Segmentation Sémantique d'Images pour la Conduite Autonome basée sur la Distillation de Connaissance Auto-Attentive

Ayoub Karine, Thibault Napoléon, Maher Jridi

Segmentation Sémantique d'Images pour la Conduite Autonome basée sur la Distillation de Connaissance Auto-Attentive

Ayoub Karine[1] Thibault Napoléon[2] Maher Jridi[1]

[1]L@bISEN, Vision-AD, Yncréa Ouest, 33 Quater Chemin du Champ de Manœuvre
44470 Carquefou, France.
[2]L@bISEN, Vision-AD, Yncréa Ouest, 20 rue Cuirassé Bretagne
29200 Brest, France.

ayoub.karine@isen-ouest.yncrea.fr, thibault.napoleon@isen-ouest.yncrea.fr,
maher.jridi@isen-ouest.yncrea.fr

Résumé

Dans cet article, nous utilisons la distillation de connaissances pour réduire la complexité de calcul des réseaux de neurones et ainsi permettre leur intégration dans des systèmes portables comme les voitures autonomes. Une nouvelle méthode est proposée pour entraîner un réseau de neurones de petite taille (l'étudiant) sous la supervision d'un réseau plus grand (l'enseignant) pour la segmentation sémantique de scènes. La principale nouveauté réside dans le transfert des connaissances de l'enseignant à l'étudiant à travers un module auto-attentif extrait des cartes de caractéristiques. Ce module capture l'association sémantique présente entre les différents canaux des cartes de caractéristiques, qui est une connaissance importante pour l'entraînement de l'étudiant. Les expérimentations sur la base de données Cityscapes démontrent l'efficacité de la méthode proposée.

Mots Clef

Conduite autonome, segmentation sémantique, distillation de connaissance, réseau auto-attentif, apprentissage profond.

Abstract

In this paper, we exploit the principle of Knowledge distillation to reduce the computational complexity of neural networks for a suitable embedding in self-driving cars. A new method is proposed for training small-size neural network (student) with the supervision of a large one (teacher) for semantic scene segmentation. The main novelty consists of transferring the knowledge from the teacher to the student through a self-attention module extracted from the feature maps. This module captures the semantic association between the different feature map channels which is an effective knowledge for the student training. Experiments on Cityscapes database demonstrate the effectiveness of the proposed method.

Keywords

Autonomous driving, semantic segmentation, knowledge distillation, self-attention, deep learning.

1 Introduction

La conduite autonome est un sujet de recherche important en intelligence artificielle. Elle peut être réalisée grâce au dernières innovations dans le domaine de traitement d'images. Une des techniques utilisées est la segmentation sémantique des scènes qui consiste à attribuer une classe à chaque pixel de l'image.

Au cours de la dernière décennie, les réseaux de neurones profonds entièrement convolutionnels (FCN) [1] ont été considérés comme une solution robuste pour les tâches de segmentation sémantique. Inspirés de la conception des FCN, plusieurs méthodes plus profondes ont obtenu de bonnes performances de segmentation, telles que PSPNet [2] et DeepLab [3]. Cependant, du fait du nombre important des paramètres mis en jeu dans ces méthodes, leur usage est limité sur le calcul en périphérie de réseau (*edge computing*) ce qui ne convient pas pour les applications portables comme celles de la conduite autonome qui a besoin d'embarquer les calculs au sein même du véhicule autonome. Pour surmonter ces limitations, plusieurs approches ont été proposées pour concevoir des réseaux compacts [4]. Ces approches peuvent être divisées en quatre grandes catégories.

La première catégorie cherche à réduire l'impact du nombre d'opérations arithmétiques utilisées dans les réseaux FCN, en particulier les multiplications et les divisions. Cela est généralement réalisé en utilisant d'une part un élagage pour supprimer les paramètres redondants du modèle [5] et d'autre part la quantification [6] pour réduire encore davantage la complexité des FCN. La deuxième catégorie vise à remplacer le réseau de base (*backbone*) des architectures FCN par un plus petit, par exemple en utilisant un ResNet-18 plutôt qu'un ResNet-152. Dans la

troisième catégorie, des réseaux de segmentation spécifiques aux applications temps réel sont proposés tels que ENet [7], ERFNet [8], ESPNet [9] ou ICNet [10]. Bien que ces trois catégories soient efficaces en termes de temps de calcul, elles produisent des performances de segmentation médiocre, car les réseaux compacts sont entraînés indépendamment des réseaux plus lourds. Pour surmonter ces limites, la quatrième catégorie investigue le principe de la distillation de connaissances (KD : *Knowledge Distillation*) qui a fait ses preuves en classification d'images [11, 12].

Le reste de l'article est organisé comme suit : un état de l'art sur la distillation de connaissances pour la segmentation sémantique est présenté dans la section 2, ensuite le principe de la méthode proposée est décrit dans la section 3, la section 4 rapporte les résultats expérimentaux, enfin, la section 5 propose une conclusion de l'article.

2 Distillation de connaissances

L'idée principale de la distillation de connaissances est d'utiliser les connaissances (par exemple les probabilités douces) d'un modèle enseignant (lourd) pour superviser l'entraînement d'un modèle étudiant (compact). Les travaux précédents [13, 14] montrent que l'adaptation des méthodes à base de distillation de connaissances utilisées en classification ne permet pas d'obtenir des résultats satisfaisants pour la segmentation sémantique. Cependant, étant donné que la méthode a montré un potentiel important [15], le développement de méthodes de distillation de connaissances adaptées à la segmentation sémantique a suscité un intérêt particulier. Le défi est de trouver un compromis entre la qualité de la segmentation et ses performances computationnelles. Le type de connaissances à distiller devrait, contrairement à la tâche de classification, modéliser les informations de contexte. Dans ce sens, Xie *et al.* [16] ont proposé deux types de connaissances calculées à partir de la sortie des logits : d'ordre zéro et de premier ordre. Le premier représente les probabilités de chaque pixel tandis que pour le second une différence entre le pixel central et son voisinage, avec une connexité d'ordre 8, est calculée. D'autres méthodes, quant à elles, utilisent la connaissance des cartes de caractéristiques intermédiaires. Dans les travaux de [17] on retrouve un auto-encodeur pour compresser la connaissance de l'enseignant. La différence entre les connaissances de l'enseignant et de l'étudiant est alors obtenue grâce à une carte d'affinité. Lieu *et al.* [13] utilisent de leur côté deux distillations de connaissance structurées (SKD : *Structured Knowledge Distillation*) : une distillation par paires et une distillation globale. La première calcule une similarité en utilisant un graphe d'affinité, tandis que la seconde, la distillation globale, vise à forcer l'étudiant à générer des sorties similaires à celles de l'enseignant via un réseau antagoniste génératif (GAN : *Generative adversarial networks*). La variation de la caractéristique intra-classe (IFVD : *Intra-class Feature Variation Distillation*) des pixels ayant la même étiquette est employée dans [14] et le transfert de connaissance se fait à travers l'utilisation d'une distance en cosinus. Récemment, Shu *et al.* [18] proposent de normaliser les cartes d'activation de chaque canal et cherchent à minimiser la divergence de Kullback-Leibler (KLD) entre elles. Inspirés par [13], les auteurs de [14] et [18] ont utilisé la distillation globale. Fen *et al.* [19] conçoivent un schéma de distillation composé de deux connaissances basées sur les similarités : une dans la dimension des pixels et une dans celle des catégories. Le point commun de ces méthodes est l'utilisation de la perte liée à la segmentation entre une image segmentée et la vérité terrain qui lui est associée, en plus des schémas de distillation qui sont proposés. Cependant, ces méthodes ne prennent pas en compte la dépendance entre les canaux des cartes de caractéristiques, qui est pourtant une information importante pour les tâches de segmentation, car chaque carte de canal représente une réponse spécifique à une classe.

3 Méthode proposée

Dans cet article, nous proposons d'utiliser une distillation auto-attentive pour la segmentation sémantique. L'idée principale est d'inciter l'étudiant à imiter les interdépendances entre les canaux des cartes de caractéristiques de l'enseignant. Cela est réalisé en mettant à jour les cartes de caractéristiques des réseaux étudiant et enseignant à travers une somme pondérée de toutes les cartes des canaux. Aussi, nous introduisons l'erreur quadratique moyenne entre les deux modules d'auto-attention comme une perte à minimiser dans le processus d'apprentissage.

3.1 Vue d'ensemble

Un aperçu de la méthode proposée est illustré à travers le diagramme présenté dans la figure 1. L'enseignant et l'étudiant partagent la même architecture. Cependant, le réseau de base de l'enseignant est plus profond que celui de l'étudiant. Nous soulignons que le réseau enseignant est figé tandis que celui de l'étudiant est entraîné avec le module auto-attentif proposé, la distillation de connaissances et la vérité terrain associée à la segmentation.

3.2 Module auto-attentif

Supposons que $A \in \mathbb{R}^{C\times H\times W}$ est une carte de caractéristiques extraite de l'enseignant ou de l'étudiant, où C est le nombre de canaux, H et W représentent respectivement la hauteur et la largeur. Tout d'abord, A est réarrangée en $A \in \mathbb{R}^{C\times N}$ avec $N = H \times W$. Ensuite, A est multipliée par sa transposée. La carte d'attention des canaux $X \in \mathbb{R}^{C\times C}$ est obtenue en appliquant la fonction softmax :

$$x_{ji} = \phi(A_i \cdot A_j) = \frac{\exp(\frac{A_i \cdot A_j}{\mathcal{T}})}{\sum_{i=1}^{C} \exp(\frac{A_i \cdot A_j}{\mathcal{T}})} \qquad (1)$$

où x_{ji} représente l'impact du $i$ème canal sur le $j$ème canal dans A et $\mathcal{T}$ est la température. Ensuite, une multiplication de matrices est effectuée entre A et la transposée de X. Enfin, les résultats sont réarrangés en $\mathbb{R}^{C\times H\times W}$, multipliés

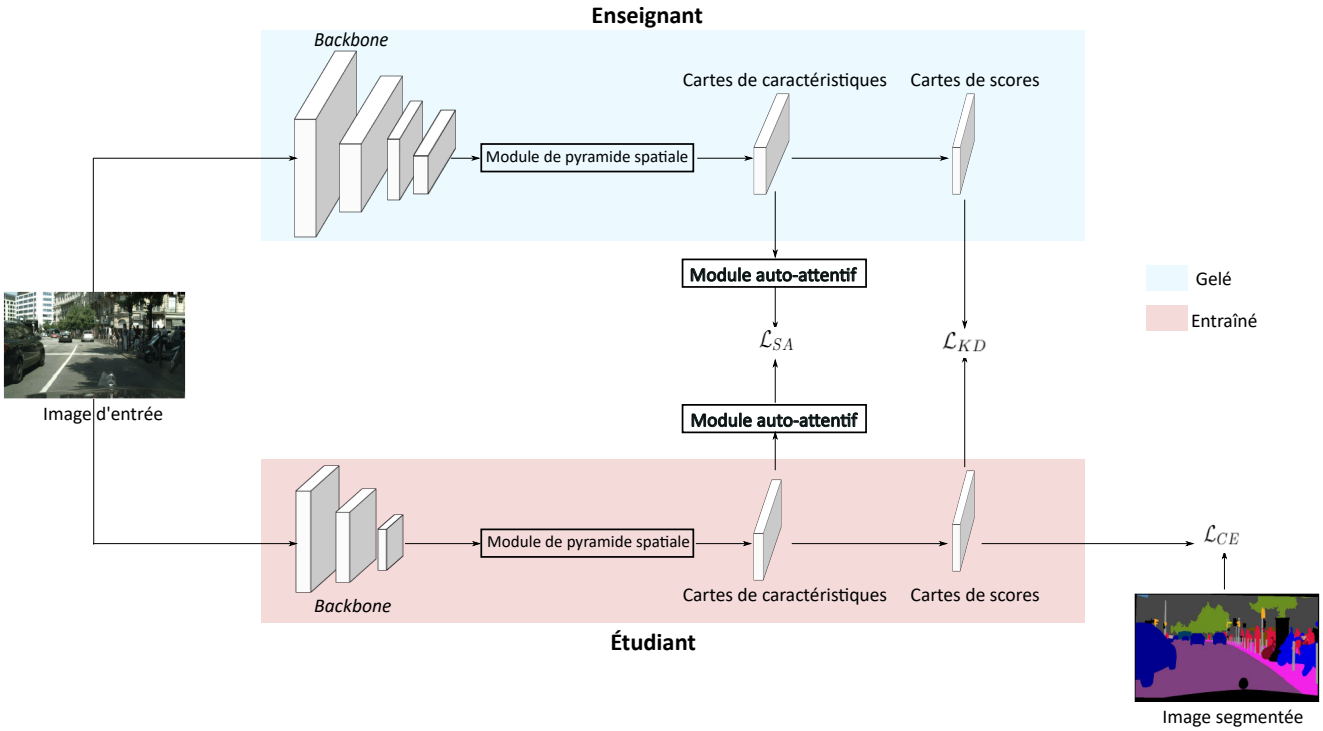


FIGURE 1 – Diagramme de la méthode proposée. Le module d'auto-attention est utilisé pour capturer les relations entre deux canaux quelconques. Ensuite, nous appliquons le transfert de connaissances de ce module, de l'enseignant vers l'étudiant. Aussi, une fonction de perte pour la distillation de connaissances, basée sur la divergence de Kullback-Leibler, et l'entropie croisée, sont incluses dans la perte totale pour aider l'étudiant à imiter l'enseignant.

par un paramètre d'échelle β et additionnés à A pour obtenir la nouvelle carte de caractéristiques $E_j \in \mathbb{R}^{C\times H\times W}$ pour le canal associé à A_j :

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{2}$$

L'équation 2 montre que les interdépendances entre les différents canaux de A sont obtenues par une somme pondérée de tous les canaux et de la carte de caractéristiques d'origine.
L'objectif de la distillation proposée est d'inciter l'étudiant à imiter les informations d'interdépendance entre canaux de l'enseignant. Par conséquent, la perte entre les nouvelles cartes de caractéristiques de l'enseignant E_j^T et de l'étudiant E_j^S devrait être minimisée. Pour cela, nous utilisons la perte quadratique moyenne ($\mathcal{L}_2$) :

$$\mathcal{L}_{SA} = \frac{1}{C} \sum_{j=1}^{C} \|E_j^T - E_j^S\|_2 \tag{3}$$

3.3 Optimisation

Pour entraîner le réseau étudiant sous la supervision du réseau enseignant, la fonction de perte totale suivante est calculée à chaque *epoch* :

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{SA} + \lambda_2 \mathcal{L}_{KD} + \mathcal{L}_{CE} \tag{4}$$

où λ_1 et λ_2 sont les poids associés aux pertes, $\mathcal{L}_{CE}$ est la perte d'entropie croisée entre l'image segmentée et la vérité terrain, $\mathcal{L}_{KD}$ est la divergence de Kullback-Leibler (KLD) entre le softmax ϕ des cartes de score de l'enseignant y^T et de l'étudiant y^S [18] :

$$\mathcal{L}_{KD} = \varphi(y^T, y^S) = \frac{\mathcal{T}^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W\cdot H} \phi(y_{c,i}^T) \cdot \log\left[\frac{\phi(y_{c,i}^T)}{\phi(y_{c,i}^S)}\right] \tag{5}$$

4 Expérimentations

Dans cette section, nous présentons d'abord la configuration expérimentale, à savoir, les jeux de données utilisés, les détails de mise en œuvre et les métriques pour l'évaluation. Enfin, nous présentons les résultats de notre nouvelle approche de distillation de connaissances pour la segmentation sémantique au regard des méthodes d'état de l'art.

4.1 Configuration expérimentale

Base de données. Pour évaluer notre méthode, nous utilisons la base de données Cityscapes [20]. Il s'agit d'un jeu de données complexe de grande envergure pour la compréhension sémantique des scènes urbaines, collectées dans 50 villes d'Allemagne. Cette base de données est principalement utilisée pour les méthodes liées à la conduite autonome. Elle est composé de 19 classes et d'une résolution d'image de 2048×1024 pixels. La base de données

contient 5000 images réparties comme suit : 2975 images pour d'entraînement, 500 images pour la validation et 1525 images pour le test.

Détails de l'implémentation. Le but de notre approche de distillation de connaissances est de transférer les capacités de segmentation d'un réseau très profond à un réseau plus léger. L'architecture de segmentation PSPNet [2] est fréquemment utilisée dans l'état de l'art, car elle propose un module de pyramide spatiale (*Pyramidal Pooling Module* - PPM) qui permet d'extraire à la fois les informations locales et globales du contexte, ce qui améliore la représentation des caractéristiques pour une meilleure segmentation sémantique. Ainsi, nous choisissons d'utiliser une architecture de segmentation PSPNet [2] aussi bien pour le réseau de l'enseignant (lourd) que pour celui de l'étudiant (compact). Comme *backbone*, nous choisissons le réseau ResNet [21], qui est connu pour offrir plusieurs profondeurs et donc plusieurs complexités d'inférence. Ainsi, nous choisissons pour notre processus de distillation de connaissances un ResNet-101 pour l'enseignant et un ResNet-18 pour l'étudiant respectivement.

Les valeurs des différents paramètres des équations 1, 2 et 4 sont fixées à travers une recherche de type *grid-search*. Les performances optimales de segmentation sont obtenues avec les valeurs suivantes :

- La température $\mathcal{T}$ dans l'équation 1 fixée à 4
- Le paramètre β dans l'équation 2 fixé à 0.4
- Les poids (λ_1, λ_2) des pertes dans l'équation 4 fixés à 14 et 3.

Nous soulignons que notre méthode est implémentée en utilisant la bibliothèque `PyTorch`.

Métrique. Pour étudier la robustesse de la méthode proposée et ses performances par rapport à l'état de l'art, nous utilisons la mesure de l'intersection sur l'union (IoU) :

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

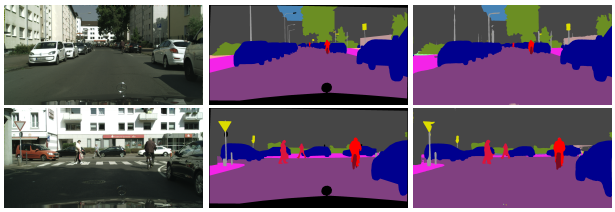
Cette métrique est bien adaptée à la comparaison des masques de classe de la vérité terrain A avec ceux prédits B en vérifiant leur chevauchement. Formellement, nous rapportons la moyenne de l'intersection sur l'union ($mIoU$) de toutes les classes en tant que mesure de performances globale de l'approche de distillation que nous proposons.

4.2 Résultats expérimentaux

Pour démontrer l'efficacité de la méthode proposée, nous avons comparé dans la table 1 notre architecture avec et sans utilisation du module d'auto-attention (SA). Comme indiqué dans cette table, les performances de segmentation sont améliorées de 1.07. Cela indique donc que le module d'auto-attention ajouté aide efficacement le réseau compact à extraire de meilleures caractéristiques.
Nous rapportons dans la table 2 une comparaison de la méthode proposée avec différentes méthodes de segmentation et de distillation de l'état de l'art. Plusieurs remarques peuvent être évoquées à la lecture de cette table. Étant donné que ENet [7] et ESPNet [9] sont les modèles les plus légers avec une taille de 0.4 Mégaoctets (Mo), ils obtiennent des résultats peu performants pour la segmentation, 58.3% et 60.3% respectivement. En revanche, PSPNet-R101 (PSPNet avec ResNet101) [22] donne de bonnes performances en segmentation avec 78.5% pour la $mIoU$. Cependant, il nécessite $70.43M$ de paramètres pour l'entraînement. Lorsque l'on remplace le réseau de base *backbone* ResNet101 par un ResNet18 dans PSPNet, les performances chutent drastiquement de 78.5% à 69.1% pour l'ensemble de validation et de 78.4% à 67.6% pour l'ensemble de test. Cela est dû à la petite taille du réseau ($13.07M$) et à son entraînement séparé du réseau PSPNet-R101. Comme on peut le voir, la distillation auto-attentive proposée améliore l'étudiant d'origine (sans distillation) de 4.68% sur l'ensemble de validation et de 4% sur l'ensemble de test. De plus, l'écart $mIoU$ entre les réseaux enseignant et étudiant est réduit de 9.4% à 4.72% et de 10.8% à 6.8% respectivement. Comparée à la méthode SKD [13], la méthode proposée permet des améliorations systématiques sur l'ensemble de validation et une amélioration légère sur l'ensemble de test. Ces résultats démontrent l'efficacité de la distillation auto-attentive proposée. Des résultats qualitatifs sont illustrés dans la figure 2.

TABLE 1 – Comparaison de notre modèle sans et avec le module d'auto-attention sur l'ensemble de validation de la base de données Cityscapes.

$Val\ mIoU$ (%)	
Sans auto-attention	Avec auto-attention
72.71	73.78



(a) Image d'entrée (b) Vérité terrain (c) Notre méthode

FIGURE 2 – Résultats qualitatifs de segmentation sémantique sur le jeu de validation Cityscapes.

5 Conclusion

Dans cet article, nous avons proposé une nouvelle méthode de distillation de connaissances pour le problème de la segmentation sémantique d'images. L'objectif était de trouver un compromis entre la qualité de la segmentation et sa complexité computationnelle afin de faciliter le déploiement de ces approches dans les véhicules autonomes. Contrairement aux méthodes de distillation de connaissances existantes, qui ignorent les interdépendances entre les canaux des cartes de caractéristiques, nous avons cap-

TABLE 2 – Étude comparative de notre méthode avec les méthodes de l'état de l'art sur la base de données Cityscapes. R101 et R18 désignent respectivement Resnet101 et Resnet18 et Ens. et Etu. désignent respectivement l'enseignant et l'étudiant.

Méthode	Params (M)	$mIoU$ (%)	
		Val	Test
Comparaison avec des méthodes de segmentation			
SegNet [23]	29.5	-	57.0
ENet [7]	0.4	-	58.3
ERFNet[8]	2.1	-	68.0
ESPNet [9]	0.4	-	60.3
ICNet [10]	26.5	-	69.5
BiseNet [24]	49.0	-	74.7
RefineNet [25]	118.4	-	73.6
SwiftNet [26]	24.7	-	76.5
FCN [1]	134.5	-	65.3
Comparaison avec des méthodes de distillation			
Ens. : PSPNet-R101	70.43	78.5	78.4
Étu. : PSPNet-R18	13.07	69.1	67.6
Étud. + SKD [13]	13.07	72.7	71.4
Étu. + SA (nous)	**13.07**	**73.78**	**71.6**

turé cette information à travers un module auto-attentif. De plus, l'erreur quadratique moyenne a été utilisée pour mesurer l'écart entre les connaissances de l'enseignant et de l'étudiant. Des expériences sur la base de données Cityscapes ont démontré l'efficacité de la méthode proposée. Dans de futurs travaux, nous proposons d'appliquer notre méthode à d'autres tâches complexes de vision par ordinateur comme la détection d'objets ou la segmentation de vidéos.

Références

[1] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4) :640–651, 2017.

[2] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[4] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv :1710.09282*, 2017.

[5] Song Han, Huizi Mao, and William J Dally. Deep compression : Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv :1510.00149*, 2015.

[6] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828, 2016.

[7] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet : A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv :1606.02147*, 2016.

[8] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet : Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1) :263–272, 2017.

[9] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet : Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.

[10] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.

[11] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[12] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*, 2(7), 2015.

[13] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[14] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision*, pages 346–362. Springer, 2020.

[15] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence : A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[16] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv :1810.08476*, 2018.

[17] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019.

[18] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021.

[19] Yingchao Feng, Xian Sun, Wenhui Diao, Jihao Li, and Xin Gao. Double similarity distillation for semantic image segmentation. *IEEE Transactions on Image Processing*, 30 :5363–5376, 2021.

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[22] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet : Object context network for scene parsing. *arXiv preprint arXiv :1809.00916*, 2018.

[23] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12) :2481–2495, 2017.

[24] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet : Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet : Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[26] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019.