Estudo de Estratégia de Aprendizado Auto-supervisionado para Aprimoramento da Consistência Temporal em Modelo de Segmentação Semântica Baseado em Deep Learning

Felipe M. Barbosa[1], **Fernando S. Osório**[1]

[1]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo (USP) - São Carlos - SP - Brasil

{felipe.manfio.barbosa,fosorio}@usp.br

Abstract. *Deep Learning-based Semantic segmentation is a task of utmost importance in visual perception for autonomous mobile robots. However, great part of the current research explores single-frame perception. This approach, besides neglecting the possibilities offered by the use of temporal data, leads to unstable models. In light of that, and considering the high cost of data labeling, new learning alternatives try to leverage the widely-available non-labeled temporal data. Therefore, in this work, we study the application of a self-supervised auxiliary supervision strategy for the promotion of temporal stability in semantic segmentation models. The results demonstrate that this strategy promotes model's precision and stability, even when utilizing data from distinct datasets.*

Resumo. *Segmentação semântica por meio de Deep Learning tem extrema importância na percepção visual para robôs móveis autônomos. Contudo, grande parte da pesquisa atual se baseia percepção quadro-a-quadro. Tal abordagem, além de negligenciar as possibilidades oferecidas pelo uso de dados temporais, resulta em modelos instáveis. Diante disso, e do alto custo do processo de rotulação, novas alternativas de aprendizado exploram a ampla disponibilidade de dados temporais não-rotulados. O presente trabalho estuda a aplicação de supervisão auxiliar auto-supervisionada para promoção da estabilidade temporal em modelos de segmentação. Os resultados demonstram que tal estratégia promove a precisão e estabilidade, mesmo utilizando dados de bases distintas.*

1. Introdução

Segmentação semântica pode ser definida como uma tarefa de classificação densa, em que a cada pixel de uma imagem é associada uma dada classe. Soluções iniciais propunham a extração manual de características das imagens – descritores HOG e SIFT [Seyedhosseini and Tasdizen 2016], assim como medidas estatísticas de média e entropia[Shinzato and Wolf 2010]–, seguida por uma etapa de classificação. Tais abordagens, contudo, além de envolverem passos separados para extração e classificação, apresentavam limitações em cenários para os quais as mesmas não haviam sido calibradas.

Com o advento das técnicas de *Deep Learning* e, mais especificamente, das Redes Neurais Convolucionais (CNNs), obervou-se uma mudança de paradigma no tocante ao desenvolvimento de modelos de segmentação. Um passo importante nesse sentido foi dado no ano de 2015, com a proposição das *Fully Convolutional Networks* (FCNs)

[Long et al. 2015], as primeiras a abordar a segmentação semântica como uma tarefa de classificação densa de imagens, segundo a óptica do *Deep Learning*.

Com base em tais avanços, e em trabalhos correlatos (U-net [Ronneberger et al. 2015] e Segnet [Badrinarayanan et al. 2017]), fez-se possível a segmentação de imagens em quaisquer resoluções, de forma automática e fim-a-fim (*end-to-end*). Dessa forma, atingiu-se um novo patamar em termos de precisão, dando origem uma linha de pesquisa dedicada a atingir valores cada vez maiores de acurácia.

Contudo, os ganhos em precisão são usualmente obtidos ao custo de arquiteturas mais complexas e altos requisitos computacionais, tornando tais modelos inaptos a aplicações com poder computacional limitado e que exijam percepção em tempo real, como é o caso de robôs móveis autônomos e, mais especificamente, veículos autônomos. Como consequência, tem-se observado um interesse crescente na proposição de modelos leves e rápidos, ou seja, voltados à eficiência. Tais estratégias têm por objetivo garantir a percepção do ambiente em tempo hábil, permitindo, por exemplo, identificar possíveis situações de risco e dar ao sistema tempo hábil para agir de modo a evitar maiores danos. Tal característica é particularmente importante no contexto de veículos autônomos.

Entretando, independentemente da linha adotada (precisão ou eficiência), a maior parte das contribuições têm como foco a percepção quadro-a-quadro, não considerando dados temporais (quadros sequenciais) no processamento do modelo. Isso leva a modelos com considerável instabilidade, tanto em curto quanto em longo prazos (figura 1). Outro ponto que merece atenção diz respeito às métricas utilizadas por tais modelos, que consideram apenas a precisão local como forma de avaliação. Em aplicações de processamento em tempo real e sistemas críticos, contudo, a estabilidade se faz tão importante quanto a precisão e, dessa forma, também deve ser avaliada.

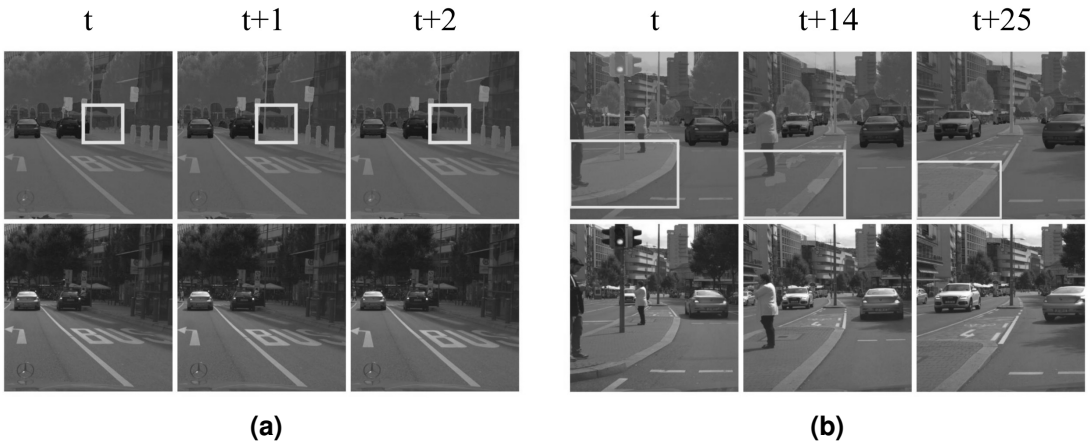

Figura 1. Modelos de percepção quadro-a-quadro podem resultar em instabilidade em (a) curto e (b) longo prazos. Melhor visualizado em cores.

Finalmente, a maior parte da literatura em segmentação semântica baseada em *Deep Learning* faz uso da abordagem de aprendizado supervisionado, partindo do pressuposto da disponibilidade de rótulos para os dados a serem aplicados no treinamento e validação dos modelos. Para bases de tamanho reduzido, tal suposição é razoável. Contudo, quando consideramos, por exemplo, bases de dados referentes a ambientes urbanos para navegação autônoma, o cenário muda completamente. Tais conjuntos são compostos

por milhares de imagens obtidas a partir de centenas, ou mesmo milhares de trechos de vídeos capturados sob diferentes condições e contendo diferentes classes de elementos. Enquanto tais características são benéficas do ponto de vista do poder de representatividade dos dados, sua dimensão torna impraticável a rotulação completa da base – contexto que já recebeu a denominação de ”curse of dataset annotation” [Xie et al. 2016].

Diante de tal cenário, faz-se necessário o uso de abordagens de aprendizado alternativas, que explorem o potencial da grande disponibilidade de dados sequenciais não rotulados para aprimorar a estabilidade de modelos de segmentação, sem a necessidade de processos dispendiosos de rotulação e sem comprometer a eficiência de tais arquiteturas.

Nesse sentido, o uso de aprendizado auto-supervisionado, atrelado a estratégias de supervisão auxiliar para promoção da consistência temporal, é uma alternativa promissora, recente, e ainda pouco explorada na literatura. Nosso objetivo neste estudo é verificar a aplicação de tal estratégia a modelos de segmentação eficientes baseados em percepção quadro-a-quadro, de modo a avaliar o impacto da mesma tanto na precisão, quanto na estabilidade dos modelos em questão. Diferentes cenários são avaliados, considerando: (i) diferentes ponderações das componentes de precisão e estabilidade nas função de custo do modelo; (ii) bases de dados de diferentes origens para uso na tarefa de supervisão auxiliar – ambas relacionadas à percepção em ambiente urbano.

2. Trabalhos Relacionados

A seguir, são apresentados os principais trabalhos relacionados, no tocante ao projeto orientado à eficiência, métricas alternativas de avaliação e aprendizado auto-supervisionado.

2.1. Modelos Orientados à Eficiência

Apesar de recente, a linha de pesquisa em segmentação semântica voltada à eficiência é bastante vasta. As principais abordagens podem ser divididas, segundo o nível de aplicação, em: estratégias a nível de entrada, nível de operação e nível de arquitetura.

Neste trabalho, utilizamos como base uma estratégia a nível de arquitetura que envolve o uso de extratores de características assimétricos. Particularmente, estruturas *multi-encoder* assimétricas, em que cada ramificação possui profundidades e larguras particulares, têm sido amplamente adotadas, com destaque à família de modelos BiSeNet.

Proposta em 2018, o modelo BiseNet [Yu et al. 2018] realiza a extração de características através de dois ramos de codificação com estruturas e finalidades diferentes. O ramo de codificação espacial possui poucas camadas de processamento e trabalha com mapas com maior dimensionalidade, visando à extração de informações com maior riqueza estrutural. O ramo contextual, por outro lado, possui uma estrutura de processamento mais profunda e com menores dimensões, a fim de captar maior riqueza semântica. Dentre as diversas adaptações propostas ao modelo original, no presente estudo empregamos o modelo BiseNet V2 [Yu et al. 2021] como base para os experimentos.

2.2. Métricas Alternativas de Avaliação

Desde a proposição dos primeiros modelos de *Deep Learning* voltados à Segmentação Semântica, a métrica Intersecção sobre União (*Intersection over Union*, *IoU*) [Everingham et al. 2015] vem sendo adotada como a principal forma de avaliação.

Contudo, a precisão a nível de pixel pode, intrinsecamente, ser uma medida equivocada. De fato, [Shi et al. 2022] argumenta que em regiões próximas a bordas entre classes há alta incidência de erros de rotulação (por agentes humanos), enquanto segundo [Oršić and Šegvić 2021] tais regiões representam o maior número de pixels presentes nas imagens. Sendo assim, partindo da suposição que os rótulos podem estar enviesados, não é viável avaliar a precisão de forma direta, ou ainda, avaliar somente medidas de precisão.

Outra limitação inerente às métricas de avaliação tradicionais é a ausência da noção de temporalidade. Nesse sentido, trabalhos recentes no ramo da Segmentação Semântica voltada à percepção em ambientes urbanos têm aplicado a métrica denominada Consistência Temporal (*Temporal Consistency*, *TC*) – equação 1 – como forma de avaliação da estabilidade das previsões do modelo – Tabela 1.

$$TC_t \;=\; mIoU(m_t, \tilde{m}_t) \tag{1}$$

Em que m_t é a saída original no tempo t e $\tilde{m}_t$ é a saída propagada do tempo $t-1$ para t. Cabe ainda ressaltar que qualquer métrica poderia ser utilizada no lugar de *mIoU*, a depender da finalidade. Seu valor médio é obtido por meio do cálculo da média considerando todos os T quadros em um dado vídeo, como na equação 2.

$$mTC \;=\; \frac{1}{T-1}\sum_{t=2}^{T} TC_t \tag{2}$$

Embora recente e pouco adotada, tal métrica tem papel essencial na avaliação de modelos de percepção voltados a aplicações que exigem estabilidade temporal, como veículos autônomos.

Tabela 1. Valores de Consistência Temporal relatados por contribuições recentes na área de Segmentação Semântica de ambientes urbanos.

Ano	Método	Backbone	Pré-treinamento	Resolução	mIoU (Cityscapes)		TC
					Val	Test	
2020	MobileNetV2+ALL [Liu et al. 2020]	MobileNetV2			73.9		69.9
2021	STT-BiSe18 [Li et al. 2021]	ResNet18			75.8		71.4
2021	CSRNet [Xiong et al. 2021]	ResNet18	Cityscapes	1024x2048	75.9		75.3

2.3. Aprendizado Auto-Supervisionado

Apesar de intuitivo, o uso de aprendizado supervisionado nem sempre é possível. Em cenários de escassez de rótulos e abundância de dados não-rotulados, faz-se necessária a exploração de métodos de aprendizado que explorem o potencial presente em tais dados.

O paradigma de aprendizado auto-supervisionado propõe a exploração somente de dados não-rotulados, visando a extrair representações úteis a partir dos mesmos. Usualmente, aplica-se algum tipo de transformação aos dados brutos, como rotações ou remoção de determinadas regiões, de modo que o modelo tente, durante o processo de otimização, recuperar o dado original ou identificar a transformação aplicada [Zhang et al. 2016, Lee et al. 2017]. Embora seja geralmente adotado como etapa de pré-treinamento, o aprendizado auto-supervisionado pode ser empregado para promover a otimização do modelo na própria etapa de treinamento. A estratégia estudada no presente

trabalho se aproxima daquelas utilizadas em [Varghese et al. 2020, Varghese et al. 2021, Liu et al. 2020], em que aprendizado auto-supervisionado é aplicado como supervisão auxiliar para penalizar previsões consecutivas inconsistentes.

Entretanto, os casos anteriores utilizam dados do mesmo domínio para o cálculo das diferentes componentes da função de custo. Nosso estudo, portanto, tem como uma de suas principais contribuições a investigação do uso de dados de diferentes domínios para cálculo das componentes de precisão e consistência durante o processo de otimização.

3. Proposta

Diante do exposto, a proposta do presente estudo é verificar a efetividade de uma função de supervisão auxiliar (*TC Loss*) para promoção da consistência temporal de um modelo-base (BiSeNet V2). A configuração proposta é ilustrada na figura 2.

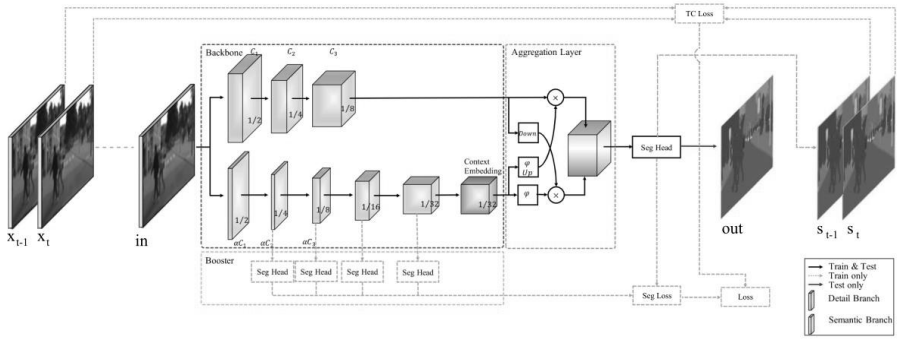

Figura 2. Modelo proposto, construído com base na arquitetura BiseNet V2.

Diferentemente de contribuições anteriores, além da influência dos pesos atribuídos a cada componente na função de perda, estudamos o uso de diferentes bases de dados para o cálculo das diferentes funções de custo empregadas no processo de otimização – enquanto a entrada principal (*in*) provém da base A, as entradas auxiliares (x_{t-1} e x_t) podem ser selecionados a partir das bases A ou B (não simultaneamente).

O procedimento empregado para o cálculo da função de custo auxiliar é ilustrado no diagrama da figura 3. Dadas as entradas sequenciais x_{t-1} e x_t, é cálculado o fluxo óptico $o_{t-1 \rightarrow t}$, por meio da rede FlowNet 2.0 [Ilg et al. 2017]. Em seguida, o mapa de segmentação gerado para o quadro no tempo $t-1$ (s_{t-1}) é propagado para o tempo t por meio do fluxo óptico previamente calculado, gerando o mapa s'_{t-1}. Finalmente, calcula-se a similaridade entre a saida gerada para o tempo t (s_t), e a saída propagada (s'_{t-1}).

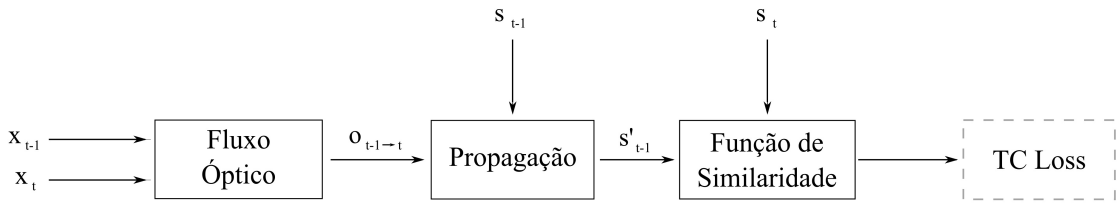

Figura 3. Etapas de cálculo da consistência temporal como supervisão auxiliar.

Cabe ressaltar que, durante a inferência, as estruturas auxiliares utilizadas para treinamento são descartadas. Desse modo, não há impacto na eficiência do modelo-base.

4. Metodologia

Nesta seção, são descritas as características do ambiente de execução, as bases de dados utilizadas e os procedimentos de treinamento e inferência.

4.1. Configuração

Todos os experimentos foram realizados utilizando o serviço *Google Colaboratory* – assinatura com acesso à GPU Nvidia Tesla T4 e 12GB de memória RAM.

A linguagem Python e a biblioteca Pytorch foram empregadas para implementação dos modelos de *Deep Learning*. Mais especificamente, utilizou-se a biblioteca de código aberto MMSegmentation [Contributors 2020], baseada em Pytorch, que fornece uma grande variedade de modelos-base, arquiteturas e pesos pré-treinados.

4.2. Bases de dados

O presente estudo visa a percepção em ambientes urbanos. Dessa forma, foram selecionadas duas bases de dados que se enquadram nesse contexto: (i) Cityscapes, amplamente utilizada na literatura da área; (ii) ZED2, capturada pelos autores a partir de um sensor de visão estéreo ZED2 (Fig. 4). Dados de ambas as bases são exemplificados na figura 5.

Figura 4. Sensor de visão estéreo ZED2.

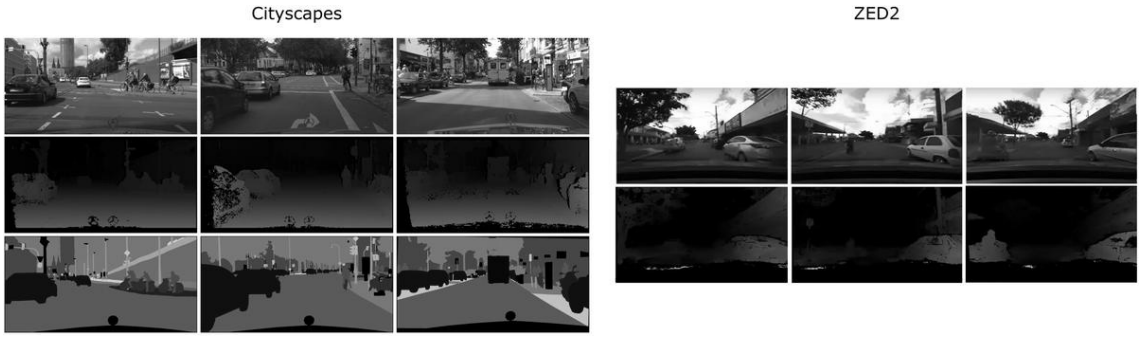

Figura 5. Exemplos de dados presentes nas bases Cityscapes e ZED2.

4.2.1. Cityscapes

A base de dados Cityscapes [Cordts et al. 2016] é amplamente utilizada na pesquisa em segmentação semântica de ambientes urbanos. Ela fornece dados 2D e de visão estéreo. além de rótulos esparsos – 5.000 imagens rotuladas com precisão e 20.000 imagens rotuladas grosseiramente. Quanto à diversidade de dados, a base provê trechos de vídeo capturados em 50 cidades alemãs, em diferentes horas do dia, estações do ano e condições climáticas. São fornecidos, ainda, mapas de disparidade referentes a cada cena.

4.2.2. ZED2

O segundo conjunto diz respeito aos dados adquiridos com o sensor ZED2 – Fig. 4. Como principais características, podemos destacar: dados captados em boas condições de visibilidade (dia, com céu limpo ou nublado) e ausência de rótulos. Assim como no caso da base de dados Cityscapes, são fornecidos mapas de disparidade.

4.3. Treinamento

Seguindo o procedimento adotado em [Varghese et al. 2021], foram empregadas 30 épocas para calibração do modelo a partir de pesos pré-treinados na base Cityscapes. Como otimizador, utilizou-se o método *Stochastic Gradient Descent* (SGD), com *momentum* $0,9$ e decaimento 10^{-4}. A taxa de aprendizado inicial foi de 10^{-2}, com uma política de atualização polinomial definida por $(1 - \frac{iter}{iters_{max}})^p$, com $p = 0,9$.

Como estratégias de aumento de dados, utilizou-se: corte (*random crop*), espelhamento (*random flip*), distorção fotométrica, normalização e preenchimento (*padding*).

Por conta das limitações de memória e processamento, adotou-se um *batch size* igual a 4. No tocante à resolução, foram utilizadas imagens com dimensões 256×256.

4.4. Inferência

A inferência é realizada em imagens com resolução 512×1024, sem técnicas de aumento de dados. Avalia-se a precisão e consistência dos modelos com base no valor médio da Intersecção sobre União (*mIoU*) e na Consistência Temporal (*TC*), respectivamente.

5. Resultados Experimentais

A seguir, são apresentados os resultados obtidos, tanto em termos de precisão (*mIoU*), quanto de estabilidade (*TC*). Avalia-se a influência do peso atribuído à função de custo auxiliar (consistência temporal) e da base de dados utilizada para o seu cálculo.

Para estudar a influência do peso atribuído à função de consistência temporal, são considerados os valores de $0,3$, $0,5$ e $0,7$. A figura 6 ilustra os resultados médios tanto em termos de precisão, quanto em termos de consistência temporal.

Para a base ZED2, observa-se um compromisso entre precisão e estabilidade, de modo que quanto maior o peso atribuído à função de custo auxiliar, maior é a consistência temporal e menor a precisão atingidas. Para a base Cityscapes, contudo, o uso de um peso maior resultou na melhoria de ambas as métricas. Finalmente, enquanto para a base Cityscapes houve melhoria da precisão em todos os casos, o uso da base de dados ZED2 só apresentou resultados satisfatórios para o menor peso utilizado.

As figuras 7 e 8 validam a análise anterior, permitindo uma avaliação mais detalhada – por classe – dos resultados. Observa-se pela figura 7 que, de modo geral, a aplicação da função de custo auxiliar promoveu o aumento da precisão das configurações envolvendo a base Cityscapes. No caso da base ZED2, por outro lado, a atribuição de importâncias maiores à função de custo auxiliar resultou na degradação do desempenho. Uma observação interessante, contudo, é que o uso da base de dados ZED2 (peso $0,3$) entregou melhoria significativa da precisão para as classes *pessoa*, *bicicleta* e *carro*. Em particular, os resultados para a classe *pessoa* até mesmo ultrapassaram os obtidos com

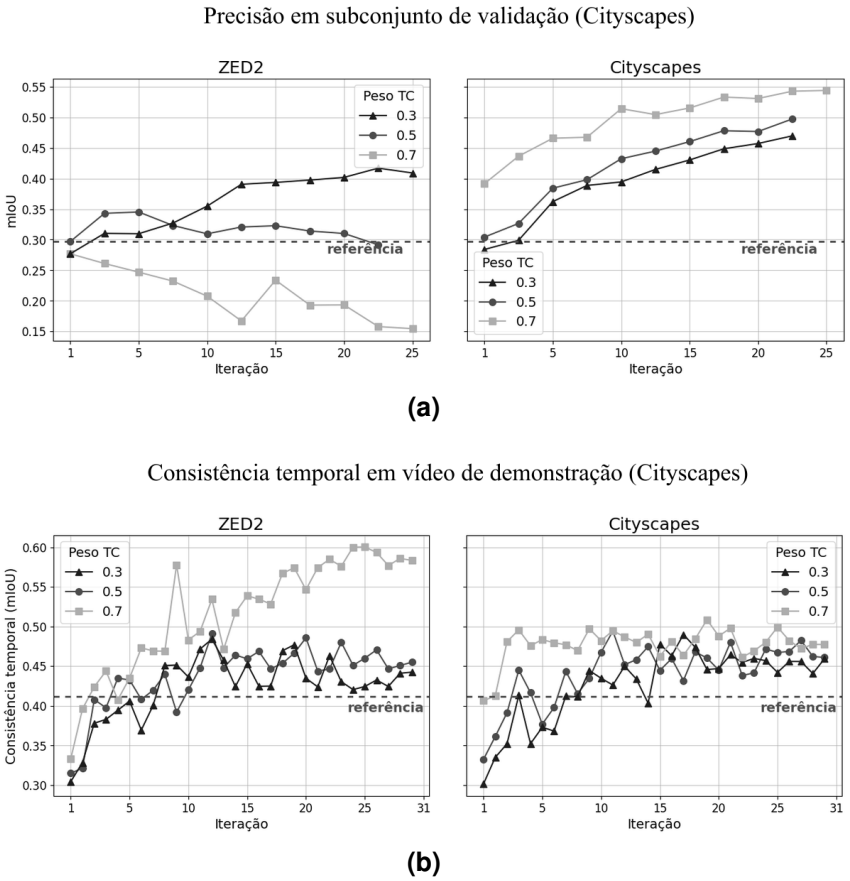

Figura 6. Resultados de (a) precisão e (b) estabilidade para as configurações testadas. A linha tracejada define o valor de referência (modelo pré-treinado).

a base Cityscapes. Tais resultados são favoráveis à configuração ZED2 $(0,3)$, dada a importância de tais classes à navegação segura em ambientes urbanos.

A análise dos resultados da figura 8 evidencia um comportamento consistente apenas para a base Cityscapes que, de forma geral, apresentou aumento da consitência temporal atrelado ao aumento do peso atribuído à função de custo auxiliar. Nos demais casos – ZED2 – houve até mesmo degradação da estabilidade do modelo quando considerando os maiores pesos. Outro resultado interessante, e que novamente destaca a configuração ZED2 (peso $0,3$), refere-se às classes *bicileta* e *pessoa*, que representam usuários vulneráveis no ambiente urbano, e para as quais os resultados em termos de estabilidade foram próximos, e até mesmo superiores (classe *pessoa*) aos das demais configurações.

Finalmente, observa-se pela figura 9 que, de modo geral, a base de dados Cityscapes promove um processo de calibragem mais robusto. Destaca-se, ainda, que apesar de ter alcançado valores de precisão menores do que os obtidos pela base Cityscapes no subconjunto de validação, a combinação da base ZED2 e do fator de ponderação $0,3$ levou a um aumento de aproximadamente 10 pontos percentuais de precisão sobre a referência.

5.1. Considerações

Uma consideração importante acerca das análises realizadas anteriormente é que, por conta de a métrica de consistência temporal ter natureza não-supervisionada, a corretude

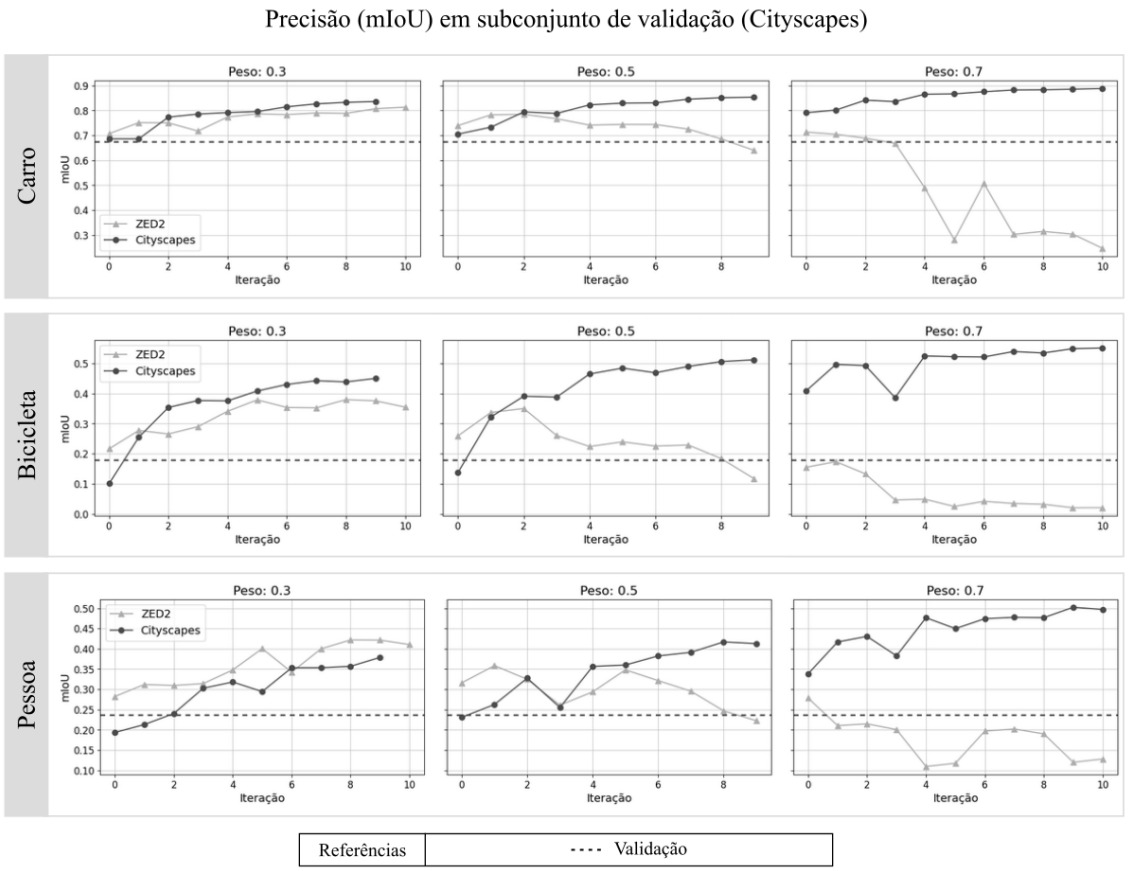

Figura 7. Precisão por classe no subconjunto de validação da base Cityscapes.

dos rótulos não tem impacto em seu cálculo, apenas a consistência entre saídas consecutivas. Dessa forma, valores elevados de consistência temporal podem advir de erros de percepção consecutivos/consistentes. Sendo assim, é de fundamental importância a análise em conjunto da precisão do modelo, de forma a garantir que tais valores de estabilidade não estejam atrelados a valores de precisão degradados.

6. Conclusão

Apesar de promoverem grandes avanços em termos de precisão e eficiência, uma limitação dos modelos de segmentação semântica atuais diz respeito à sua instabilidade. Além disso, a maior parte dos trabalhos na área emprega aprendizado supervisionado, apesar de o cenário ser de escassez de rótulos.

Motivados por tais fatos, neste trabalho estudamos o uso de técnicas de aprendizado auto-supervisionado para aprimoramento da precisão e estabilidade de modelos de segmentação leves, de modo a tirar proveito da grande disponibilidade de dados não-rotulados e das relações temporais entre os mesmos. Em particular, nossa principal contribuição frente aos demais trabalhos na área diz respeito ao estudo de diferentes bases de dados para cálculo das supervisões principal (segmentação semântica, supervisionada) e auxiliar (consistência temporal, auto-supervisionada).

Segundo os resultados, o melhor desempenho é obtido quando a mesma base usada no pré-treinamento (Cityscapes) é empregada no cálculo da supervisão auxiliar.

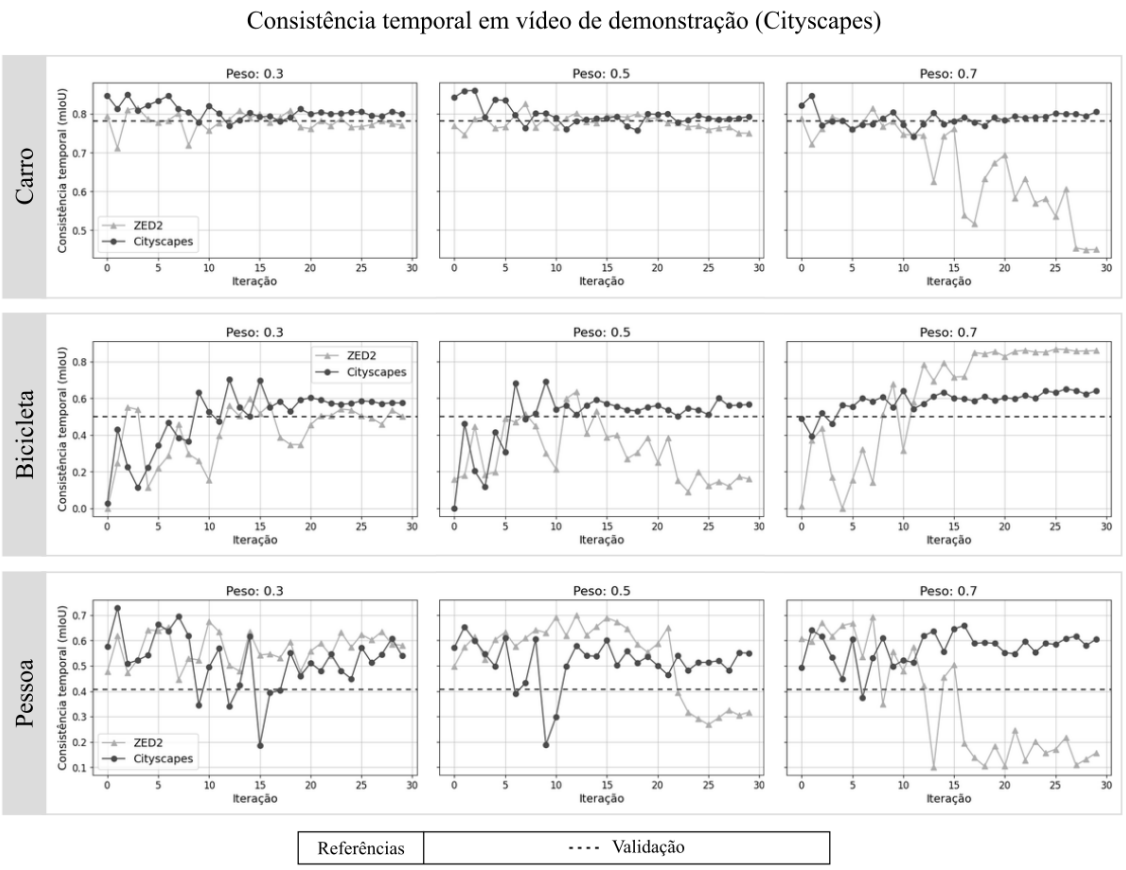

Figura 8. Consistência temporal por classe em vídeo demo na base Cityscapes.

Contudo, pode-se destacar como um resultado relevante o fato de a base de dados ZED2, combinada a pesos menores, promover tanto a precisão, quanto a estabilidade geral do modelo. Isso é uma evidência promissora de que o desempenho de um modelo não está limitado à disponibilidade de dados rotulados, ou mesmo ao montante de dados provenientes da base originalmente utilizada para treinamento. Seu aprimoramento – tanto em termos de precisão quanto estabilidade – é possível a partir de novos dados adquiridos posteriormente, em diferentes condições de captura – sensores, ângulo, iluminação.

A partir de tais conclusões, é possível ainda indagarmos se o uso de dados simulados/artificiais podem ter o mesmo efeito benéfico no aprimoramento de modelos pré-treinados, discussão esta que reservamos a estudos futuros.

Agradecimentos

Agradecemos à FUNDEP e ao projeto Rota2030 SegurAuto pelo suporte financeiro fornecido no decorrer do projeto.

Referências

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

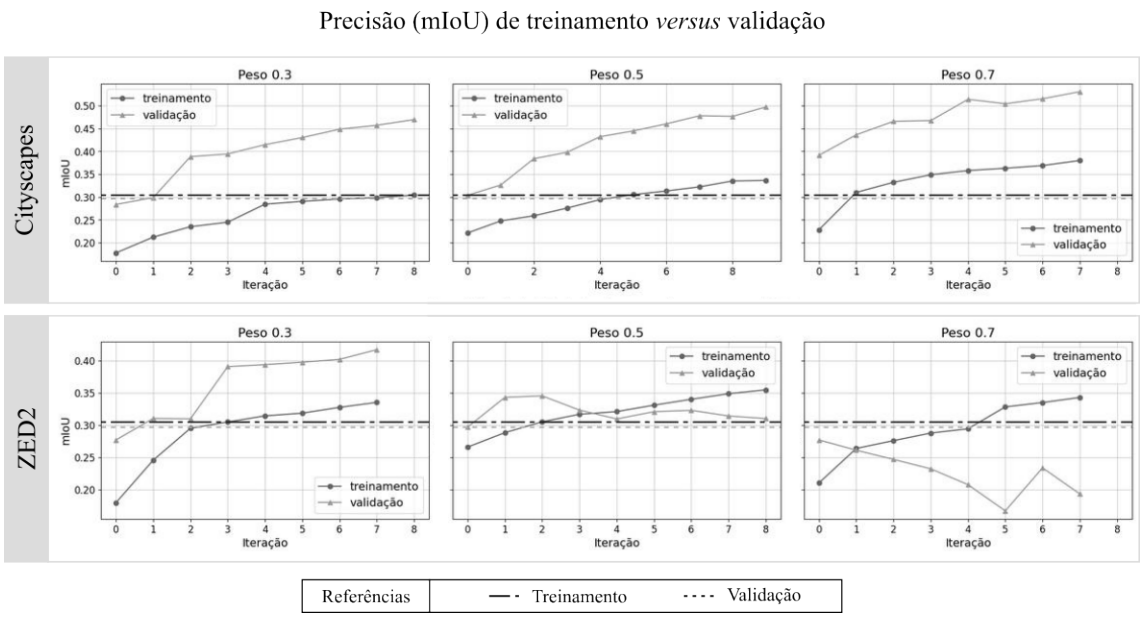

Figura 9. Evolução da precisão de treinamento e validação na base Cityscapes.

Contributors, M. (2020). MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655.

Lee, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 667–676.

Li, J., Wang, W., Chen, J., Niu, L., Si, J., Qian, C., and Zhang, L. (2021). Video semantic segmentation via sparse temporal transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 59–68, New York, NY, USA. Association for Computing Machinery.

Liu, Y., Shen, C., Yu, C., and Wang, J. (2020). Efficient semantic video segmentation with per-frame inference. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 352–368, Cham. Springer International Publishing.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 3431–3440.

Oršić, M. and Šegvić, S. (2021). Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, 110:107611.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Seyedhosseini, M. and Tasdizen, T. (2016). Semantic image segmentation with contextual hierarchical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):951–964.

Shi, W., Xu, J., Zhu, D., Zhang, G., Wang, X., Li, J., and Zhang, X. (2022). Rgb-d semantic segmentation and label-oriented voxelgrid fusion for accurate 3d semantic mapping. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):183–197.

Shinzato, P. Y. and Wolf, D. F. (2010). Statistical analysis of image-features used as inputs of an road identifier based in artificial neural networks. In *2010 Latin American Robotics Symposium and Intelligent Robotics Meeting*, pages 19–24.

Varghese, S., Bayzidi, Y., Bär, A., Kapoor, N., Lahiri, S., Schneider, J. D., Schmidt, N., Schlicht, P., Hüger, F., and Fingscheidt, T. (2020). Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1369–1378.

Varghese, S., Gujamagadi, S., Klingner, M., Kapoor, N., Bär, A., Schneider, J. D., Maag, K., Schlicht, P., Hüger, F., and Fingscheidt, T. (2021). An unsupervised temporal consistency (tc) loss to improve the performance of semantic segmentation networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 12–20.

Xie, J., Kiefel, M., Sun, M.-T., and Geiger, A. (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697.

Xiong, J., Po, L.-M., Yu, W. Y., Zhao, Y., and Cheung, K.-W. (2021). Distortion map-guided feature rectification for efficient video semantic segmentation. *IEEE Transactions on Multimedia*, pages 1–1.

Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., and Sang, N. (2021). Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 334–349, Cham. Springer International Publishing.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham. Springer International Publishing.