neste trabalho obtiveram os menores valores para a raiz do erro quadrático médio (RMSE), quando comparadas com as outras arquiteturas.

Embora as implementações utilizando UNet++ tenham demonstrado resultados significativos nas métricas RMSE, δ_1, e δ_2, em geral, os resultados ficaram inferiores em comparação com os obtidos pela U-Net tradicional. Além disso, as implementações UNet++ exigem uma carga computacional mais elevada devido a suas convoluções densas e aninhadas. Esse desempenho sugere que, para aproveitar plenamente o potencial do aninhamento de seus blocos, podem ser necessários ajustes arquiteturais mais refinados.

V. Conclusão

Este trabalho explorou a similaridade entre a tarefa de estimativa de profundidade monocular e a segmentação semântica por meio da comparação de redes e codificadores concebidos inicialmente para a tarefa de segmentação. A partir da avaliação quantitativa das diversas arquiteturas, pode-se observar a influência da escolha do codificador no desempenho da estimativa de profundidade. Os resultados derivados da combinação entre a arquitetura U-Net e o codificador Mix Transformer (MiT-B2), proposto originalmente para a tarefa de segmentação, são particularmente promissores, visto ter apresentado desempenho superior em relação as outras abordagens avaliadas. Tal análise é válida também quando comparada com outras redes, inclusive aquelas dotadas de estruturas mais complexas. Além disso, a TransUnet também se mostrou promissora na tarefa de estimativa de profundidade. Os resultados obtidos neste trabalho reforçam a hipótese da similaridade entre segmentação semântica e estimativa de profundidade monocular, fornecendo pistas para o desenvolvimento de soluções mais eficazes e adequadas para casos onde há restrições de recursos computacionais, uma vez que o aprendizado supervisionado foi utilizado. Como trabalhos futuros, pretende-se avaliar as arquiteturas em outras bases de dados, além de investigar estratégias de otimização para que as arquiteturas propostas sejam mais eficazes em cenários de recursos computacionais restritos.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

[1] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digital Signal Processing*, vol. 123, p. 103441, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200422000586

[2] R. Huang and M. Sun, "Network algorithm real-time depth image 3d human recognition for augmented

reality," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 307–319, Nov. 2020. [Online]. Available: https://doi.org/10.1007/s11554-020-01045-z

[3] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. Lopez, "Multimodal end-to-end autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 537–547, Jan. 2022. [Online]. Available: https://doi.org/10.1109/tits.2020.3013234

[4] H. Itoh, M. Oda, Y. Mori, M. Misawa, S.-E. Kudo, K. Imai, S. Ito, K. Hotta, H. Takabatake, M. Mori, H. Natori, and K. Mori, "Unsupervised colonoscopic depth estimation by domain translations with a lambertian-reflection keeping auxiliary task," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 6, pp. 989–1001, May 2021. [Online]. Available: https://doi.org/10.1007/s11548-021-02398-x

[5] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1438–1447, May 2020. [Online]. Available: https://doi.org/10.1109/tmi.2019.2950936

[6] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, and S. Mattoccia, "On the synergies between machine learning and binocular stereo for depth estimation from images: a survey," 2020. [Online]. Available: https://arxiv.org/abs/2004.08566

[7] J. Xie, C. Lei, Z. Li, L. E. Li, and Q. Chen, "Video depth estimation by fusing flow-to-depth proposals," 2019. [Online]. Available: https://arxiv.org/abs/1912.12874

[8] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.12.089

[9] Y. J. Jung, A. Baik, J. Kim, and D. Park, "A novel 2d-to-3d conversion technique based on relative height-depth cue," in *SPIE Proceedings*, A. J. Woods, N. S. Holliman, and J. O. Merritt, Eds. SPIE, Feb. 2009. [Online]. Available: https://doi.org/10.1117/12.806058

[10] K. Han and K. Hong, "Geometric and texture cue based depth-map estimation for 2d to 3d image conversion," in *2011 IEEE International Conference on Consumer Electronics (ICCE)*, 2011, pp. 651–652.

[11] H. Yan, X. Yu, Y. Zhang, S. Zhang, X. Zhao, and L. Zhang, "Single image depth estimation with normal guided scale invariant deep convolutional fields," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 80–92, 2019.

[12] S.-P. Tseng and S.-H. Lai, "Accurate depth map estimation from video via mrf optimization," in *2011 Visual Communications and Image Processing (VCIP)*, 2011, pp. 1–4.

[13] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2032924, 2022. [Online]. Available: https://doi.org/10.1080/08839514.2022.2032924

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020. [Online]. Available: https://doi.org/10.1109/tmi.2019.2959609

[15] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," 2020.

[16] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*. Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[18] N. S. Punn and S. Agarwal, "Modality specific u-net variants for biomedical image segmentation: a survey," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5845–5889, Mar. 2022. [Online]. Available: https://doi.org/10.1007/s10462-022-10152-1

[19] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention unet for building segmentation in remote sensing images," *Science China Information Sciences*, vol. 63, no. 4, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s11432-019-2791-7

[20] K. Cao and X. Zhang, "An improved res-unet model for tree species classification using airborne high-resolution images," *Remote Sensing*, vol. 12, no. 7, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/7/1128

[21] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[22] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 572–588. [Online]. Available: https://doi.org/10.1007/978-3-030-58529-7_34

[23] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9250–9256.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, 2012, pp. 746–760. [Online]. Available: https://doi.org/10.1007/978-3-642-33715-4_54

[25] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv e-prints*, vol. abs/1812.11941, 2018. [Online]. Available: https://arxiv.org/abs/1812.11941

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] Y. Yang, Y. Wang, C. Zhu, M. Zhu, H. Sun, and T. Yan, "Mixed-scale unet based on dense atrous pyramid for monocular depth estimation," *IEEE Access*, vol. 9, pp. 114 070–114 084, 2021.

[30] H.-T. Duong, H.-M. Chen, and C.-C. Chang, "URNet: An UNet-based model with residual mechanism for monocular depth estimation," *Electronics*, vol. 12, no. 6, p. 1450, Mar. 2023. [Online]. Available: https://doi.org/10.3390/electronics12061450

[31] S. Saxena, A. Kar, M. Norouzi, and D. J. Fleet, "Monocular depth estimation using diffusion models," 2023.

[32] A. Jan and S. Seo, "Monocular depth estimation using res-UNet with an attention model," *Applied Sciences*, vol. 13, no. 10, p. 6319, May 2023. [Online]. Available: https://doi.org/10.3390/app13106319

[33] L. Guzzo and K. Gazolli, "Utilizando a arquitetura unet++ na estimativa de profundidade monocular," in *Anais do L Seminário Integrado de Software e Hardware*. Porto Alegre, RS, Brasil: SBC, 2023, pp. 131–142. [Online]. Available: https://sol.sbc.org.br/index.php/semish/article/view/25068

[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[39] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, sep 2022. [Online]. Available: https://doi.org/10.1145/3505244

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 213–229. [Online]. Available: https://doi.org/10.1007/978-3-030-58452-8_13

[41] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[42] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[43] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014.

[44] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," 2018.

[45] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4676–4689, sep. [Online]. Available: https://doi.org/10.1109%2Ftip.2018.2832296

[46] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 283–291.

[47] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "SOSD-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, Jun. 2021. [Online]. Available: https://doi.org/10.1016/j.neucom.2021.01.126

[48] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 238–255. [Online]. Available: https://doi.org/10.1007/978-3-030-01249-6_15