





## Article

# FASSD-Net Model for Person Semantic Segmentation

Luis Brandon Garcia-Ortiz <sup>1,\*</sup>, Jose Portillo-Portillo <sup>1</sup>, Aldo Hernandez-Suarez <sup>1</sup>, Jesus Olivares-Mercado <sup>1</sup>, Gabriel Sanchez-Perez <sup>1</sup>, Karina Toscano-Medina <sup>1</sup>, Hector Perez-Meana <sup>1</sup> and Gibran Benitez-Garcia <sup>2</sup>

<sup>1</sup> Instituto Politecnico Nacional, ESIME Culhuacan, Mexico City 04440, Mexico; jportillo@ipn.mx (J.P.-P.); alhernandezsu@ipn.mx (A.H.-S.); jolivares@ipn.mx (J.O.-M.); gasanchezp@ipn.mx (G.S.-P.); ltoscanom@ipn.mx (K.T.-M.); hmperez@ipn.mx (H.P.-M.)

<sup>2</sup> Department of Informatics, The University of Electro-Communications, Chofu-shi 182-8585, Japan; gibran@ieee.org

\* Correspondence: lgarciao1400@alumno.ipn.mx

**Abstract:** This paper proposes the use of the FASSD-Net model for semantic segmentation of human silhouettes, these silhouettes can later be used in various applications that require specific characteristics of human interaction observed in video sequences for the understanding of human activities or for human identification. These applications are classified as high-level task semantic understanding. Since semantic segmentation is presented as one solution for human silhouette extraction, it is concluded that convolutional neural networks (CNN) have a clear advantage over traditional methods for computer vision, based on their ability to learn the representations of appropriate characteristics for the task of segmentation. In this work, the FASSD-Net model is used as a novel proposal that promises real-time segmentation in high-resolution images exceeding 20 FPS. To evaluate the proposed scheme, we use the Cityscapes database, which consists of sundry scenarios that represent human interaction with its environment (these scenarios show the semantic segmentation of people, difficult to solve, that favors the evaluation of our proposal), To adapt the FASSD-Net model to human silhouette semantic segmentation, the indexes of the 19 classes traditionally proposed for Cityscapes were modified, leaving only two labels: One for the class of interest labeled as person and one for the background. The Cityscapes database includes the category “human” composed for “rider” and “person” classes, in which the rider class contains incomplete human silhouettes due to self-occlusions for the activity or transport used. For this reason, we only train the model using the person class rather than human category. The implementation of the FASSD-Net model with only two classes shows promising results in both a qualitative and quantitative manner for the segmentation of human silhouettes.

**Keywords:** semantic segmentation; person class; deep learning; human silhouette; cityscapes



check for updates

**Citation:** Garcia-Ortiz, L.B.; Portillo-Portillo, J.; Hernandez-Suarez, A.; Olivares-Mercado, J.; Sanchez-Perez, G.; Toscano-Medina, K.; Perez-Meana, H.; Benitez-Garcia, G. FASSD-Net Model for Person Semantic Segmentation. *Electronics* **2021**, *10*, 1393. <https://doi.org/10.3390/electronics10121393>

Academic Editors: Athanasios Voulodimos and Fabio Grandi

Received: 7 May 2021

Accepted: 3 June 2021

Published: 10 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

There are many high-level computer vision tasks which relay in human detection in video sequences, such as intelligent video surveillance. The objective of an intelligent video surveillance system (IVVS) is to efficiently detect an interesting event from a large number of videos in order to prevent dangerous situations [1], where the main interest, of course, is the normal and abnormal behavior of human beings [2]. The applications of IVVS is becoming more specific, e.g., environmental home monitoring related to human activities, such as remote monitoring and automatic fall detection for elderly people at home [3]. Another main application for IVVS is video storage and retrieval, where the surveillance system may be prone to record the video if human beings are in the scene, saving time, data storage and, of course, resources. Nowadays, another major application for high-level computer vision and IVSS is Human Computer Interface (HCI), of which identity recognition and specifically, human identification is based on gait analysis. Although the applications seem to be very broad, all of them share the same issues, i.e., human detection

in video, human pose estimation, human tracking analysis and understanding of time series data [4]. To address the task of human detection in video sequences, the researchers first split the video sequences considering each frame in the video and performing different approaches; for example, image classification, for which the main objective is to assign one or more category labels for a whole image, which results in identifying which objects exist in the image under analysis, e.g., semantic concepts such as a person, car, road, and building are detected, but without the locations of the objects in the image.

In order to obtain the region of the objects, the next approach, i.e., the object detection, assigns the category labels and also locates the objects with annotated rectangles in the images; nevertheless, the rectangles may contain some pixels belonging to other classes or the background. However, to achieve more specific and meaningful results, this research proposes another approach named semantic segmentation. The main objective is to assign each pixel a predefined category label and in consequence, partition each object region from the background region. Although there are many studies using different predefined labels or classes, the present work focuses on two labels—human and background; the idea is using this model in future works to extract specific features for human activities comprehension and human identification. To address the main issues in this high-level semantic comprehension of human activities based on video surveillance, it is necessary to extract the foreground (human silhouettes) from the scene, at pixel level.

Since the semantic segmentation seems to be the natural solution for human silhouette extraction, we determine that the enormous success of recent state-of-the-art approaches for semantic segmentation is based on Convolutional Neural Networks (CNN). The distinctive advantage over traditional machine learning methods is the ability to learn appropriate feature representations for segmentation tasks in an end-to-end training fashion instead of using hand-crafted features that require domain expertise [5] and cannot adjust itself for an incorrect prediction [6]. Of course, there is a vast evolution and research in CNN [7], leading to different algorithms and methods focused on specific objectives, such as real time approaches, accuracy, reducing the number of parameters, computational cost, low energy consumption and storage memory.

As stated above, many high-level tasks for understanding human interaction in video sequences, are based on accurate semantic segmentation of human silhouettes; this also requires that the implementation can be executed on high-resolution images and in real time; therefore, this paper proposes the use of the novel neural network entitled FASSD-Net model [8] adapted specifically for the semantic segmentation of two classes of interest—“person” (human silhouette) and “background”—encouraging the use of human silhouettes in future applications; for example, the human identification [9] by gait analysis with a holistic approach or translating Mexican Sign Language into text.

The main contributions are summarized as follows:

- Adaptation of the FASSD-Net model for two-class semantic segmentations (“human silhouette” and “background”).
- Reduction of the computational complexity of the original FASSD-Net model that requires 45.1 GFLOPS to segment 19 classes, to 11.25 GFLOPS for two-class segmentation.

## 2. Methods and Materials

### 2.1. FASSD-Net

Although different algorithms are available for semantic segmentation, we focused on FASSD-Net (Fast and Accurate Semantic Segmentation with dilated asymmetric convolutions) [8], which was proposed as a solution to generate a semantic segmentation in real time considering a validation of urban landscapes [8].

The authors of the FASSD-Net model declare two main contributions over the baseline Harmonic DenseNet (HarDNet): the Dilated Asymmetric Pyramidal Fusion (DAPF) module, and the Multi-resolution Dilated Asymmetric (MDA) module [8].

Both modules exploit contextual information without excessively increasing the computational complexity using asymmetric convolutions. As a result, the FASSD-Net provides the following advantages:

- Reduced computational complexity allowing its use in real time applications.
- State-of-the-art result of mean intersection over union (mIoU) in the validation of urban landscapes.
- Better learning by using two different stages of the network, simultaneously refining spatial and contextual information.
- Three versions of the model, FASSD-Net, FASSD-Net-L1 and FASSD-Net-L2, to maintain a better tradeoff between speed and accuracy.

The baseline model for FASSD-Net is the FC-HarDNet-70, based on HarDNet for the task of semantic segmentation [10]. The FC-HarDNet-70 is a U-shape-like architecture [11], which is composed by five encoders blocks and four decoder blocks, all of them HarDBlock (Harmonic Dense Block), which are specifically designed to address the problem of the GPU memory traffic. In the architecture of the FASSD-NET, the following items can be found: Encoder, Decoder, MDA and DAPF modules, as shown in the Figure 1.

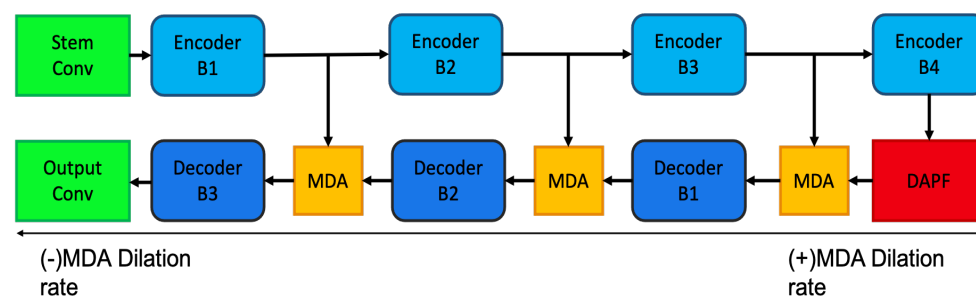


Figure 1. FASSD-NET architecture.

The DAPF module is designed to produce an increase to the receptive field of the last stage of the network (encoder), as shown in Figure 2, obtaining high-quality contextual characteristics. It is possible to change the number of pyramidal feature maps within the DAPF and thus be able to fit with the number of input feature maps, which significantly reduces the computational complexity for this module [8].

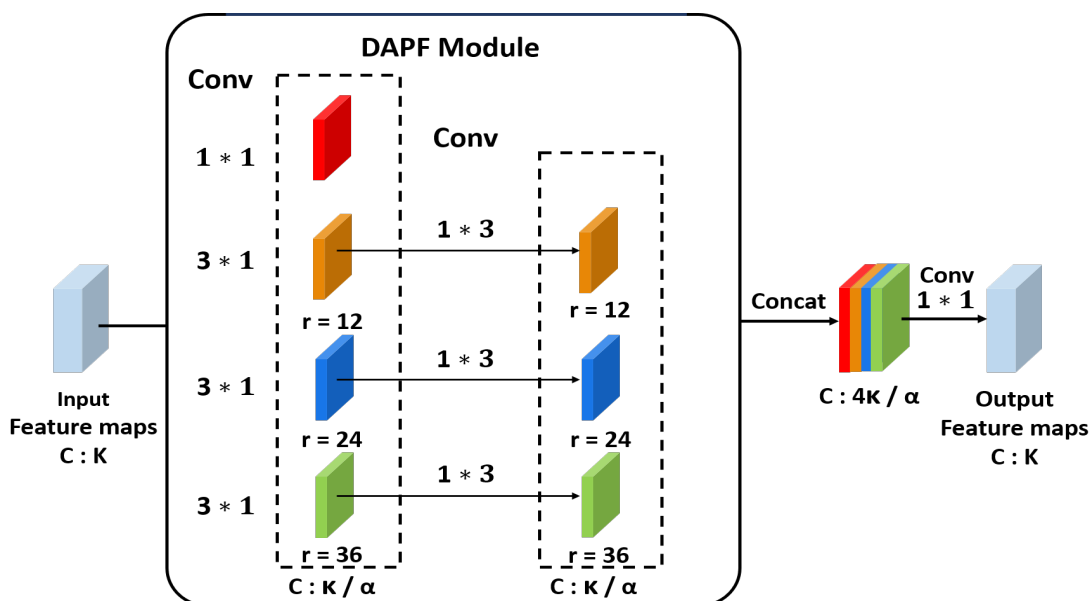


Figure 2. The DAPF Module.

The second module is the Extended Assimilation of Resolution Multiple (MDA). In this module, the feature maps, the asymmetric branch and the non-asymmetric branch, are simultaneously processed, where it seeks to exploit the contextual information from feature maps entry and retrieval of details, taking advantage of the use of dilated convolutions. In contrast, the non-asymmetric branch focuses on refining the details [8].

The FASSD-Net uses the decoder modules to recover lost information while shrinking the resolution at the encoder, in such a way that it concatenates the feature maps encoder with sampled feature maps from the decoder at each stage, to form a ladder-shaped structure. This allows the decoder at each stage to re-learn the relevant characteristics that are lost when grouped in the encoder [8].

## 2.2. Dataset

### 2.2.1. Cityscapes

The Cityscapes dataset is composed of a set of stereo video footage recorded on streets of 50 different cities. It contains 5000 images which have a high-quality pixel level, as well as about 20,000 additional images. The annotations are divided based on high frequency of occurrence within the images obtained, leaving 19 classes for evaluation [12].

The images are divided into different sets—training, validation, and testing. The images only serve as training data. The data are not divided randomly, but in a way that guarantees each division to be representative of the variability of different scenarios of street scenes. The underlying division criteria imply a balanced distribution of the geographic location and population, as well as the size of individual cities [12].

### 2.2.2. Database Pre-Processing

The Cityscape dataset has 30 visual classes, from which 19 of them are widely used for evaluation purposes. In order to fit the FASSD-Net model, the images in the Cityscapes training dataset with their respective labels are pre-processed by changing the indexes of the other 18 classes, leaving only two labels: one for background and another for the class of interest labeled as person. In addition, all those scenes with a “void” label are also assigned as background. This pre-processing stage is repeated in the Synscape database to perform the experiments [12].

Performing an analysis of the composition of classes in the Cityscapes training set, it is notorious that they have the category "Human" and that includes the rider class; however, the rider class, derived from segmentation problems, is further used in other tasks of interest, that require semantic segmentation of people in a scene. This is because people in the rider class can include drivers, passengers, or riders of bicycles, motorcycles, scooters, skateboards, horses, Segways, (inline) skates, wheelchairs, road cleaning cars, or convertibles [12].

Please note that a visible driver of a closed car can only be seen through the window. If we consider some people in the rider class, we will have instances of a class that will lack some extremity (legs), biasing the training of scenes containing people with some kind of occlusion; therefore, we avoided using the rider class.

## 2.3. Model Training

The Cityscapes database is trained with the proposed FASSD-Net network [8], modifying the original code that trains the model with a number of classes equal to 19. After customizing the network, it will be trained only with 2 classes to later evaluate it. Performance is primarily measured in intersection over union (IoU), mean intersection over union (mIoU) and frames per second (FPS). Experiments are carried out in order to improve the recognition of the person class, for which each of the databases are pre-processed, homogenizing the indices of other classes. The training is carried out with each of the aforementioned databases.

#### 2.4. Implementation Details

The implementation of the proposed FASSD-NET model is carried out with the same configuration used in [13]: PyTorch 1.0 with CUDA 10.2, the training setup use Stochastic Gradient Descent (SGD) with  $5 \times 10^{-4}$  weight drop, and 0.9 boost is used as the optimizer. We employ the "poly" learning rate strategy  $lr = (initial_{lr}) \times (\frac{iter}{total_{iter}})^{0.9}$  and an initial learning rate of 0.02. Total Cross Iter the entropy loss is calculated following the online start-up strategy [14].

We trained the model for 200,000 iterations with a size of batch 8, setting the initial learning rate at 0.02. The inference speed (in FPS) was measured on an Intel Core i9-9900K desktop with one NVIDIA GTX 2080ti. The speed was calculated from the average FPS rate of 10,000 iterations measured on images of size  $1024 \times 2048 \times 3$ . As shown in [8], considering the use of the 19 traditional classes from Cityscapes, the original FASSD-Net model, the computational complexity reported in GFLOPS is 45.1, and the required number of parameters is 2.85 M; however, the FASSD-Net model trained to segment human silhouettes, which means only two classes rather than traditional 19, has a computational complexity of 11.25 GFLOPs; the difference from the original model is due to the reduction in the number of classes. The number of parameters required by the two-class model is 2.84 M, which is relatively close to the number of parameters of the original FASSD-Net. The time required for training with a NVIDIA GTX 2080ti is up to 31 h.

#### 2.5. Methodology of Experiments

The Cityscapes database [12] is divided into 2975 training and 500 validation images with publicly available annotations, as well as 1525 test images with annotations withheld for benchmarking purposes [14–20].

As an example of our proposal, we used the Cityscape dataset [12] to measure its performance accuracy in a qualitative and quantitative manner. To train the FASSD-Net model, we used a "from scratch" approach using only Cityscapes training set images [12], and a "pretrained" approach, in which the weights are initialized for the "from scratch" approach.

### 3. Results and Discussions

This research work uses data sets belonging to Cityscapes [12] that represent human interaction with various urban landscapes with their respective labels for proofs of concept, in order to adapt the FASSD-Net model [8] to it. Once the tests were carried out, the indices of the other 18 classes were modified, leaving only two labels: one for background and another for the class of interest labeled as person. When performing an analysis of the composition of the classes in the training set From Urban Landscapes, it is noted that they have the category "Human" and that they include the rider class; however, the use of the rider class leads to segmentation problems for future uses of interest that require semantic segmentation of people in a scene. The formation of the FASSD-Net model [8] with only two classes presents promising results for the segmentation of human silhouettes, preparing the data for other applications such as human identification. That is the reason to use only the validation set to measure the accuracy of the proposed models and it is necessary to perform the pre-processing stage during the validation set, to obtain two labels only—background and person.

#### 3.1. Evaluation Methods

As a metric to evaluate semantic segmentation reported in Table 1 we use Intersection-over-Union (IoU), which is one of the most frequently used metrics. Doing the calculation by class  $IoU_c$ , with the following equation  $IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c}$  where  $TP_c$ ,  $FP_c$  and  $FN_c$ , indicates the number of true positive, false positive and false negative pixels by class [21].

After obtaining the IoU by class, the Mean IoU is calculated from the following:  $mIoU = average \times (IoU_c) \forall c \text{ in Class}$ .

**Table 1.** Per-class IoU (%) results for different models on the Cityscapes Validation set. The column of Person Class is highlighted.

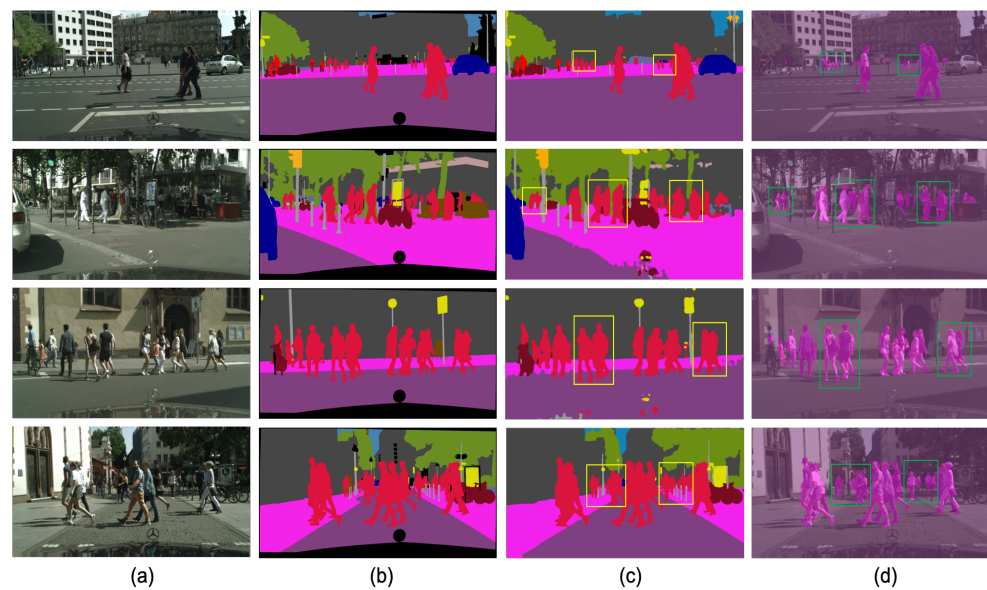
Dataset	Person IoU	Background IoU	Mean IoU
Cityscapes	79.86%	99.74%	89.80%

The Per-class IoU (%) results for different models on the Cityscapes Validation set are shown in Table 2. The column of Person Class is highlighted, the first row shows the results obtained by the ERFNet model [15] in its proposed “from scratch” training strategy where they only use Cityscapes images resulting in an IoU of 73.0% for the Person class, while the second row shows its second proposed “pre-trained” strategy, where the weights are initialized by training the network using a larger dataset such as ImageNet, obtaining an IoU of 75.2% for the Person class [15]. The third row shows the results obtained by the Fully convolutional Residual Network (FCRN) model [14], resulting an IoU of 77.1% for the person class, while the fourth row shows the results obtained by the FCRN considering the application of its online bootstrapping method, obtaining an IoU of 79.8% for the Person class [14]. Finally, row 5 shows the result obtained by the ContextNet model, with its branch a full resolution (cn124) obtaining 70.9% for the Person class. The training and evaluation conditions of the methods presented in Table 2, can be easily reproduced, allowing a fair comparison between previous models and the proposed model, obtaining, in our case, an IoU of 79.86% [16], achieving the same level of that obtained in fCRN+Bs [14] (Table 2, row 4). Although the quantitative results seem similar, in the rest of this section, a qualitative evaluation will be performed to determine the performance of the proposed method, compared to some existing ones.

**Table 2.** Per-class IoU (%) results for different models on the Cityscapes Validation set. Classes: Road, Sidewalk, Building, Wall, Fence, Pole, Traffic Light, Traffic Sign, Vegetation, Terrain, Sky, Person, Rider, Car, Truck, Bus, Train, Motorcycle and Bicycle, respectively. The column of Person Class is highlighted.

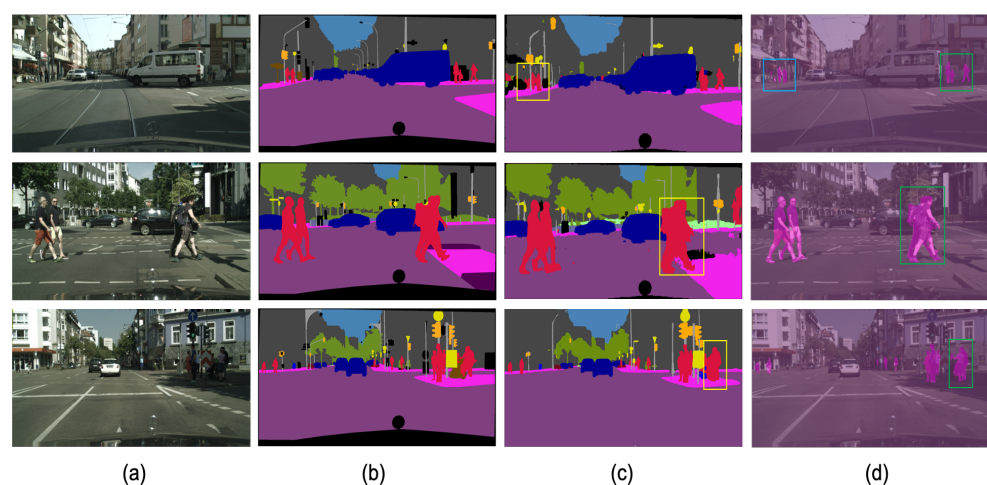
Models	Ro	Si	Bu	Wa	Fe	Po	TL	TS	Ve	Te	Sk	Pe	Ri	Ca	Tr	Bu	Tn	Mo	Bi
Scratch (VAL) [15]	97.4	80.6	90.3	55.8	50.1	57.5	58.6	68.2	90.9	61.2	93.1	<b>73</b>	53.2	91.8	59.1	70.1	66.7	44.9	67.1
Pretrained (VAL) [15]	97.5	81.4	90.9	54.6	54.1	59.8	62.5	71.6	91.3	62.9	93.1	<b>75.2</b>	55.3	92.9	67	77.4	59.8	41.9	68.4
FCRN [14]	97.4	80.3	90.8	47.6	53.8	53.1	58.1	70.2	91.2	59.6	93.2	<b>77.1</b>	54.4	93	67.1	79.4	62.2	57.3	72.7
fCRN+Bs [14]	97.6	82	91.7	52.3	56.2	57	65.7	74.4	91.7	62.5	93.8	<b>79.8</b>	59.6	94	66.2	83.7	70.3	64.2	75.5
ContextNet [16]	97.4	79.6	89.5	44.1	49.8	45.5	50.6	64.6	90.2	59.4	93.4	<b>70.9</b>	43.1	91.8	65.2	71.9	64.5	41.95	66.1

In Figure 3, column (a)—The original images; column (b)—The ground truth; the column (c)—The segmentation results obtained by model LBN-AA DASPP in the Cityscapes validation set [12]; and column (d)—Our segmentation results. The results showing an improvement in the segmentation of the human silhouette have been highlighted by rectangles in green, achieving mainly greater details in the segmentation of the limbs, i.e., legs, feet, arms, and neck. While yellow rectangles have highlighted some errors of the obtained segmentation, it can be noticed that there are regions that are segmented as the Person class, contrasting with the original image and the ground truth where these people are not observed. Yellow rectangles highlight some gross segmentation of human silhouettes.



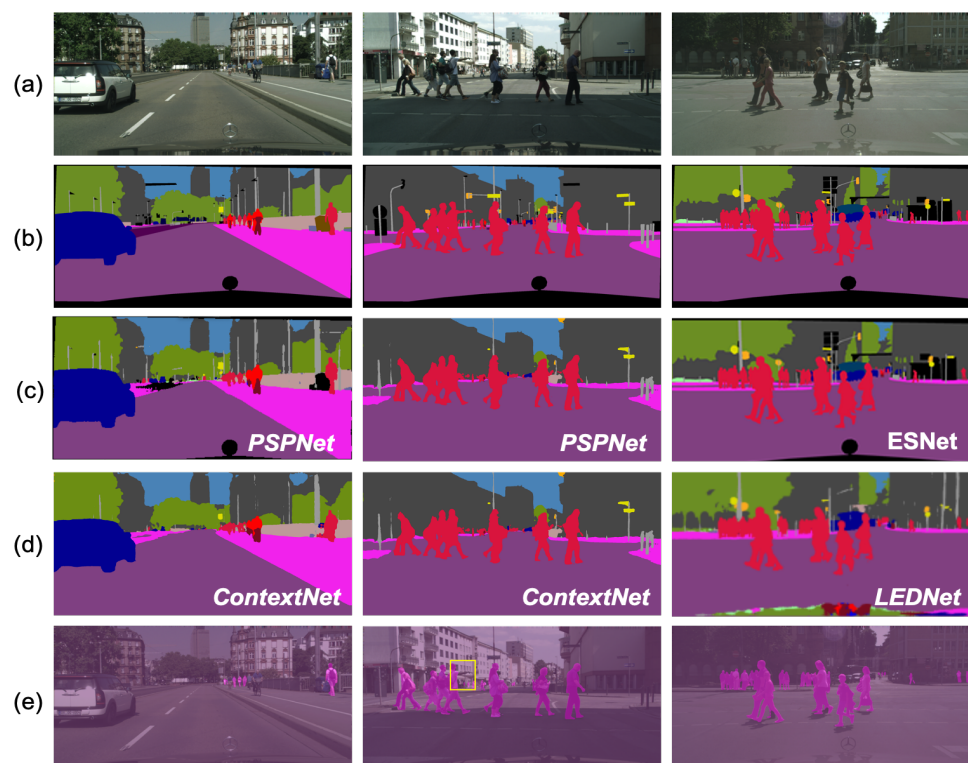
**Figure 3.** Visual results from the Cityscapes Validation set [12], from which a comparison between the predictions can be observed, made by the model LBN-AA DASPP [17]: First column (a)—Original image; second column (b)—Groundtruth; third column (c)—LBN-AA DASPP Results; fourth column (d)—Results obtained with the FASSD NeT model.

In Figure 4 column (c)—Some results obtained by the ERFNet model are presented in images of the Cityscapes validation set [12]; while the column (d)—The results obtained with our proposal. In the column Results, some green regions have been framed by rectangles where the segmentation quality obtained by our algorithm exceeds the segmentation obtained by the ERFNet model. The differences in segmentation results are observed, involving better segmentation details, specifically in the extremities, i.e., leg–foot separation, arm–torso separation, in general a more detailed segmentation. In the first row, the column (d) of Figure 3 in our results, an error segmentation can be observed, which is highlighted by a blue rectangle, because in the ground truth, there is no human silhouette is labeled in the region in which our model presents a positive result. However, a thorough analysis, as shown in Figure 6, explains that there are actually two human silhouettes in difficult lighting conditions, so it was even a challenge to the manual labeling process for tag assignments in Ground Truth.



**Figure 4.** Visual results in the Cityscapes Val set [12] in which we have a comparison with the predictions made by the ERFNet model.

Figure 5 presents the segmentation results obtained with models ESNet [22], PSPNet [23], LEDNet [24], ContextNet [16] and our models, respectively applied to the same images, achieving a better qualitative comparison of the results. Row (a)—The original images to use and Row (b)—The ground truth. The segmentation results for the original image in the first column are listed below: The first column of row (c)—The result obtained using the model PSPNet [23]; The first column of row (d)—The result obtained using ContextNet [16] and the first column of row (e)—The result with our model. Our model is not able to recognize the person who is riding a bicycle as human, and fails in the detection of the person behind the bicycle. Although it is important to remember that we intentionally left the rider class out of the training of our model, this is why it presents problems to segment human silhouettes on bicycles or motorcycles. Considering the original image in the second column, we present its segmentation results: in row (c) second column shows the obtained result using PSPNet [23]; in row (d) second column shows ContextNet [16] and in Row (e) second column shows the segmentation results of our model. The results observed for the original image of the second column show that our model presents a better and more detailed segmentation of people, even in distant human silhouettes. However, it still presents some obvious errors in the extremities in all models, particularly the arm of some of the silhouettes framed in a yellow rectangle in our model. Finally, the segmentation results for the original image of the third column are presented: the third column row (c) column shows the segmentation results obtained using the ESNet [22] model; the third column row (d)—The segmentation results using the LEDNet model [24] and the third column row (e)—The results obtained with our model. It can be observed that in general terms, the application of a model specifically trained to detect people performs a finer detection of human silhouettes than its counterpart that uses multiple classes for training.

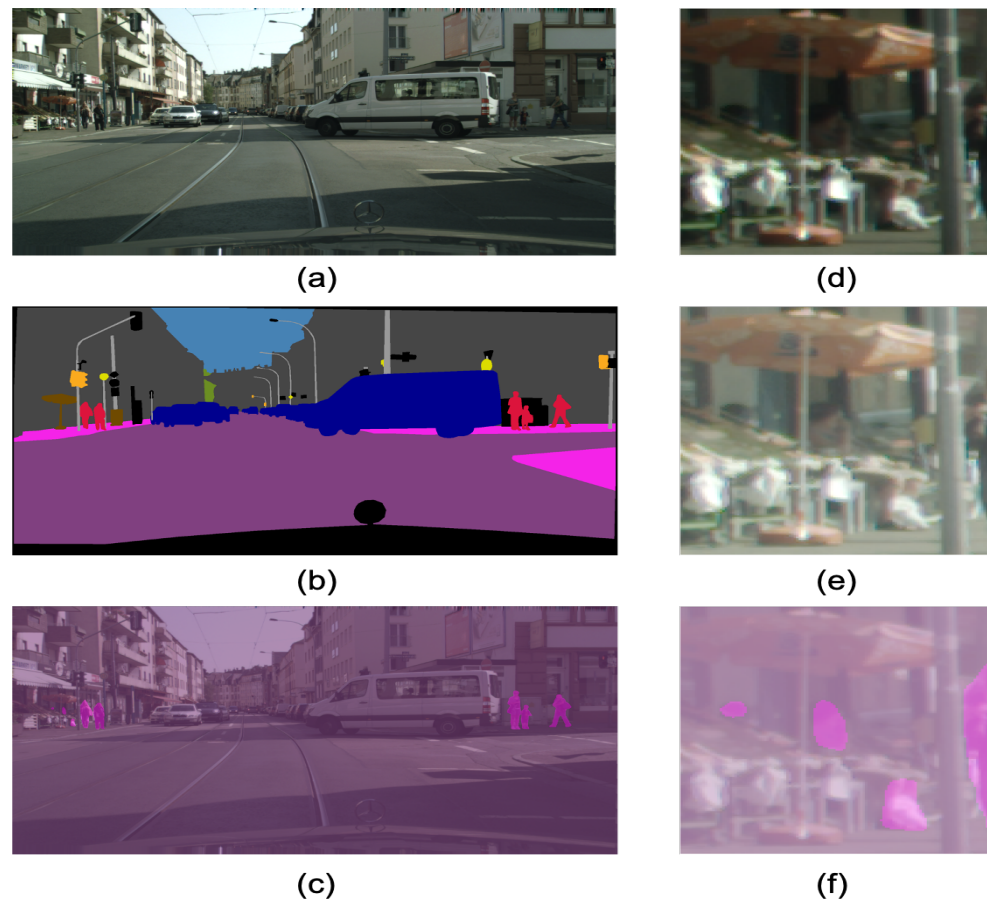


**Figure 5.** Visual results in the Cityscapes validation set [12] were compared with segmentation results for ESNet.

The Figure 6 shows—in the first row first column (a)—the original image of the validation set. The second row first column shows the ground truth (b), and the third row first column (c) shows the result obtained with the proposed FASSDNet model [8]. The first row second column (d) shows the original section where our model marks the existence of



a human being, by making an improvement in the brightness of the image, as shown in the second row second column (e); where the silhouette of a human being sitting and also a head are observed, both being detected by the proposed model in third row second column (f). Although these silhouettes are not recorded in the ground truth, this may result in a decrease in the results obtained, even if the identification was correct.



**Figure 6.** Important case where the ground truth does not contain a specific instance of human silhouette.

#### 4. Conclusions and Future Work

The FASSD-Net model trained only with two classes, presents promising results for human silhouette segmentation, preparing the data for further applications such as Human Identification and automatic fall detection for elderly people at home. The evaluation results show that the proposed scheme using the FASSD-Net provides better results, quantitatively as well as qualitatively, compared to other previously proposed schemes. Furthermore, it may be more accurate if the number of images used in the training model is bigger. Nevertheless, the annotation of the data is a problem, such that the database must be homogenized for training the data. Therefore, in future work, it will be possible to obtain a new training model using more databases, in addition to using the segmented images for other high-level semantic understanding tasks such as person identification, analysis of the march and extraction of ideogram characteristics belonging to the Mexican Sign Language.

**Author Contributions:** L.B.G.-O., G.B.-G. and J.P.-P. developed the proposed model and carried out the analysis of the final results. J.O.-M., A.H.-S. and G.S.-P. investigated others models that evaluated the performance of the proposed model. K.T.-M. and H.P.-M. development and analysis of the database, whose results are presented in the evaluation results' sections. All authors participated in the write-up and review of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project has received funding from the Secretaría de Educación, Ciencia, Tecnología e Innovación.

**Acknowledgments:** We thank the Instituto Politécnico Nacional for support the present research and provide the resources from Sección de Estudios de Posgrado e Investigación. We also express our gratitude to Secretaría de Educación, Ciencia, Tecnología e Innovación promote the development of knowledge in this matter.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Mabrouk, A.B.; Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* **2018**, *91*, 480–491. [[CrossRef](#)]
2. Han, H.; Ma, W.; Zhou, M.C.; Guo, Q.; Abusorrah, A. A Novel Semi-supervised Learning Approach to person Re-Identification. *IEEE Internet Things J.* **2020**, *8*, 3042–3052. [[CrossRef](#)]
3. Koshmak, G. Remote Monitoring and Automatic Fall Detection for Elderly People at Home. Ph.D. Thesis, Mälardalen University, Västerås, Sweden, 2015.
4. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors* **2019**, *19*, 1005. [[CrossRef](#)] [[PubMed](#)]
5. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J. Recent advances in convolutional neural networks. *Pattern Recogn.* **2018**, *77*, 354–377. [[CrossRef](#)]
6. Sultana, F.; Sufian, A.; Dutta, P. Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. In *Knowledge-Based Systems*; Jones & Bartlett Publishers: Burlington, MA, USA, 2020; pp. 201–202.
7. Xia, Y.; Yu, H.; Wang, F.Y. Accurate and robust eye center localization via fully convolutional networks. *IEEE/CAA J. Automat. Sin.* **2019**, 1127–1138. [[CrossRef](#)]
8. Rosas-Arias, L.; Benitez-Garcia, G.; Portillo-Portillo, J.; Sanchez-Perez, G.; Yanai, K. Fast and Accurate Real-Time Semantic Segmentation with Dilated Asymmetric Convolutions. *ICPR 2021*, 1–8. [[CrossRef](#)]
9. Han, H.; Zhou, M.; Shang, X.; Cao, W.; Abusorrah, A. KISS+ for rapid and accurate person re-identification. *IEEE Transact. Intell. Transport. Syst.* **2020**, *99*, 394–403.
10. Chao, P.; Kao, C.Y.; Ruan, Y.S.; Huang, C.H.; Lin, Y.L. HardNet: A Low Memory Traffic Network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3552–3561.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
12. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
13. Rosas-Arias, L.; Benitez-Garcia, G.; Portillo-Portillo, J.; Olivares-Mercado, J.; Sanchez-Perez, G.; Yanai, K. FaSSD-Net: Fast and Accurate Real-Time Semantic Segmentation for Embedded System. In Proceedings of the ITS World Congress, T-ITS 2021, Hamburg, Germany, 11–15 October 2021.
14. Wu, Z.; Shen, C.; Hengel, A.v.d. High-performance semantic segmentation using very deep fully convolutional networks. *arXiv* **2016**, arXiv:1604.04339.
15. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transact. Intell. Transport. Syst.* **2017**, *19*, 263–272. [[CrossRef](#)]
16. Poudel, R.P.; Bonde, U.; Liwicki, S.; Zach, C. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv* **2018**, arXiv:1805.04554.
17. Dong, G.; Yan, Y.; Shen, C.; Wang, H. Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Transact. Intell. Transport. Syst.* **2020**. [[CrossRef](#)]
18. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5229–5238.
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
20. Siam, M.; Gamal, M.; Abdel-Razek, M.; Yogamani, S.; Jagersand, M.; Zhang, H. A comparative study of real-time semantic segmentation for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 587–597.
21. Han, H.Y.; Chen, Y.C.; Hsiao, P.Y.; Fu, L.C. Using Channel-Wise Attention for Deep CNN Based Real-Time Semantic Segmentation With Class-Aware Edge Information. *IEEE Transact. Intell. Transport. Syst.* **2020**. [[CrossRef](#)]
22. Wang, Y.; Zhou, Q.; Xiong, J.; Wu, X.; Jin, X. Esnet: An efficient symmetric network for real-time semantic segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Xian, China, 8–11 November 2019; pp. 41–52.

- 
23. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
  24. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2019; pp. 1860–1864.