VisiTherS: Visible-thermal infrared stereo disparity estimation of human silhouette

Noreen Anwar, Philippe Duplessis-Guindon, Guillaume-Alexandre Bilodeau
LITIV, Polytechnique Montréal
noreen.anwar@polymtl.ca, philippe.duplessis-guindon@polymtl.ca, gabilodeau@polymtl.ca

Wassim Bouachir
Data Science Laboratory, University of Québec (TÉLUQ)
wassim.bouachir@teluq.ca

Abstract

This paper presents a novel approach for visible-thermal infrared stereoscopy, focusing on the estimation of disparities of human silhouettes. Visible-thermal infrared stereo poses several challenges, including occlusions and differently textured matching regions in both spectra. Finding matches between two spectra with varying colors, textures, and shapes adds further complexity to the task. To address the aforementioned challenges, this paper proposes a novel approach where a high-resolution convolutional neural network is used to better capture relationships between the two spectra. To do so, a modified HRNet backbone is used for feature extraction. This HRNet backbone is capable of capturing fine details and textures as it extracts features at multiple scales, thereby enabling the utilization of both local and global information. For matching visible and thermal infrared regions, our method extracts features on each patch using two modified HRNet streams. Features from the two streams are then combined for predicting the disparities by concatenation and correlation. Results on public datasets demonstrate the effectiveness of the proposed approach by improving the results by approximately 18 percentage points on the $\leq$ 1 pixel error, highlighting its potential for improving accuracy in this task. The code of VisiTherS is available on GitHub at the following link: https://github.com/philippeDG/VisiTherS.

1. Introduction

The objective of this paper is to propose a method for estimating pixel disparities between a visible image and a thermal infrared image. The idea of combining these two types of images is to benefit from each of them for tasks, such as object detection. If both images form a stereo pair, estimating disparity allows depth estimation that can further be used to improve detection itself or subsequent tasks, like tracking. The estimation of these pixel disparities can be used to align the visible-thermal infrared stereo thereby generating an augmented image. These augmented images are particularly useful in challenging scenarios such as low light, fog, and smoke, and can significantly enhance the accuracy of object detection and tracking. In contrast to classical stereo, pixel pattern-based approaches are insufficient in visible-thermal infrared stereo, highlighting the need for more advanced techniques for estimating pixel disparities.

(a) RGB rectified image (b) LWIR rectified image

Figure 1. A rectified visible and thermal infrard image from the LITIV 2018 dataset [17].

To deal with this particular challenge, we focus on estimating disparities of human silhouettes. We assume that those disparities are estimated from a sparse set of points, that is, our method is designed for sparse stereoscopy. Human silhouettes can be captured in both visible and thermal images, where the silhouette in thermal images is formed from the body's heat emission, and in the visible images by the color on the person. In this case, relying solely on pixel patterns as in classical stereo is insufficient. For example, the shirt logo present in the visible image (Figure 1a) is absent in the thermal image (Figure 1b). In classical stereo, the shirt logo would have been an effective mean

of estimating the disparity. We can also observe that the heat emitted through the shirt is unevenly distributed across surface. Consequently, the thermal image displays some intensity variation in the shirt region, while the visible image depicts uniform black coloring. We aim to propose a precise and efficient method that can estimate the disparity between the pixels in two human silhouettes.

This paper introduces a new convolutional neural network (CNN) architecture, called VisiTherS (standing for Visible-Thermal infrared Stereo), for estimating the disparity between visible and thermal image pairs. We propose to use a high-resolution network for extracting features as we believe that it can capture better the relationships between the pixels in both types of images. For this purpose, we selected HRNet to obtain a series of feature maps with strong semantic meaning at different scales. We investigated two ways to use the feature maps. First, we concatenated the features at the different scales of the last stage, which are adjusted to match the highest resolution feature map size. The resulting feature maps can take advantage of both high-resolution information and multi-scale information, providing a more comprehensive representation of the input data.Second,we concatenated high-resolution feature maps of two stages, thus retaining only the high-resolution information. We show that both of these strategies to exploit high-resolution features significantly improve results compared to the best SOTA methods.

Our proposed method consists of two HRNet streams, where each patch of the image in the stereo pair has its own feature extractor.While both streams have the same structure, but there is no weight sharing between them. VisiTherS takes two small square patches as input and extracts features from them, resulting in a feature vector for each image patch. To enhance the robustness of our network for predicting disparities, we employ two fusion techniques on the feature vectors. Firstly, we perform a correlation product between both vectors, forwarding the result to the correlation head. Secondly, we perform a concatenation between the two vectors, forwarding the result to the concatenation head. Better results are generally obtained by using the two fusion techniques simultaneously compared to only using correlation or concatenation. The correlation and concatenation heads consist of fully connected layers outputing the probability of both patches being the same or not. Each classification head has its own loss function, and during testing, we employ both classification heads to obtain the disparity predictions.

Our contributions can be summarized as follows:

- We propose VisiTherS, a new CNN architecture based on two streams composed of a high-resolution convolutional neural network feature extractor. Our architecture extracts features from both image domains and uses two fusion processes to compute the probability of the input patches being the same.
- Our findings show that our model is highly effective in performing disparity estimation between visible and thermal image pairs and that high-resolution features are a good choice for this kind of task. This represents a significant improvement over existing approaches and highlights the potential of our novel CNN architecture in advancing the field of disparity estimation.

2. Literature review

Stereo estimation can be achieved through two primary approaches: sparse stereo and dense stereo. These two approaches are the main methods used to perform the stereo estimation. Sparse stereo estimation involves selecting two regions from the original images, rather than inputting the entire images into the network. That is, only small patches around the disparity points are fed into the network. The objective is to find the corresponding patch in the other image. The disparity is calculated by measuring the pixel distance between the coordinates of these two patches. Since dense disparity labels are not required, this approach is applicable to both dense and sparse datasets, although it is generally slower in the case of dense stereo estimation.

Several papers have explored this approach in visible-visible stereo, including the pioneering work of Zbontar and LeCun [23], where a CNN is used to learn the similarity between a 9×9 region on the left and right images, with the goal of determining the disparity between these regions. Luo et al. [15] built upon this approach by creating a feature vector for the left image patch and a feature volume for the right patch and using correlation products to calculate the probability distribution of the disparity. Kendall et al. [13] proposed the GC-Net, which was the first end-to-end architecture using a Siamese network for feature extraction and 3D convolution for disparity mapping. Other methods have then been developed, including those using spatial pyramid pooling modules for feature extraction, and hourglass networks for cost volume regularization and disparity regression [7].

In dense stereo, disparities are estimated for each pixel of the images. To effectively train a machine learning model and reduce the likelihood of it overfitting, it is necessary to have datasets that contain a large number of densely labeled examples. The first that proposed an end-to-end dense stereo model were Mayer et al. [16]. This work had a huge impact on the field since they created a densely annotated dataset FlyingThings3D. This dataset consists of images having a disparity value at every pixel, which leads to a lot of subsequent work using the end-to-end method. Their method, called DispNet, is inspired by FlowNet [10] for compression and decomposition, respectively. The compression part is built with convolutions that result in a final

reduction factor of 64. The decompression then resizes the disparity maps gradually in a non-linear way, taking into consideration the characteristics in the compression step. The final result of the network is a disparity map with the same image size as the input image.

Prior to the rise of neural networks, Visible-thermal infrared stereo relied on matching feature points, often using SIFT [14] as a feature descriptor. MSIFT [6] was then introduced to improve the correlation between RGB channels in RGB (visible)-NIR (Near-infrared) pairs of images. However, some methods have opted to use window-based methods, such as mutual information [20], HOG [8], SSD [5], LSS [19], to find image matches. Among these window-based methods, Bilodeau et al. [5] found that mutual information is the most accurate approach [17].

In recent studies of visible-thermal infrared stereo, Beaupré et al. [3] proposed a novel method using two Siamese networks to compute the disparity from visible to thermal and vice versa. The Siamese networks have shared parameters, and their architecture is similar that of Luo et al [15]. The method involves comparing a small patch of a visible image with a patch in the thermal image of the same height, but of the full width of the original image. The correlation is done with every possible translation to find the corresponding disparity. The same principle applies to the other Siamese network, however the small patch correspond to is the thermal image at the given disparity while the wider image is the visible image. To select the final disparity, a summation layer is used, to sum up, the prediction vector from each network branch, with the final disparity being the maximum element.

In a subsequent work by Beaupré et al. [4], a modified approach was proposed, yielding to significant improvements over the previous method. Unlike the previous method, this approach does not share weights between the two feature extraction branches. This change was made due to the dissimilar nature of the two types of images used in visible-thermal infrared stereo matching. Unlike the typical inputs used in Siamese networks, the visible and thermal images are dissimilar in terms of color, shape, and contrast. The only aspect they have in common is the shape of the objects, which is not even exactly the same due to the differences in how the images are captured. Therefore, parameter sharing between feature extractors is not appropriate in this case. This approach served as an inspiration for our work. In the work of Duplessis-Guindon et al. [11], an approach was proposed for estimating the disparity of people in a scene using segmentation masks obtained from both visible and thermal images. Masks helped estimate the disparities at the object boundaries.

Visible-infrared stereo matching is not limited to thermal infrared, as there have been studies on Visible-Near infrared (NIR) stereo as well. Aguilera et al. [2] investigated the effectiveness of three different CNN architectures compared to the traditional methods mentioned earlier for this task. Building on their previous work, Aguilera et al. [1] introduced quadruplet networks that take two matching pairs of images, providing two pairs of positive examples and four pairs of negative examples for training. However, similarly to visible-thermal infrared stereo, there is a shortage of datasets for Visible-NIR stereo. To address this problem, Zhi et al. [22] created a method that transforms a visible image into the NIR spectrum and uses the resulting image for self-supervised learning.

3. Proposed method

Our method is inspired by the work of [4]. Figure 2 visually depicts the overall architecture of our model. It is composed of two streams, one for the visible (RGB) and one for thermal infrared (LWIR) patches. In both, features are extracted using a high-resolution CNN. Features are then fused and patches are classified. Our architecture is detailed in the following.

3.1. Feature extractor

In this section, we explain in detail the feature extraction part of our architecture. It requires two patches as input, an RGB and an LWIR patch. These patches are sized 36×36 to capture the surrounding context of the image around a point where we wish to calculate disparity. These patches are referred to in the following as P_{RGB} and P_{LWIR}. As shown in Figure 2, each patch is processed by its own feature extractor with different learned weights. Each feature extractor outputs a $36 \times 36 \times 64$ feature map represented by F_{RGB} and F_{LWIR} as illustrated in Figure 2.

For feature extraction, we selected HRNet [21] to obtain high-resolution features. This is motivated by the fact that visible and thermal infrared are different, and we believe that more expressive feature maps are required to match them. Traditional CNN backbone architectures reduce resolution between convolution layers, leading to less information in the final feature maps. HRNet major objective is to align input and output resolution. HRNet maintains resolution after each convolution and each stage adds a new feature map scale. The network output is a concatenation of these feature maps. All feature maps are resized to match the original input size. The final feature map, therefore, has a large number of channels. The original HRNet [18] network performs a series of convolutions on this final feature map to reduce its dimensions. However, our goal in introducing this feature extractor is to have the best possible resolution. We therefore only scaled the number of channels to have as output a feature map of size $36 \times 36 \times 64$. The last dimension of the feature map represents the number of feature channels. In our HRNet architecture, we removed the

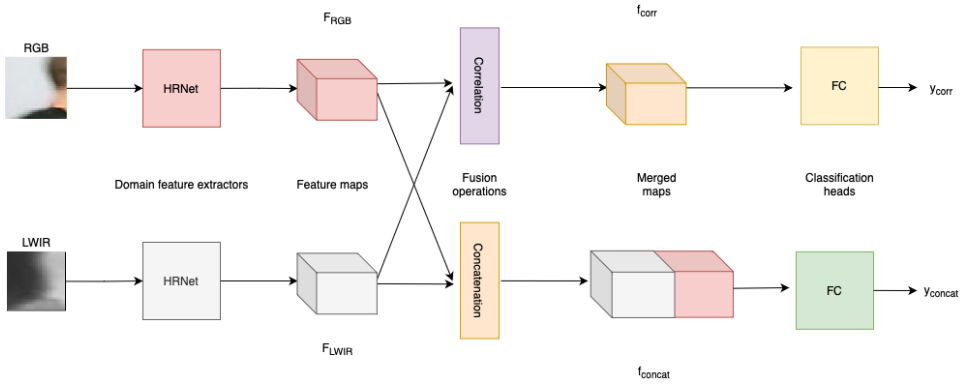


Figure 2. Overview of our proposed network architecture. RGB: visible, LWIR: Thermal infrared. Features are first extracted by two independent streams for RGB and LWIR. Features are then combined together using concatenation and correlation. Correlated features and concatenated features are then fed to two separate classification heads. Classification results are then combined (not shown in the figure).

upper layers from the original and kept only the first three stages.

We investigated two ways of exploiting the feature maps generated by HRNet. In the first, we concatenate features from several scales at the last stage. This is illustrated in Figure 3a. In the figure, yellow feature maps are 36×36, orange feature maps are 18×18, and red feature maps are 9×9. To obtain the concatenated feature map, we concatenated the three feature maps of the last stage and adjust their sizes to match the highest resolution. This results in a concatenated output of the three multiscale feature maps, which formed the final feature map F_{RGB} or F_{LWIR}, depending on the stream.

In the second way presented in Figure 3b, we concatenate the highest resolution features from several stages. More precisely, we are concatenating the high-resolution feature map of the last stage with that of the previous stage. Therefore, we only keep high-resolution information. This results in a concatenated output of the two high-resolution feature maps, which formed the final feature map F_{RGB} or F_{LWIR}, depending on the stream.

3.2. Classification heads

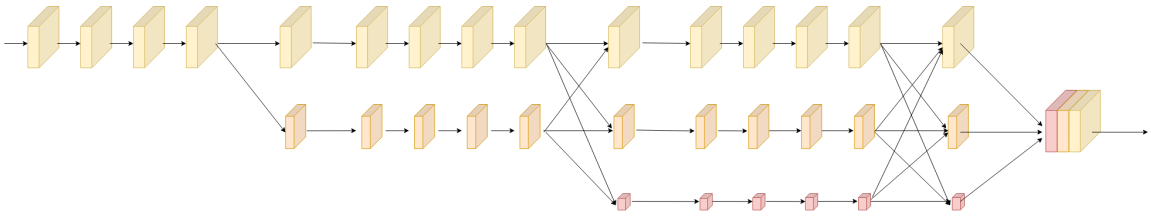
In our proposed method, we employ two distinct fusion operations on feature maps, namely correlation and concatenation, as described in [12]. These fusion operations are widely used in disparity estimation for integrating image features. While both operations have their advantages, each also presents certain limitations. Specifically, the correlation fusion operation is characterized by its computational speed and memory efficiency; however, it may result in the loss of some features from both spectra during the fusion process. On the other hand, the concatenation operation does not lead to any loss of features, but it entails a trade-off between the computational time and memory space required for its implementation. The correlation operation outputs a $36 \times 36 \times 64$ feature map, represented by f_{corr} in Figure 2. The concatenation operation outputs a $72 \times 36 \times 64$ feature vector, which is represented by f_{concat} in Figure 2. Both f_{corr} and f_{concat} are going through separate fully connected networks (FCNs). The weights are not shared between each fully connected network and each output a classification vector. These are represented by y_{corr} and y_{concat} in Figure 2. Both FCNs generate a 2D probability vector and this vector represents the likelihood that two patches are either identical or different.

3.3. Training losses

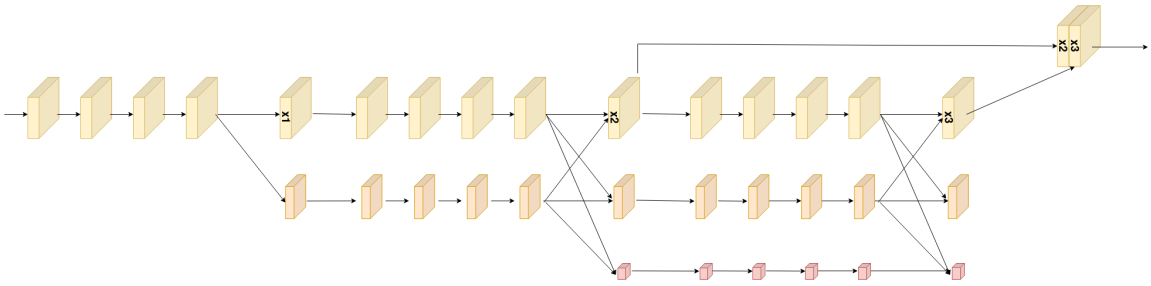
The network can learn by training on two corresponding image patches of 36×36 pixels (P_{RGB} and P_{LWIR}), one for the visible spectrum and one for the thermal spectrum. During inference, the network is fed with a 36×36 patch in the visible spectrum and it tries to locate the corresponding patch within a larger thermal image patch. In other words, the network learns to associate the two types of images and can use this knowledge to identify the location of a visible patch within a thermal patch.

To train our network, we employ two separate loss functions, one for the correlation head and another for the concatenation head. This allows us to optimize the network performance based on both fusion schemes. They are given by

$$loss_{corr} = -1/N \sum_{i=1}^{N} gt^i log(y^i_{corr}), \qquad (1)$$



(a) Concatenation of feature maps of various scales from the last stage.



(b) Concatenantion of feature maps from the last two stages.

Figure 3. Our proposed versions of feature extractors. For 36×36 image patches, neglecting the number of channels, yellow feature maps are 36×36, orange feature maps are 18×18, and red feature maps are 9×9.

and

$$loss_{concat} = -1/N \sum_{i=1}^{N} gt^i log(y^i_{concat}), \tag{2}$$

where N represent the number of data points, gt^i the ground-truth, which is 0 or 1 if the patches are the same, and y^i_{corr} and y^i_{concat} are the similarity probabilities.

The total loss function is given by the sum of both losses in both heads by

$$loss_{total} = loss_{corr} + loss_{concat}. \tag{3}$$

3.4. Disparity estimation

To evaluate the disparity, a maximum disparity value d_{max} is established. To form a wider thermal patch, with the same height of 36 and width of $36 + d_{max}$, half of this distance is added to both sides of the center point of the patch. With this, the network is able to perform d_{max} translations of a 36×36 thermal image patch while the visible patch remains the same.

After passing these patches in the feature extractor, F_{RGB} will be a feature map of size $36 \times 36 \times 64$ and F_{LWIR} will be a feature map of $(36 + d_{max}) \times 36 \times 64$. Next, the F_{LWIR} are passed through the fusion operations and passed through the fully connected layers, as explained earlier. The resulting y_{corr} and y_{concat} correspond to the probability of the patches being the same or different.

For every possible disparity value in the enlarge thermal patch, there is now a matching probability indicating whether the 36×36 patch at this disparity value corresponds to the visible patch or not. The disparity is then the index $\hat{d}$ with the highest probability. This is given by

$$\hat{d}_{corr} = argmax(y_{corr}), \tag{4}$$

and

$$\hat{d}_{concat} = argmax(y_{concat}). \tag{5}$$

The final disparity is an average of the best disparity from each branch $\hat{d}_{corr}$ and $\hat{d}_{concat}$ and is given by

$$\hat{d} = \frac{\hat{d}_{corr} + \hat{d}_{concat}}{2}. \tag{6}$$

4. Experiments

In this section, we provide a detailed overview of the experimental setup, datasets used for training and testing our model, as well as our results with comparison with other state-of-art methods. We also present an ablation study.

4.1. Implementation details

Our network is built using the PyTorch framework, with a default patch size of 36×36 with a maximum disparity of 64 for testing. The HRNet backbone was pre-trained on ImageNet [9]. We only use the first three stages.

We employ the Adam optimizer for backpropagation. We use a gradient step of 0.001. We trained for 200 epochs with a batch size of 24, as it is the maximum that fits on an RTX 2080 GPU.

4.2. Dataset and metrics

We used two datasets: the LITIV 2014 [17] dataset and the LITIV 2018 [17] dataset. The limited availability of visible-thermal infrared datasets pose a significant challenge for training CNNs. In our particular case, despite using the LITIV datasets, the number of ground-truth points is only slightly above 40,000, which is inadequate for robust training without data augmentation. Therefore, we use two data augmentation techniques. The first data augmentation technique consists of assigning the same disparity as a ground-truth point to its immediate neighbors [4]. Therefore, for a pixel with a Manhattan distance of one, we consider that they all have the same disparity. This makes the dataset 5 times bigger. Another technique used to generate more data is mirroring over the y axis. This additionally doubles the number of data points.

Our method was evaluated with cross-validation and trained/tested using different folds, mixing both datasets for training, validation and testing. We used the same folds as Beaupré et al. [4]. The datasets feature several actors moving in a room. It is to be noted that a few files are missing from the original datasets. Therefore, for a fair comparison, we re-ran the Beaupré et al. [4] method on the slightly incomplete dataset. We have evaluated our method with the recall metric given by

$$Recall = \frac{1}{N}\sum_{i=1}^{N} |\hat{d}^i - gt^i| \leq n, \tag{7}$$

where N stands for the number of points to be evaluated, $\hat{d}^i$ represents the evaluated disparity at a given point, gt^i is the ground-truth at the same given point, and lastly, n represents the allowed correspondence error in pixels.

4.3. Comparison with state-of-the-art methods

The performance of our proposed VisiTherS approach, which incorporates both scale concatenation (VisiTherS-scales) and stage concatenation (VisiTherS-stages), was evaluated against several state-of-the-art (SOTA) methods on the LITIV 2014 and LITIV 2018 datasets. Tables 1 and 2 present the results of these evaluations. The tables report the mean of three folds. VisiTherS obtains SOTA results on both datasets, with significantly improved performance for the ≤ 1 $pixel\ error$ and ≤ 3 $pixel\ error$, particularly for the LITIV 2014 dataset. Given, the low standard deviation, this performance is observed across all folds. This validates our hypothesis that high-resolution features are important for matching the content of dissimilar modalities, like thermal infrared and visible images. Comparing our proposed two versions of feature exploitation strategies, we can observe that they give results that are quite similar with a small advantage to VisiTherS-scales for the ≤ 1 $pixel\ error$ on LITIV 2014 and the reverse on LITIV 2018. This suggests that incorporating multiple scales can improve the correspondence process since the complexity of the content of patches may differ across scales, but considering different stages can give equivalent results. On the LITIV 2018 dataset, 4D-MultispectralNet that uses object masks is not far behind VisiTherS for the ≤ 1 $pixel\ error$, but having high-resolution features proves to be globally a better strategy. Adding masks to VisiTherS did not improve our results.

It should be noted that the results obtained with the Domain Siamese CNN method [4] differ slightly from those reported in the corresponding paper, as the code was rerun. It yields slightly lower results for ≤ 3 $pixel\ error$ precision, but for ≤ 1 $pixel\ error$ and ≤ 5 $pixel\ error$ precision, the results are higher than their initial study due to differences in the dataset. Considering both the new results and the originals, our proposed method outperforms Domain Siamese CNN significantly showing the benefit of high-resolution features.

4.4. Ablation study

4.4.1 Ablation study of feature fusion

Previous studies showed that using concatenation and correlation of features simultaneously gave better results than each separately [4]. Our new approach was able to validate this observation. In this study, while both convolutional neural networks (CNNs) extracted features from each patch, only one feature fusion operation was performed at a time to observe its performance. This study was performed with VisiTherS-scales. Results are presented in Table 3. They indicate that generally, the combination of both fusion methods yields superior performance compared to each fusion operation used separately. However, the correlation fusion method outperformed the concatenation method and the combined method (VisiTherS-scales) for the third fold of LITIV 2014. Nevertheless, by comparing the results for LITIV 2014 in Table 3, it can be observed that combining the two operations gives better results than using the concatenation or correlation operation for most folds. For LITIV 2018, the correlation operation outperforms the combined operations for the second fold. The correlation operation performs better in terms of recall metric across all three precision values. In general, correlation is a more efficient approach than concatenation. However, the result is improved when both are used together.

4.4.2 Comparison of the two proposed feature extractors

We conducted ablation studies on both versions of our proposed feature extractors. By concatenating the full resolution of the last two stages, we achieved better results,

Table 1. Results on LITIV 2014 compared to SOTA Methods. The results are the mean of the 3 folds [4] domain with standard deviation. †We re-ran their code with the slightly incomplete dataset. ‡: results on the complete dataset. VisiTherS-scales: feature maps concatenated from three scales. VisiTherS-stages: high-resolution feature maps concatenated from two stages. **Bold: Best Result**

Method	≤ 1 *pixel error*	≤ 3 *pixels error*	≤ 5 *pixel errors*
Domain Siamese CNN [4] †	56.3 ± 3.6	89.9 ± 0.4	98.5 ± 0.4
Siamese CNN ‡ [3]	-	77.9 ± 5.0	-
St-Charles [17] ‡	48.2 ± 4.0	-	-
Mutual Information [5] (40×130) ‡	-	83.3	-
Mutual Information [5] (20×130)‡	-	77.5	-
Mutual Information [5] (10×130) ‡	-	64.9	-
Fast Retina Keypoint [5](40×130) ‡	-	64.1	-
Local Self-Similarity [5, 17](40×130)‡	22.6 ± 10.7	73.4	-
Sum of Squared Difference [5](40×130) ‡	-	65.6	-
4D-MultispectralNet [11]	57.5 ± 2.3	88.7 ± 1.0	98.6 ± 0.4
VisiTherS-scales (ours)	$\mathbf{75.0 \pm 0.7}$	96.2 ± 0.4	99.6 ± 0.2
VisiTherS-stages (ours)	74.1 ± 1.2	$\mathbf{96.9 \pm 0.6}$	$\mathbf{99.8 \pm 0.1}$

Table 2. Results of all our methods on the 2018 LITIV dataset. The results are the mean of the three folds with standard deviation. **Bold: Best Result.**

Methods	≤ 1 *pixel error*	≤ 3 *pixels error*	≤ 5 *pixels error*
DASC Sliding Window [17]	10.4	-	-
Multispectral Cosegmentation [17]	26.5	-	-
Domain Siamese CNN [4] †	44.2	-	-
4D-MultispectralNet [11]	60.5 ± 4.4	87.4 ± 2.0	98.7 ± 0.1
VisiTherS-scales (ours)	63.3 ± 7.0	92.6 ± 2.3	99.7 ± 0.2
VisiTherS-stages (ours)	$\mathbf{63.6 \pm 5.4}$	$\mathbf{94.8 \pm 2.6}$	$\mathbf{99.9 \pm 0.1}$

as demonstrated in Table 4. Comparing the results on the LITIV2014 dataset, we observed an improvement in precision from 96.24 to 96.94 for ≤ 3 *pixel errors*. However, for ≤ 3 *pixel errors*, the precision dropped slightly from 75.00 to 74.14, which can be considered relatively similar as the standard deviation overlaps. The precision for ≤ 5 *pixel errors* improved slightly from 99.61 to 99.87. Regarding the results on the LITIV 2018 dataset, VisiTherS-stages always gets better results compared to the VisiTherS-scales.

4.4.3 Impacts of the choice of layers

We tested the accuracy of the high-resolution layer according to each stage in HRNet. We can see the results in the table 5. In this table, $x1$ represents the first stage full resolution output, $x2$ represents the second stage full resolution output, and $x3$ represents the third stage full resolution output (see figure 3b). We can see that the best results are split between $x2$ and $x3$. Indeed, for the ≤ 1 *pixel error*, the last stage has better performance. However, for ≤ 3 *pixel error* and ≤ 5 *pixel error*, stage $x2$ is better. This, therefore, justifies our choice to use the output of $x2$ and $x3$ in VisiTherS-stages.

5. Conclusion

This paper introduces a new method for visible-thermal infrared disparity estimation. The proposed model is designed with two versions of feature extractors that employ two streams to extract features independently for each visual and thermal infrared image patch. The first version concatenates features of different scales in one layer, while the second version concatenates high-resolution features of different stages. The model combines the extracted features from both images using two operations, namely correlation and concatenation, to jointly enhance the network performance. Overall, the proposed model, VisiTherS, offers a novel solution for disparity estimation with promising results. Experimental evaluation on public datasets reveals that the proposed method surpasses several SOTA methods.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2020-04633].

Table 3. Ablation study of fusion methods in terms of recall for several pixel errors. **Bold: Best Result.**

	Correlation			Concatenation			VisiTherS-scales (both operations)		
	$\leq 1\ pixel$	$\leq 3\ pixels$	$\leq 5\ pixels$	$\leq 1\ pixel$	$\leq 3\ pixels$	$\leq 5\ pixels$	$\leq 1\ pixel$	$\leq 3\ pixels$	$\leq 5\ pixels$
LITIV2014-fold1	73.44	**96.22**	99.76	69.22	93.32	99.11	**75.61**	95.97	**99.80**
LITIV2014-fold2	68.23	94.89	99.37	64.88	95.45	99.41	**75.08**	**96.06**	**99.67**
LITIV2014-fold3	77.15	95.87	99.66	**80.93**	**97.68**	**99.86**	74.30	96.06	99.36
LITIV2018-fold1	68.39	92.79	99.62	60.70	90.72	99.50	**68.64**	**94.65**	**99.92**
LITIV2018-fold2	**59.33**	**92.21**	**99.71**	56.07	87.14	97.73	55.40	90.16	99.52
LITIV2018-fold3	64.14	**93.06**	99.55	59.95	87.20	97.80	**65.98**	92.83	**99.64**

Table 4. Ablation study of comparison between proposed versions of feature extractors in terms of recall for several pixel errors. **Bold: Best Result.**

Dataset	Error	VisiTherS-scales	VisiTherS-stages
LITIV 2014	$\leq 1\ pixel$	$\mathbf{75.00 \pm 0.66}$	74.14 ± 1.21
	$\leq 3\ pixels$	96.24 ± 0.40	$\mathbf{96.94 \pm 0.56}$
	$\leq 5\ pixels$	99.61 ± 0.23	$\mathbf{99.87 \pm 0.04}$
LITIV 2018	$\leq 1\ pixel$	63.34 ± 7.00	$\mathbf{63.55 \pm 5.37}$
	$\leq 3\ pixels$	92.55 ± 2.26	$\mathbf{94.83 \pm 2.64}$
	$\leq 5\ pixels$	99.69 ± 0.21	$\mathbf{99.90 \pm 0.10}$

Table 5. Results on LITIV2014-fold1, according to the depth of the high-resolution stage. **Bold: Best Result.**

Error	x1	x2	x3
$\leq 1\ pixel$	73.04	76.66	**76.83**
$\leq 3\ pixels$	95.58	**96.64**	96.35
$\leq 5\ pixels$	99.26	**99.84**	99.76

References

[1] Cristhian Aguilera, Angel Sappa, and Ricardo Toledo. Cross-spectral local descriptors via quadruplet network. *Sensors*, 17:873, 04 2017. 3

[2] Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 267–275, 2016. 3

[3] David-Alexandre Beaupre and Guillaume-Alexandre Bilodeau. Siamese cnns for rgb-lwir disparity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3, 7

[4] David-Alexandre Beaupre and Guillaume-Alexandre Bilodeau. Domain siamese cnns for sparse multispectral disparity estimation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3667–3674. IEEE, 2021. 3, 6, 7

[5] Guillaume-Alexandre Bilodeau, Atousa Torabi, Pierre-Luc St-Charles, and Dorra Riahi. Thermal–visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64:79–86, 2014. 3, 7

[6] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184, 2011. 3

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[11] Philippe Duplessis-Guindon and Guillaume-Alexandre Bilodeau. 4d-multispectralnet: Multispectral stereoscopic disparity estimation using human masks. *arXiv preprint arXiv:2204.09089*, 2022. 3, 7

[12] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 4

[13] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression, 2017. 2

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004. 3

[15] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, 2016. 2, 3

[16] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 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*, pages 4040–4048, 2016. 2

[17] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Online mutual foreground segmentation for multispectral stereo videos. *International Journal of Computer Vision*, 127:1044–1062, 2019. 1, 3, 6, 7

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 3

[19] Atousa Torabi and Guillaume-Alexandre Bilodeau. Local self-similarity as a dense stereo correspondence measure for themal-visible video registration. In *CVPR 2011 WORKSHOPS*, pages 61–67. IEEE, 2011. 3

[20] P. Viola and W.M. Wells. Alignment by maximization of mutual information. In *Proceedings of IEEE International Conference on Computer Vision*, pages 16–23, 1995. 3

[21] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021. 3

[22] Tiancheng Zhi, Bernardo R. Pires, Martial Hebert, and Srinivasa G. Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[23] J. Žbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1592–1599, 2015. 2