WS-CRNN: A Hybrid Deep Learning Approach for Stereo Disparity Estimation Using Wavelet Scattering and CNN+RNN Architectures

[1*] Kheira HADJ DJELLOUL, [2] Mohamed SENOUCI, [1]Zineb HEMMAMI

[1]Electromechanic Department, Institute of maintenance and industrial safety, University of Oran 2 – Mohamed ben Ahmed, Bir el djir 31000, Oran, Algeria

[2]Department of Computer Science, University of Oran 1 Ahmed Ben Bella, El M'Naouer-31000, Oran, Algeria.

Abstract- Deep learning has become a cornerstone of modern stereo matching algorithms due to its ability to accurately model scene geometry. Disparity estimation varies depending on the specific constraints of each application: some methods rely on recurrent architectures to dynamically refine predictions by focusing on uncertain regions, while others exploit multi-scale strategies, processing images at different resolutions to capture fine details in visually complex environments.

In this context, we propose WS-CRNN, a hybrid architecture that combines the reactivity of recurrent refinement with the structural richness of multi-scale analysis. At the core of the model, the wavelet scattering transform robustly extracts both local and global features at multiple scales, while substantially reducing the dimensionality of cost volumes, thus alleviating the computational overhead typically associated with 3D convolutions.

In parallel, a recurrent neural network iteratively enhances disparity predictions through successive comparisons of left and right views, enabling precise hierarchical estimation. Compared to state-of-the-art methods, WS-CRNN achieves competitive performance while maintaining low memory and energy consumption. Overall, WS-CRNN represents a promising trade-off between algorithmic complexity and prediction quality, marking a significant step forward in the field of deep stereo vision.

Index Terms- *Deep learning, Stereo matching, Wavelet scattering, CNN, RNN, ConvLSTM, Multi-scale, Matching cost volume, Disparity map*.

1. INTRODUCTION

Human vision fundamentally differs from computer perception. The eye captures light, and the brain processes the information to interpret shapes, colors, and depth. Computer vision aims to replicate these capabilities by enabling machines to understand their environment from images. One of the major challenges in this field is 3D reconstruction from stereoscopic images, which relies on disparity estimation between two views of the same scene. It is also important to note that the quality of 3D reconstruction heavily depends on the accuracy of stereo matching. Therefore, this process is considered a complex and crucial task in computer vision.

The development of stereo matching algorithms remains one of the most challenging problems. Stereo matching has been the subject of active research, with numerous authors proposing various approaches and algorithms. Consequently, in their paper [1], H. Mohd Saad et al. presented a survey on algorithms related to stereo matching. According to this preliminary survey, two major frameworks are identified in the current development of stereo matching algorithms: traditional methods, including local and global techniques based on energy minimization, and artificial intelligence-based methods, notably convolutional neural networks (CNNs). It appears that most traditional methods are significantly less accurate than AI-based methods.

Another survey by T. Fabio, B. Luca, and P. Matteo [2] on deep stereo matching identifies several categories of architectures, each based on specific key concepts. These categories illustrate the evolution of deep stereo matching techniques and the progress made in this area.

In their study [3], Zhou K., Meng X., and Cheng B. explore both traditional and modern deep learning-based methods for stereo matching, along with the challenges associated with applying these techniques in real-world scenarios. They classified deep learning approaches into three categories: end-to-end models, Siamese networks, and Generative Adversarial Networks (GANs).

From this, we can conclude that, currently, end-to-end disparity estimation methods based on stereoscopic systems [1][2][3][4] achieve significantly superior results compared to classical methods [5].

Deep learning-based disparity estimation methods involve creating a high-dimensional (4D) cost volume [6][7][8] that covers the entire disparity range. This volume must then undergo complex filtering, typically using 3D convolutional layers, to extract correspondences and generate an accurate disparity map. However, these approaches present limitations due to their high memory consumption, computational complexity, and large number of parameters, which can slow down predictions.

To overcome these challenges, DeepPruner [9] proposes a solution by reducing the search range, which decreases the cost

volume's dimension and allows for refinement to improve accuracy. Another solution is proposed by the MSDE method [10], designed end-to-end, which uses features extracted at different scales to build cost volumes, progressively reducing the search range at higher scales.

All of the aforementioned studies have inspired us to develop a new model in this context, aiming to address some issues related to stereo matching. We propose the WS-CRNN approach, which combines two powerful techniques: wavelet scattering and the combined use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This is a multi-scale disparity estimation model, designed end-to-end, extracting features at different scales and leveraging them to generate multi-scale cost volumes.

To reduce the dimensionality of the global cost volume, we first establish a cost volume at a coarse scale considering the maximum search range. This range is then projected by bilinear interpolation to a higher scale, while generating preliminary cost volumes. To improve efficiency, a ConvLSTM is applied to the cost volume. This allows the spatial and temporal dependencies to be processed line-by-line for refining the matching result before disparity map regression.

The main contributions of this work are:

1. We describe how to use wavelet scattering for extracting spatial features at different scales from two rectified stereoscopic images, and then decompose them at various resolutions to construct a multi-resolution image pyramid.
2. Once the features are extracted, we calculate a cost volume that represents the differences between the features extracted from the left and right images.
3. After the generation of the cost volume, a ConvLSTM is applied to handle the spatial and temporal dependencies line-by-line, refining the matching result while maintaining the performance of the cost volume feature aggregation.
4. The output of the ConvLSTM enables the prediction of the disparity map at each resolution level. The predicted disparity maps will be used to learn error maps, which constitute the output of the comparison.
5. Finally, using an evaluation protocol, we analyze the performance of the approach and assess its robustness.

2. Related Work

Over the years, various algorithms and techniques have been developed for disparity estimation. These methods can be broadly classified into two categories: traditional algorithms and deep learning-based algorithms. Most traditional algorithms follow four main steps:

- Matching cost calculation
- Matching cost aggregation
- Disparity computation and optimization
- Disparity refinement

Deep learning-based algorithms, particularly convolutional neural networks (CNNs), have first altered the matching cost calculation by leveraging deep features instead of traditional ones. More recently, the four steps of classical algorithms have been replaced by end-to-end architectures that handle the entire process. Thus, learning-based methods are classified into three categories: sub-region methods, end-to-end methods, and multispectral methods [1].

The work presented in paper [1] compares prior research integrating the CNN-based approach with traditional algorithms for the standard stereo pipeline. The authors agree that the works of Zbontar and LeCun [11][12] have served as a reference for many researchers applying CNNs to the matching cost computation. Seki and Pollefeys [13] introduced SGM-Nets, a learning-based disparity estimation method, whose results surpass those of manually tuned SGM [12], confirming the usefulness of learning to enhance performance.

End-to-end learning algorithms based on CNNs are further categorized into encoder-decoder architectures and methods based on learning regularities for 3D convolutions [3]. Encoder-decoder models such as iResNet [14], DispNetC [15], and CRL [16] use two sub-networks to estimate and regularize disparity maps. However, these methods suffer from a large number of parameters and inaccurate estimation in occluded and textureless regions. To improve performance and reduce the complexity of the cost volumes, EdgeStereo [17] and DeepPruner [9] were introduced. EdgeStereo uses a shallow edge detection sub-network, while DeepPruner applies the PatchMatch method to reduce the size of the cost volumes. MCliqueNet [29] has been proposed for feature extraction in stereo disparity estimation, and its efficiency has been demonstrated. In contrast to these approaches, MSDE [10] uses the encoder-decoder architecture only in the feature extraction module while optimizing the model size through multi-scale disparity estimation and the integration of residual disparity, thus offering a more efficient and accurate solution.

Other methods based on learning regularities for 3D convolutions apply 3D convolutions to 4D cost volumes, composed of height, width, features, and disparity values [18]. Although these methods offer better performance than previous approaches [14][15][16], they require more memory and computational resources, thus increasing inference time. Among these methods are PSMNet [8], GA-Net [20], GWCNet [21], SCVNet [22], GCNet [23], and PDSNet [24]. LRCR [19] proposes an innovative end-to-end approach using two parallel LSTM networks [25], but this technique takes considerable time to estimate disparity maps.

To reduce inference time, lightweight models such as ESNet [26] and StereoNet [7] decrease the number of 3D convolutions. StereoNet uses only five convolutions to estimate an initial disparity map after downsampling the images, while AnyNet [27] applies a residual disparity map in three stages. LEAStereo [8] proposes a hierarchical neural architecture search (NAS), improving accuracy. However, all of these methods still require high-dimensional 4D cost volumes and numerous 3D convolutions, increasing their resource consumption. In comparison, the MSDE approach [10] reduces computations by leveraging low-dimensional cost volumes at different scales.

Cost volume filtering is crucial for eliminating noise from the generated volumes, but it requires significant computational resources and memory-intensive 3D convolution layers to aggregate 4D cost volumes. This is why, in the GA-Net method [20], the number of convolutions is reduced, and refinement

blocks are added to improve accuracy, although this results in a loss of speed. A method frequently used in CNNs, proposed by He K, Zhang X, Ren S, et al. [28] and adapted to ResNet, decomposes 3D convolutions into two parts: the 2D convolution for the spatial component and the 1D convolution for the temporal component. This method has been adapted by MSDE [10] to optimize cost volume filtering by replacing 3D convolutions with 2D and 1D convolutions.

In the context of the WS-CRNN method, we use the encoder-decoder structure in the feature extraction module, rather than applying it at each step of the disparity estimation pipeline. This approach allows for the design of a more compact model, thanks to multi-scale disparity estimation, improving the accuracy and efficiency of the task. The WS-CRNN model generates low-dimensional cost volumes across multiple scales, thereby reducing the computations required for filtering. Furthermore, to refine the matching results, a ConvLSTM recurrent neural network (RNN) is used to process spatial and temporal dependencies line-by-line, capturing complex relationships. This step improves the quality of disparity predictions. The output of the RNN is then used to predict the final disparity map, providing a more accurate depth estimation.

3. Architecture of the Proposed Approach

Our WS-CRNN approach generates a dense disparity map from two rectified stereo images as input, having the same dimensions, defined by height H and width W. To achieve this, we adopted a hierarchical approach aimed at establishing the correspondence between each pixel of the left image and its counterpart in the right image (and vice versa, from the right image to its counterpart in the left image).

Our architecture consists of the following steps:

1. Feature Extraction: We used wavelet scattering to extract information at different scales from the two stereo images (left and right).
2. Pyramid Network: We constructed an image pyramid from the feature maps obtained in the previous step.
3. Cost Volume (PCV – ACV): From the feature maps obtained at different levels, we computed a cost volume that represents the differences between the features of the two images. This allows evaluating potential correspondences between pixels from the left and right images.
4. Refinement by ConvLSTM: ConvLSTM is applied to capture contextual spatial dependencies, which refines and improves the pixel matching.
5. Disparity Map Prediction: The outputs of the ConvLSTM will allow for the prediction of the disparity map, indicating the positional difference between corresponding pixels in the two images.

3.1. Wavelet Scattering Network for Multi-Scale Feature Extraction

To perform precise matching between the pixels of the left and right input images, it is essential to extract unique and informative features for each pixel. Our feature extraction method relies on the use of wavelet scattering, a technique inspired by S. Mallat [30][31], which allows for extracting information at different scales. Each scale is obtained through a wavelet transform applied to the image.

The wavelet scattering (WS) approach decomposes the image at different resolutions, following a multi-resolution technique, thus enabling the capture of both local (in higher scales) and global (in lower scales) details of the image. To meet computational requirements in terms of processing time and memory space, we designed a three-level extraction model, with each level consisting of two steps:

1. **Wavelet Transform**: A wavelet convolution, which replaces traditional convolutions to extract features at each level.
2. **Wavelet Dimensionality Reduction**: A pooling process using wavelet coefficients, which reduces the image size while preserving essential information, unlike classical max pooling.

These two steps result in image downsampling, where relevant information is captured through multi-scale analysis. This process generates three feature maps at three distinct resolution levels. Each map is a fraction of the input image size ($1/2^r$), so the shape of each feature map ($H/2^{2r}$, $W/2^{2r}$), where r=1,2,3 and H and W are the height and width of the input image, respectively.

3.2. Cost Volume Estimation

The cost volume is a crucial step in stereo vision. It is generated by comparing the pixels or features of the two images at different disparities. Each level of the volume represents a disparity hypothesis. By calculating these costs for various disparities, a three-dimensional representation of the pixel costs from the two input images is obtained.

Our approach to disparity volume estimation is based on previous research [10][29], where the cost volume is constructed in three stages:

1. Grouping potential correspondence candidates.
2. Efficient aggregation of correspondence features.
3. Refining the cost volume.

Based on this strategy, we design our cost volume as illustrated in **Figure -1-** over three scales through the following steps:

1. A preliminary cost volume (PCV) for both the left and right images is generated at a coarse scale (with reduced dimensions) and a resolution of $1/2^{2r}$ (r = 3).
2. Refinement through two parallel stacked ConvLSTM networks and prediction of preliminary left and right disparities.
3. Calculation of left and right error maps.
4. Propagation of disparities and generation of adjusted left and right cost volumes at a higher level (r = 2, r = 1).

3.2.1. Preliminary Cost Volume (PCV)

After extracting the feature maps from the scattering of both the left and right images, we construct two preliminary cost volumes for the left and right images (PCV_L and PCV_R) (see **Figure -1-**) to obtain a preliminary disparity map estimate at the coarse scale (r = 3). In this step, two constraints must be considered:

1. **Aggregation Region Constraint**: The aggregation region is typically limited to a region around a central pixel, in the form of a fixed-size window. In our method, we used the image reduced by $1/2^{2r}$ (r = 3).at the coarse scale as the search area to determine correspondences.
2. **Epipolar Constraint**: This geometric constraint reduces the search space for correspondences to a single dimension, as homologous points lie along the same line, called the epipolar line. This is why we used rectified images.

Thus, these two modeled constraints allowed us to evaluate the relationships between pixels in the same plane and those in the aggregation region. To compute the similarity between the features of each pixel (x, y) in the left image and its horizontal counterpart in the right image (and vice versa, between the right image and the left image), we chose the L1 norm (Manhattan Distance) [33], which measures the sum of the absolute differences between the corresponding pixel features. This choice is relevant in our case, as the L1 norm computes the distance by following the coordinate axes, as if moving strictly along these axes.

This processing generates a 2D disparity map for each disparity value. These 2D maps are then concatenated to create a 3D volume. Once the PCV_L and PCV_R are generated, a preliminary left disparity map (D_L) and a preliminary right disparity map (D_R) are predicted, after a high-performance ConvLSTM processing and disparity refinement, to improve precision and eliminate errors caused by information loss due to the reduction of the input images.

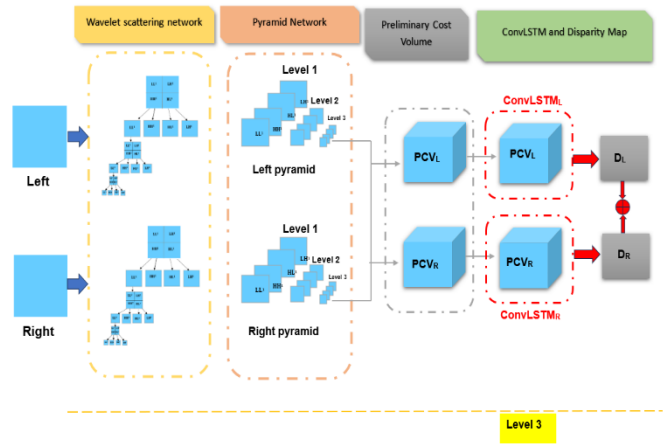

Figure -1- The process of generating the left PCV and right PCV at scale 3 after feature extraction using wavelet scattering and the construction of a pyramidal network. Disparity refinement is achieved through parallel ConvLSTM on the PCVs.

3.2.2. Adjusted Cost Volume

Given that the left (PCV_L) and right (PCV_R) disparity maps have been carefully refined by ConvLSTM at scale r=3, preliminary disparity maps, D_L and D_R, have been generated, accompanied by error maps, E_L and E_R, which will then be used as inputs for the next scale. Inspired by the work of J. Kang et al. [34] and A. Alghoul [10], adjusted cost volumes, ACV_L and ACV_R, were constructed at higher scales r=2 and r=1 as illustrated in **Figure -2-**. These volumes exploit bilinear interpolation of the disparity maps D_L and D_R, allowing for a significant increase in the resolution of the preliminary maps while facilitating the comparison of features between the left and right images. The correlation between these images is measured by a small window convolution, followed by L1 normalization on the depth dimension. The adjusted cost volumes ACV_L and ACV_R, as well as the augmented error maps, are then passed as input to further ConvLSTM networks, ensuring progressive and accurate refinement of the disparities at each scale.

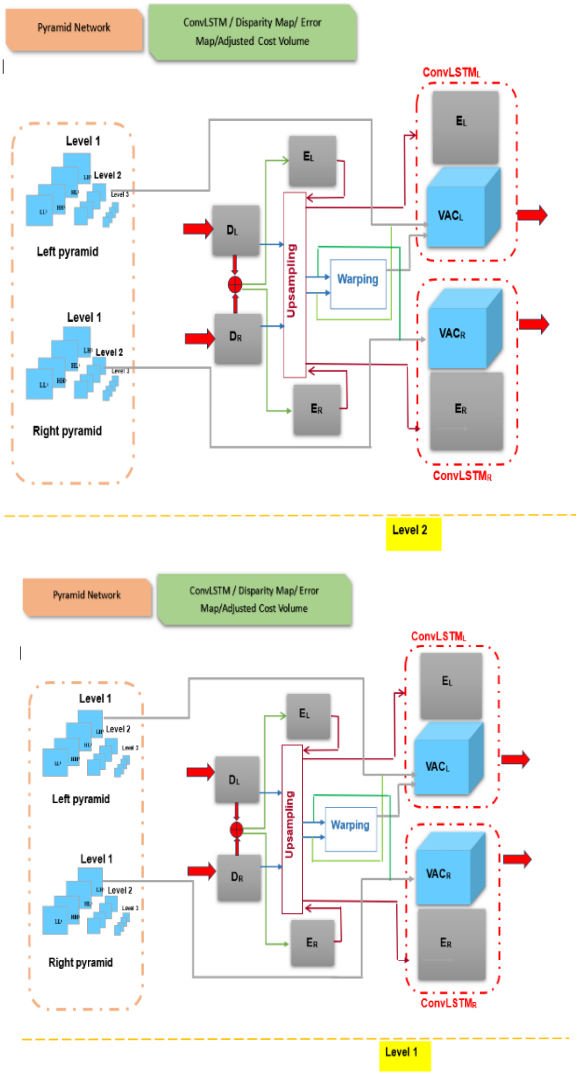

Figure – 2 -The process of generating the left and right Adjusted Cost Volumes (ACV) at scale 2 and scale 1.

3.2.3. Refinement by Stacked ConvLSTM and Disparity Prediction

The WS-CRNN model iteratively refines the disparity estimation using two stacked ConvLSTMs in parallel, processing the left and right views. At each resolution level, a ConvLSTM receives as input the matching cost volume (PCV_L or ACV_L and PCV_R or ACV_R) and the previous error map (E_L and E_R) to generate the left and right disparity maps (D_L and D_R). These maps

are then compared to produce new error maps, which are reintroduced into the model to progressively improve the uncertain regions.

ConvLSTM networks [35][36] are particularly effective at capturing contextual spatial information while reducing model redundancy, thus optimizing their role in filtering and refining the results. A ConvLSTM follows the same logic as a classic LSTM but applies spatial convolutions instead of matrix multiplications in its input, forget, and output gates:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i)$$
$$i_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_{io})$$
$$H_t = o_t \circ \tanh(C_t)$$

Where:

- *: represents the convolution operation instead of matrix multiplication.
- X_t: represents the input tensor obtained by concatenating the matching cost volume PCV and the error map generated at step t.
- H_{t-1} and C_{t-1}: correspond to the hidden state and memory of the ConvLSTM at step t−1.
- W_{xi}, W_{xf}, W_{xc}, W_{xo}: Convolution filters associated with the inputs..
- W_{hi},W_{hf},W_{hc},W_{ho}: Convolution filters applied to the hidden state.
- W_{ci},W_{cf},W_{co}: Parameters for the cell memory
- **σ**: Sigmoid function that constrains values between 0 and 1.
- i_t, f_t et o_t **:** represent the input gate, forget gate, and output gate, which regulate the flow of information within the model.

The ConvLSTM hidden state tensor is then passed through simple convolutional layers to produce a cost tensor of dimensions ($H/2^{2r} * W/2^{2r} * d_{max}$), where r=1,2,3 and H and W are the height and width of the input image, and d_{max} is the maximum disparity considered. By taking the negative of each value in this cost tensor, we obtain a score tensor:

$$S = -C$$

Where:

- **C**: is the cost tensor.
- **S**: is the score tensor.

A softmax normalization is applied to the score tensor to obtain a probability tensor that represents the probability of each possible disparity for each pixel:

$$P(d) = \frac{e^{S(d)}}{\sum_{d' \in D} e^{S(d')}}$$

Where:

- P(d): is the probability associated with disparity d.
- S(d): is the score for the cost corresponding to disparity d.

Rather than simply taking the disparity with the highest probability, a weighted average of the disparities is calculated to ensure better accuracy:

$$\hat{d} = \sum_{d \in D} d \cdot P(d)$$

Where:

- $\hat{d}$: is the final estimated disparity.
- **P(d)**: is the probability of disparity d.
- **d** represents each possible disparity value.

The left and right disparity maps (D_L and D_R), generated by the stacked ConvLSTMs of the left and right views inspired by Jie Z [36], are first converted into the coordinate system of the opposite view (D'_L and D'_R). Then, the initial disparity map and its converted version (D_L and D'_L, D_R and D'_R) are merged and processed through a series of convolutional layers, followed by a sigmoid transformation, to produce the corresponding error map (E_L and E_R). This error map is then propagated to the next level, allowing the model to target areas requiring adjustments.

4. Evaluations and Results

To assess the relevance of our approach and evaluate the performance of our design, we utilized a stereo dataset that integrates real-world conditions, combined with a rigorous evaluation protocol based on multiple metrics.

4.1. Data

According to MSDE [10], Selective-stereo [43], and All-in-one [44], four datasets are commonly used to examine stereo matching methods: Scene Flow [46], Middlebury [42], KITTI-2025 [45], and ETHD3 [47]. These datasets offer a remarkable diversity of synthetic, indoor, outdoor, and real-world contexts, thus allowing for testing the robustness of stereo matching algorithms.

Scene Flow [46] offers a vast collection of over 39,000 pairs of synthetic stereo images, carefully divided into training and test sets.

Middlebury 2014 [42] provides a corpus of 23 indoor scenes for training and 10 scenes for testing, with flexible resolution across three levels.

KITTI-2015 [45] stands out with its 200 training pairs and 200 test pairs, accompanied by sparse disparity maps taken from real driving scenes, presenting a challenge for algorithmic accuracy.

Finally, **ETH3D [47]** provides grayscale image pairs capturing both indoor and outdoor environments, adding further complexity to the analysis.

4.2. Evaluation Criteria

The KITTI 2015 benchmark [44][50], a key reference in the evaluation of computer vision algorithms, stands out for its ability to dive deep into dynamic environments. It allows for precise measurement of algorithm performance by analyzing the error rate of pixels, a crucial metric derived from a set of test images that are meticulously annotated according to a rigorously established ground truth.

A pixel is considered correctly estimated only when the error associated with it is less than 3 pixels or 5% of its actual value a tolerance threshold that pushes algorithms to their limits.

Errors are carefully classified into different categories, each holding its own significance in the overall evaluation:

- **D1 / D2**: These measures quantify the stereo disparity errors on the first and second image of a temporal pair, two crucial elements for depth perception.

- **Fl**: The optical flow error rate, reflecting the accuracy of observed movements between successive images.
- **SF** (Scene Flow): A composite measure where disparity and optical flow combine to provide a more global understanding of errors across the entire moving scene.
- **bg / fg / all**: The errors are then separated into three distinct regions: the background, the foreground, and all annotated pixels, allowing for a more detailed evaluation of specific image areas.

Such an evaluation protocol, which blends rigor and comprehensiveness, not only ensures an objective comparison of different approaches but also becomes the essential catalyst for the development of increasingly robust methods. It pushes the boundaries of visual perception, crucial for navigation in complex and ever-evolving environments.

4.3. Implementation

We implemented WS-CRNN using the PyTorch framework [48], leveraging the RMSprop optimizer [49] for improved gradient management within the range [-1, 1]. For pretraining, we selected the Scene Flow dataset, chosen for its rich collection of synthetic stereo pairs with high-fidelity ground truth, including both the cleanpass and finalpass subsets providing an ideal training environment for stereo matching. WS-CRNN operates across multiple representation scales, simultaneously generating three feature maps at 1/4, 1/8, and 1/16 of the original input size, thereby capturing information at varying levels of granularity.

Initial training was performed over 150 epochs with an initial learning rate of 0.001, decayed by a factor of 0.9 every 10 epochs. The ConvLSTM-based refinement stage was subsequently trained for 30 epochs with a learning rate of 0.01, reduced by a factor of 10 every 10 epochs. Fine-tuning was conducted on the KITTI dataset, using an 80/20 split for training and validation, and a reduced learning rate to better adapt the model to KITTI-specific disparity distributions. The full training procedure requires approximately 15 days on a high-performance GPU (RTX 3090 or higher), owing to the increased complexity of combining SWCNN and ConvLSTM modules. This comprehensive training strategy enables WS-CRNN to achieve strong generalization while progressively enhancing prediction accuracy through recurrent refinement.

4.4. Quantitative Results

The evaluation of the **WS-CRNN model** performance is based on a thorough analysis of disparity errors, where the predicted disparity maps are compared with ground truth maps. This evaluation process follows a rigorous protocol, divided into several steps. First, the disparity maps are generated by the WS-CRNN model, taking into account the specifics of each scene and the underlying dynamics. Then, performance metrics (D1, D2, bg, fg, and all) are computed according to the evaluation protocol provided by the **KITTI 2015 benchmark** [50].

The evaluation results of **WS-CRNN** on the **KITTI 2015 benchmark** are illustrated in **Figure -3-**, and compared with various state-of-the-art methods [10]. These results reveal some interesting trends: **WS-CRNN** manages to maintain relatively low errors in terms of overall disparity (D1-all and D2-all), which is a clear sign of excellent stereo matching accuracy. This accuracy is also reflected in modest errors on the background (bg), highlighting the model's ability to accurately estimate disparities even in low-texture areas. However, another, more subtle aspect emerges: errors in the foreground (fg) are slightly higher. This trend suggests that the model faces some challenges in complex areas, where moving objects or dense, varied textures pose greater difficulties for the disparity estimation method. This phenomenon, although expected in current vision algorithms, points to areas for future improvement.

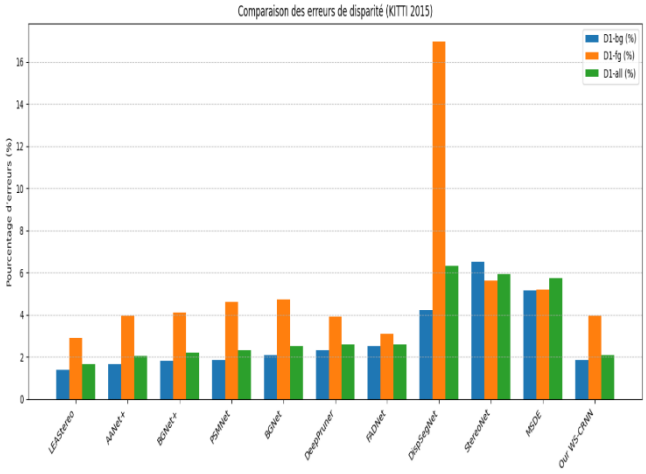

Figure -3- Comparison of disparity errors (KITTI 2015)

4.5. Qualitative Results

In this section, we present an in-depth qualitative evaluation of the WS-CRNN model applied to the KITTI 2015 dataset. **Figure – 4 -** illustrates a direct comparison between our approach and two state-of-the-art methods, including PSMNet, a model that leverages spatial pyramid pooling to capture global context at multiple resolutions, and effectively regularizes the 3D cost volume through stacked hourglass-type networks [6]. Since our model relies on ConvLSTM as proposed by [19], we also provide a direct comparison with the results from [19].

The visual results generated by WS-CRNN display disparity maps of competitive quality, both in terms of structural accuracy and robustness in ambiguous areas, such as fine borders, weakly textured surfaces, or partially occluded regions. The joint visualization of error maps and estimated disparities highlights the model's ability to maintain spatial coherence while effectively locating regions of uncertainty.

Finally, the maps produced by WS-CRNN are qualitatively very close to those generated by PSMNet and LRCR, suggesting that our model achieves performance parity with these architectures, while benefiting from the unique advantages associated with its spatio-temporal recurrent structure.

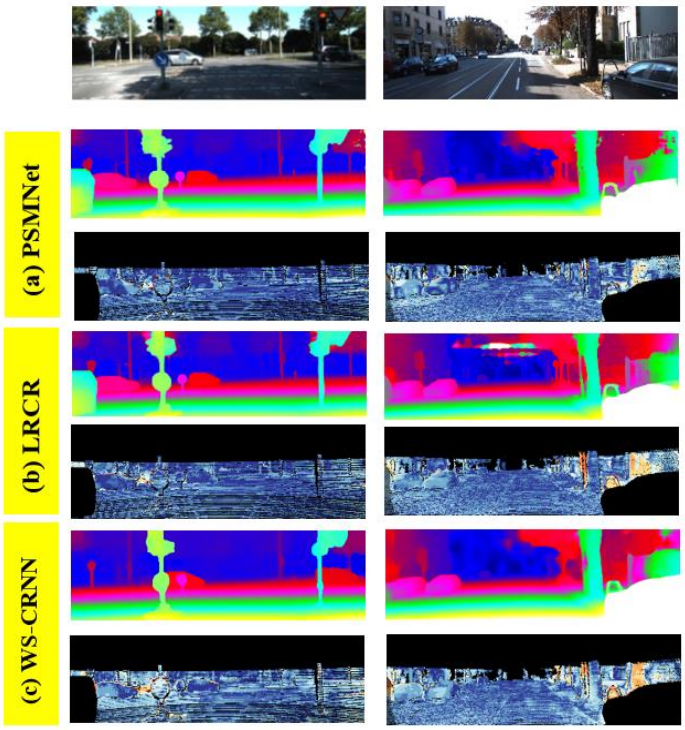

Figure 4: Shows the disparity prediction results on the KITTI 2015 test data. The first row displays the left images of the stereo pairs. The following rows show the disparity maps estimated by: (a) PSMNet, (b) LRCR, (c) WS-CRNN, along with their respective error maps.

4.6. Analysis of Configurations from the Ablation Study

As part of the analysis of the structural contributions of the WS-CRNN model, a rigorously conducted ablation study was performed to isolate the impact of three key modules by comparing their removal to the performance of the complete model. The configurations tested are detailed in **Figure 5**, where each row displays one of the five metrics (D1-all, D2-all, bg, fg, all) for each of the following ablation configurations:

- **AB1**: In this configuration, the feature extraction step using wavelet scattering was completely removed. The stereo image pairs, without the prior multi-scale processing, were fed directly into the network.
- **AB2**: Here, the ConvLSTM recurrent component responsible for refining the results was replaced by a simple 2D convolutional layer.
- **AB3**: In this configuration, instead of completely removing the multi-scale analysis, the pyramid strategy initially integrated into WS-CRNN was replaced with a more conventional pyramid decomposition, often used in earlier CNN architectures.
- **AB4**: The complete WS-CRNN model, including all the aforementioned modules, is used as the reference point. This is not an ablation itself, but it provides a comparative baseline to quantify the performance losses associated with each removal.

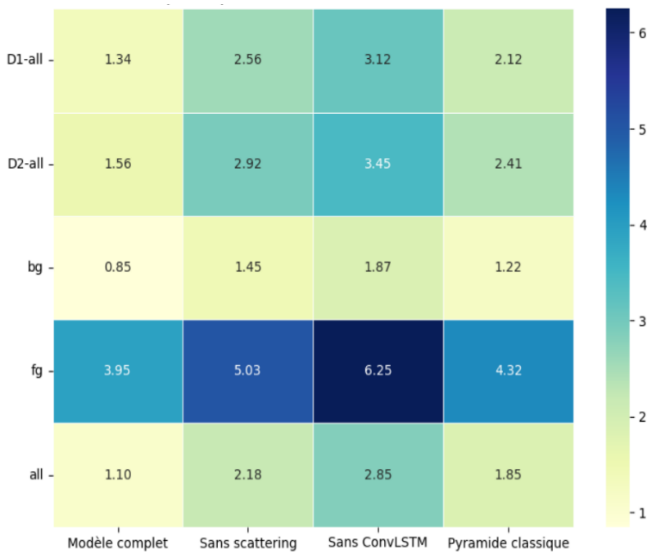

Figure – 5 -The evaluation results from the different ablation studies were obtained on the test set of the FlyingThings3D dataset.

The results of the ablation studies conducted on the WS-CRNN model using the FlyingThings3D dataset show that the complete model achieves the best performance in terms of accuracy and robustness in generating disparity maps. The removal of wavelet scattering leads to a loss of multi-scale information, increasing errors, particularly in low-texture areas. The absence of the ConvLSTM module impairs the model's ability to capture spatial and temporal dependencies, resulting in a significant increase in errors in complex areas. Lastly, using a classic pyramid decomposition, although it maintains a scale hierarchy, moderately degrades performance, especially in visually dense environments. These results highlight the importance of each component in the WS-CRNN model and their complementarity for precise and reliable disparity estimation.

5. Conclusion

In this study, we introduced WS-CRNN, an innovative neural architecture strategically designed for disparity estimation, balancing algorithmic precision and computational efficiency. The model leverages a methodological synergy between several low-dimensional cost volumes spread across multiple resolutions and a ConvLSTM recurrent module, which iteratively extracts, aggregates, and refines spatial-temporal dependencies. This foundation is further enhanced by the integration of wavelet scattering, serving as a hierarchical encoding mechanism for textures at different frequencies, thus strengthening the robustness of the initial representation.

This technological hybridization has enabled competitive performance in terms of disparity map quality, comparable to state-of-the-art approaches.

The systematic ablation studies conducted on the key components of the model empirically and quantitatively confirms their

functional complementarity. The removal of the scattering module, the elimination of the ConvLSTM, or the replacement of the multi-scale pyramid with a more conventional variant all lead, without exception, to a significant degradation in performance, highlighting the interdependence of the modules and the relevance of a deeply hierarchical modular architecture.

Looking ahead, future directions include the integration of more refined adaptive mechanisms, such as occlusion handling blocks or contextual attention modules, to reduce sensitivity to ambiguous areas and improve generalization under real-world conditions. Additionally, adapting WS-CRNN for platforms with limited resources or applications requiring real-time processing is a promising avenue, at the intersection of structural optimization and embedded intelligence.

Funding Declaration: No

Author Contribution Statement:

- Kheira HADJ DJELLOUL.: Conceptualization, - original draft. Data curation, Software, Visualization, Validation.
- Mohamed SENOUCI, Zineb HEMMAMI.: Methodology, Formal analysis, Investigation, Writing - review & editing.

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Hamid MS, Manap NA, Hamzah RA, et al**. Stereo matching algorithm based on deep learning: A survey**. Journal of King Saud University—Computer and Information Sciences. 2022. DOI: 10.1016/j.jksuci.2020.08.011

[2] T. Fabio and B. Luca and P. Matteo. **A Survey on Deep Stereo Matching in the Twenties.** arXiv:2407.07816v1 [cs.CV] 10 Jul 2024

[3] Zhou K, Meng X, Cheng B. **Review of Stereo Matching Algorithms Based on Deep Learning. Computational Intelligence and Neuroscience.** 2020. DOI: 10.1155/2020/8562323.

[4] Luo W, Schwing AG, Urtasun R. **Efficient Deep Learning for Stereo Matching**. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016. DOI: 10.1109/cvpr.2016.614

[5] Hirschmuller H. **Stereo Processing by Semiglobal Matching and Mutual Information**. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),2008. DOI: 10.1109/tpami.2007.1166

[6] Chang JR, Chen YS. **Pyramid Stereo Matching Network**. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. DOI: 10.1109/cvpr.2018.00567

[7] Khamis S, Fanello S, Rhemann C, et al. **StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction**. Lecture Notes in Computer Science, 2018. DOI: 10.1007/978-3-030-01267-0_35.

[8] Cheng X, Zhong Y, Harandi M, et al. **Hierarchical Neural Architecture Search for Deep Stereo Matching**. arXiv. arXiv. 2020, arXiv:2010.13501.

[9] Duggal S, Wang S, Ma WC, et al. **DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch**. IEEE/CVF International Conference on Computer Vision (ICCV), 2019. DOI: 10.1109/iccv.2019.00448.

[10] A. Alghoul and R. Battrawy and D. Stricker. **MSDE: Multi-scale disparity estimation model from stereo images**. Article *in* Journal of Autonomous Intelligence · January 2024. DOI: 10.32629/jai.v7i5.813

[11] Zbontar, J., LeCun, Y., 2016. **Stereo matching by training a convolutional neural network to compare image patches**. J. Mach. Learn. Res. 17, 1–32. https://DOI.org/10.1186/s13568-015-0106-7.

[12] Zbontar, J., LeCun, Y., 2015. **Computing the Stereo Matching Cost with a Convolutional Neural Network.** In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015). pp. 1592 1599. https://DOI. Org/10.1109/CVRP.2015.7298767.

[13] Seki, A., Pollefeys, M., 2017. **SGM-Nets: Semi-global matching with neural networks. In: Proceedings** – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 6640–6649. https://DOI.org/10.1109/CVPR.2017.703.

[14] Liang Z, Feng Y, Guo Y, et al. **Learning for Disparity Estimation Through Feature Constancy**. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. DOI: 10.1109/cvpr.2018.00297.

[15] Mayer N, Ilg E, Hausser P, et al**. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation**. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. DOI: 10.1109/cvpr.2016.438

[16] Pang J, Sun W, Ren JSJ, et al. **Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching.** IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.Published online October 2017. DOI: 10.1109/iccvw.2017.108

[17] Song X, Zhao X, Hu H, et al. **EdgeStereo: A Context Integrated Residual Pyramid Network for Stereo Matching.** Lecture Notes in Computer Science, 2019. DOI: 10.1007/978-3-030-20873-8_2

[18] Dovesi PL, Poggi M, Andraghetti L, et al**. Real-Time Semantic Stereo Matching**. IEEE International Conference on Robotics and Automation (ICRA), 2020. DOI: 10.1109/icra40945.2020.9196784

[19] Jie Z, Wang P, Ling Y, et al. **LRCR**: **Left-Right Comparative Recurrent Model for Stereo Matching**. IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018. DOI: 10.1109/cvpr.2018.00404

[20] Zhang F, Prisacariu V, Yang R, et al. **GA-Net: Guided Aggregation Net for End-To-End Stereo Matching**. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. DOI: 10.1109/cvpr.2019.00027

[21] Guo X, Yang K, Yang W, et al. **GWCNet: Group-Wise Correlation Stereo Network**. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. DOI: 10.1109/cvpr.2019.00339

[22] Lu C, Uchiyama H, Thomas D, et al. **SCVNet : Sparse Cost Volume for Efficient Stereo Matching**. Remote Sensing,2018. DOI: 10.3390/rs10111844

[23] Kendall A, Martirosyan H, **GCNet: Dasgupta S, et al. End-to-End Learning of Geometry and Context for Deep Stereo Regression**. IEEE International Conference on Computer Vision (ICCV), 2017. DOI: 10.1109/iccv.2017.17
[24] Tulyakov S, Ivanov A, Fleuret F. **Practical deep stereo (PDS): Toward applications-friendly deep stereo matching**. Adv Neural Inf Process Syst,2018, 5871–5881.
[25] Shi X, Chen Z, Wang H. **Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting**. arXiv. 2015, arXiv:1506.04214v2.
[26] Huang Z, Norris TB, Wang P. **ES-Net: An Efficient Stereo Matching Network**. Available online: http://arxiv.org/abs/2103.03922 (accessed on 13 October 2021).
[27] Wang Y, Lai Z, Huang G, et al. **AnyNet: Anytime Stereo Image Depth Estimation on Mobile Devices**. International Conference on Robotics and Automation (ICRA), 2019. DOI: 10.1109/icra.2019.8794003.
[28] He K, Zhang X, Ren S, et al. **Deep Residual Learning for Image Recognition**. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. DOI: 10.1109/cvpr.2016.90
[29] Gao Q, Zhou Y, Li G, et al. **Compact StereoNet: Stereo Disparity Estimation via Knowledge Distillation and Compact Feature Extractor**. IEEE Access,2020. DOI: 10.1109/access.2020.3029832
[30] Joan Bruna, Stéphane Mallat. **Invariant Scattering Convolution Networks**. Computer Vision and Image Understanding (CVIU). Mars 2012. https://DOI.org/10.48550/arXiv.1203.1513
[31] S. Mallat. **Modèles multi-échelles et réseaux de neurones convolutifs. Fev 2020; rév. 30 juin 2020.**
[32] Ben Zhang et Denglin Zhu. **Local Stereo Matching: An Approach Based on Adaptive Weighted Guided Image Filtering**. Revue internationale de reconnaissance de formes et d'intelligence artificielle Vol. 35, no 03, 2154010 (2021) https://DOI.org/10.1142/S0218001421540100
[33] Zhang, Y., & Li, X. (2013). "**A Novel Stereo Matching Algorithm Using L1-Norm for Correspondence**." *International Journal of Computer Science Issues*, 10(1), 19-25.
[34] Kang J, Chen L, Deng F, et al. **Improving disparity estimation based on residual cost volume and reconstruction error volume**. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,2020. DOI: 10.5194/isprs-archives-xliii-b2-2020-135-2020
[35] Batsos, K., Mordohai, P., 2018. **Recresnet: A recurrent residual cnn architecture for disparity map enhancement**, in: 2018 Int. Conference on 3D Vision (3DV). IEEE, 238–247.
[36] Jie, Z., Wang, P., Ling, Y., Zhao, B., Wei, Y., Feng, J., Liu, W., 2018. **Left-right comparative recurrent model for stereo matching**, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3838–3846.
[37] Lahav Lipson, Zachary Teed, Jia Deng. **RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching**. arXiv:2109.07547v1 [cs.CV] 15 Sep 2021.
[38] Gangwei Xu Xianqi Wang Xiaohuan Ding XinYang. **Iterative Geometry Encoding Volume for Stereo Matching. This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.2023**
[39] Ziyang Chen,Wei Long, He Yao and Yongjun Zhang**. MoCha-Stereo: Motif Channel Attention Network for Stereo Matching. This CVPR paper is the Open Access version, provided by the Computer Vision Foundation 2024.**
[40] Jiahao Pang Wenxiu Sun Jimmy SJ. Ren Chengxi Yang Qiong Yan**. Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching. arXiv:1708.09204v2 [cs.CV] 30 Jul 2018.**
[41] Jialiang Wang and Todd Zickler. **Local detection of stereo occlusion boundaries**. **This CVPR paper is the Open Access version, provided by the Computer Vision Foundation 2019.**
[42] Scharstein, D.; Hirschm¨uller, H.; Kitajima, Y.; Krathwohl, G.; Ne˘si´c, N.; Wang, X.; and Westling, P. 2014**. High resolution stereo datasets with subpixel-accurate ground truth.** In GCPR, 31–42. Springer.
[43] Xianqi Wang, Gangwei Xu, Hao Jia, Xin Yang . **Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching. CVPR 2024.**
https://DOI.org/10.48550/arXiv.2403.00486.
[44] Jingyi Zhou1, Haoyu Zhang1, Jiakang Yuan1, Peng Ye1,3,4, Tao Chen1, Hao Jiang2, Meiya Chen2, Yangyang Zhang2. **All-in-One: Transferring Vision Foundation Models into Stereo Matching**. AAAI 2025, https://DOI.org/10.48550/arXiv.2412.09912
[45] Menze, M.; and Geiger, A. 2015. **Object scene flow for autonomous vehicles**. In CVPR, 3061–3070. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 10.1109/CVPR.2015.7298925
[46] Mayer. N; Ilg. E; Hausser. P; Fischer. P; Cremers. D, Dosovitskiy. A and Brox. T. 2016. **A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation**. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4040–4048.
[47] Schops.T; Schonberger. JL, Galliani. S, Sattler. T, Schindler. K, Pollefeys. M and Geiger. A. 2017. **A Multiview stereo benchmark with high-resolution images and multi-camera videos**. In CVPR, 3260–3269.
[48] Paszke A, Gross S, Chintala S, et al. **Automatic differentiation in pytorch**. 2017.
[49] Lyon RF. **Neural Networks for Machine Learning. Human and Machine Hearing**. 2017. DOI: 10.1017/9781139051699.031.
[50] Page off official benchmark KITTI Scene Flow 2015. https://www.cvlibs.net/datasets/kitti/eval_scene_flow.php.