# Domain Siamese CNNs for Sparse Multispectral Disparity Estimation

David-Alexandre Beaupré and Guillaume-Alexandre Bilodeau
LITIV lab., Department of Computer and Software Engineering, Polytechnique Montreal
{david-alexandre.beaupre, gabilodeau}@polymtl.ca

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021
POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

## 1. Abstract

Multispectral disparity estimation is a difficult task since it includes the challenges found in the visible spectrum, while also having the difficulty of matching corresponding pixels with few similarities. We propose a new CNN architecture that extracts features from each patch without sharing any weights. We combine those features with two operations: correlation and concatenation. These newly merged features are then forwarded to their respective classification heads that determine if the input patches are the same or not. Experiments on the LITIV 2014 and 2018 datasets show that our network outperforms other methods. The PyTorch code to reproduce the experiments is available at **https://github.com/beaupreda/domain-networks**.

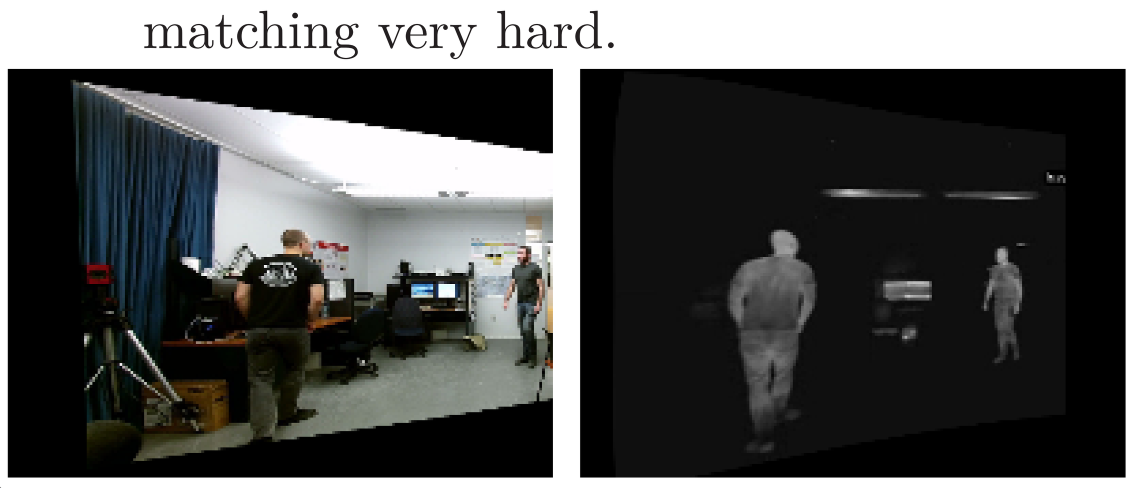
## 2. Introduction

**Benefits of working with multispectral stereo images:**

- The detection of an object, either in the visible or infrared domain, is always difficult when its contrast with the environment is low like someone wearing dark clothes at night.

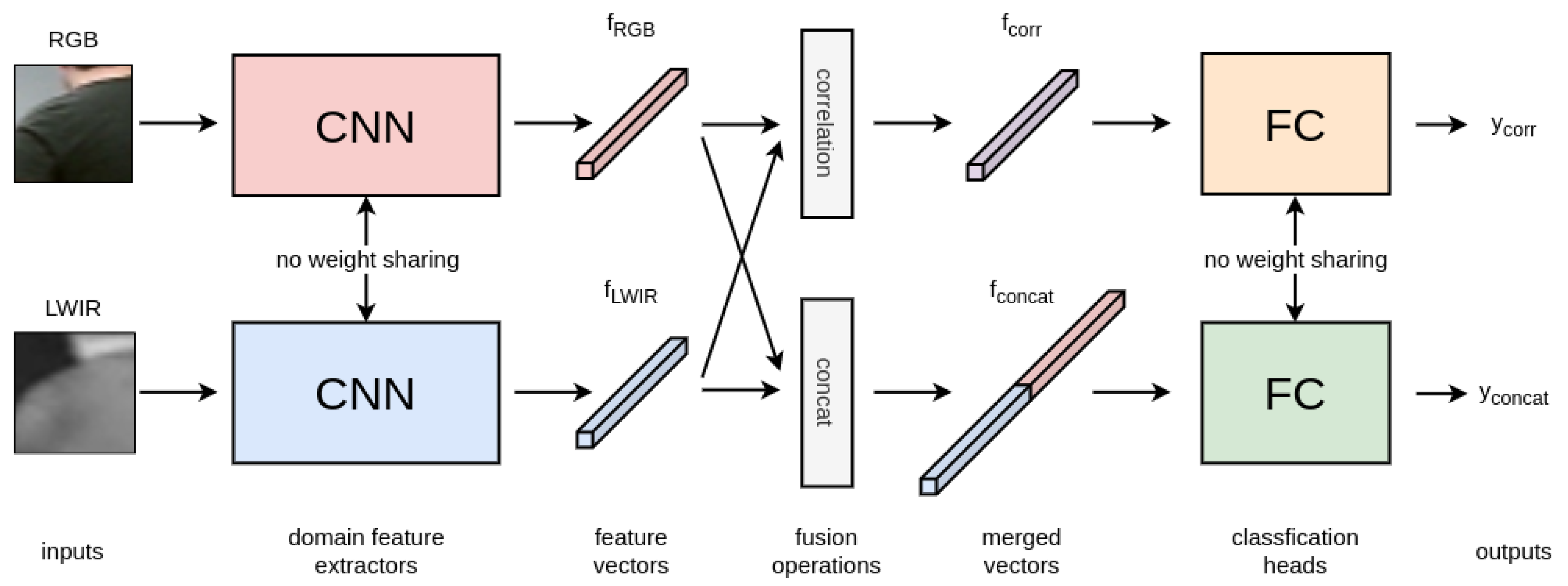
**Spectrum difference:**

- RGB-RGB pairs: there are a lot of similarities between the images (colors, textures) so matching is at its easiest.
- RGB-NIR pairs: it is more challenging than RGB-RGB, since we lose the color information, but we can still see some common textures between both images.
- RGB-LWIR pairs: it is very difficult since we only have information from objects emitting heat, and very few common textures between both images, which makes matching very hard.



## 3. Proposed method

- Extract the feature vectors ($f_{RGB}$ and $f_{LWIR}$) with a siamese CNN with **no weight sharing** for a pair of RGB and LWIR patches.
- Join the feature vectors with two fusion operations: **correlation** and **concatenation**.
- Forward each merged vector to its own classification head to get the probabilites that the input patches are similar.



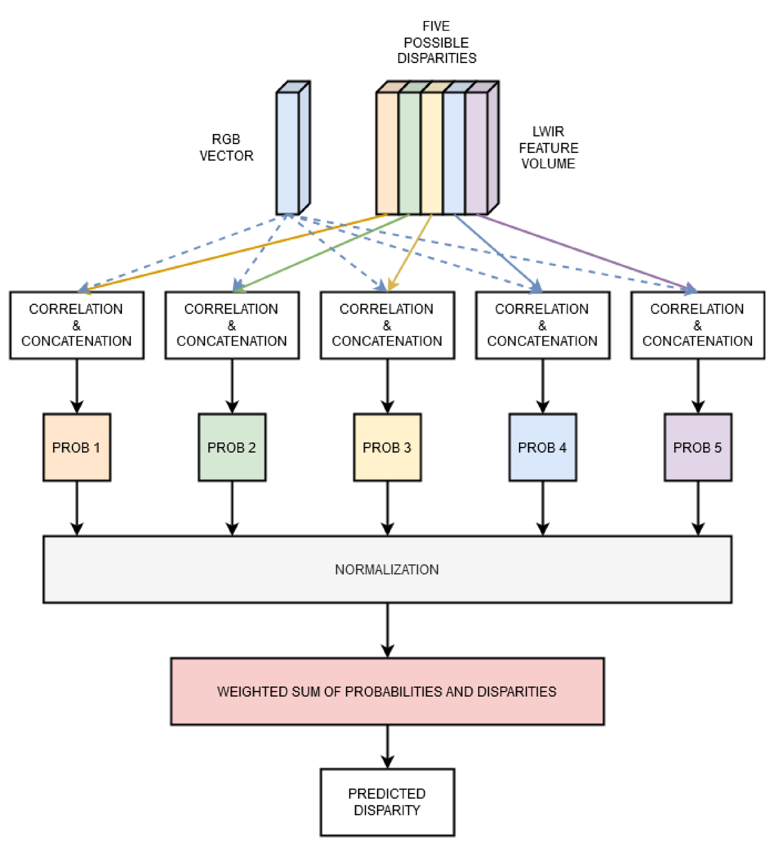

**Training:**

- Binary cross-entropy: $loss_{corr/concat} = -\frac{1}{N}\sum_{i=1}^{N} gt_i log(y_i)$

**Prediction:**

- Expand width of the LWIR patch to $disp_{max}$ and get the feature vectors from our network.
- At each disparity location in the feature volume LWIR, get the probability of finding the RGB patch and then do a regression to get the predicted disparity $\hat{d}_{corr/concat} = \sum_{d=0}^{disp_{max}} d \times p_d$.

## 3. Proposed method (continued)




## 4. Results

**Datasets:**

- LITIV 2014 [1] and LITIV 2018 [2] were used to evaluate our method, each dataset is separated into folds where we tested on one, and trained with the others.

Table 1: Ablation study on LITIV 2014 dataset illustrating the difference in recall with different configurations (correlation only, concatenation only, both). **Boldface**: best results.

| | Correlation branch only | | | Concatenation branch only | | | Corr + Concat (proposed model) | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\leq 1$ px | $\leq 3$ px | $\leq 5$ px | $\leq 1$ px | $\leq 3$ px | $\leq 5$ px | $\leq 1$ px | $\leq 3$ px | $\leq 5$ px |
| Fold 1 | 0.524 | 0.859 | 0.984 | 0.551 | 0.894 | 0.981 | **0.588** | **0.901** | **0.985** |
| Fold 2 | 0.454 | 0.854 | 0.978 | 0.472 | 0.897 | 0.985 | **0.474** | **0.904** | **0.986** |
| Fold 3 | 0.541 | 0.875 | 0.982 | 0.558 | 0.895 | 0.982 | **0.629** | **0.916** | **0.989** |

Table 2: Comparison of our proposed model against two other methods evaluated on the LITIV 2018 dataset. **Boldface**: best results.

| | Fold 1 | | Fold 2 | | Fold 3 | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | $\leq 1$ px | $\leq 4$ px | $\leq 1$ px | $\leq 4$ px | $\leq 1$ px | $\leq 4$ px | $\leq 1$ px | $\leq 4$ px |
| DASC Sliding Window [2] | 0.104 | 0.265 | 0.086 | 0.236 | 0.121 | 0.289 | 0.104 | 0.263 |
| Multispectral Cosegmentation [2] | 0.253 | 0.562 | 0.236 | 0.531 | 0.307 | 0.678 | 0.265 | 0.590 |
| Proposed Model | **0.480** | **0.943** | **0.446** | **0.877** | **0.406** | **0.972** | **0.442** | **0.930** |

Table 3: Comparison of our proposed model against classic and learned-based methods on the LITIV 2014 dataset. Patch sizes are in parentheses. **Boldface**: best results, *italic*: second best.

| Method | $\leq 3$ px |
|---|---|
| Proposed Model (36 × 36) | **0.906** |
| Siamese CNNs [3] (37 × 37) | 0.779 |
| Mutual Information [1] (40 × 130) | *0.833* |
| Mutual Information [1] (20 × 130) | 0.775 |
| Mutual Information [1] (10 × 130) | 0.649 |
| Fast Retina Keypoint [1] (40 × 130) | 0.641 |
| Local Self-Similarity [1] (40 × 130) | 0.734 |
| Sum of Squared Differences [1] (40 × 130) | 0.656 |

## 5. Conclusion

We proposed a new CNN architecture capable of estimating the disparity between images from the RGB and LWIR domains. Our CNN is able to extract robust features for matching corresponding regions, which leads to better results when compared to other classical descriptors and CNN-based methods.

## 6. References

[1] G.-A. Bilodeau and etal., "Thermal-visible registration of human silhouettes: A similarity measure performance evaluation," *Infrared Physics & Technology*, vol. 64, no. C, pp. 79–86, 2014.

[2] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Online mutual foreground segmentation for multispectral stereo videos," *IJCV*, Jan 2019.

[3] D.-A. Beaupre and G.-A. Bilodeau, "Siamese cnns for rgb-lwir disparity estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

## 7. Acknowledgements

CRSNG NSERC
NVIDIA