STRONG STEREO FEATURES FOR SELF-SUPERVISED PRACTICAL STEREO MATCHING

PIERRE-ANDRÉ BROUSSEAU & SÉBASTIEN ROY

Université de Montréal

CONTRIBUTION

We propose a hybrid method; a self-supervised feature encoder working with a classical matching algorithm.

1. A simple and practical self-supervised method to train **a feature encoder which can be readily integrated in an OpenCV stereo pipeline** and achieves competitive performance.

2. A novel method to **express permutation as a pretext task to obtain strong stereo features** that does not require hands-on knowledge of the dataset such as ground truth depth or scene content.

OVERVIEW

• Deep stereo algorithms show strong performances yet this shift from physics-model-driven to data-driven has not been followed by industrial adoption.

• **When stereo disparity is the only source of depth information, ground truth is rarely available for training supervised deep methods.**

• During training, our approach aims to **recover a strong feature representation**, i.e. it enables dense stereo algorithms to compute accurate disparity results.

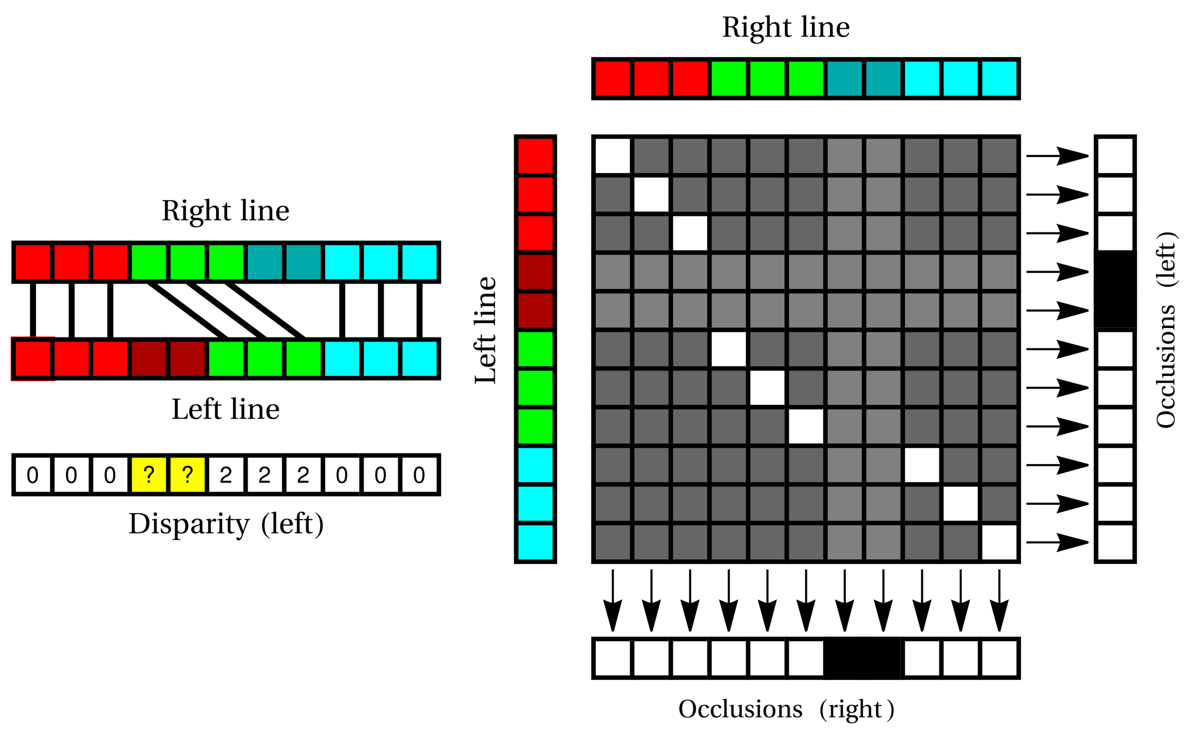
• At inference time, our method outputs a matching cost volume which is **directly integrated with industry standard classical stereo algorithms, such as the OpenCV stereoSGBM**, and leads to strong performances on natural image datasets.

PERMUTATION MODEL

The permutation provides **a natural representation of stereo constraints** by simultaneously representing:

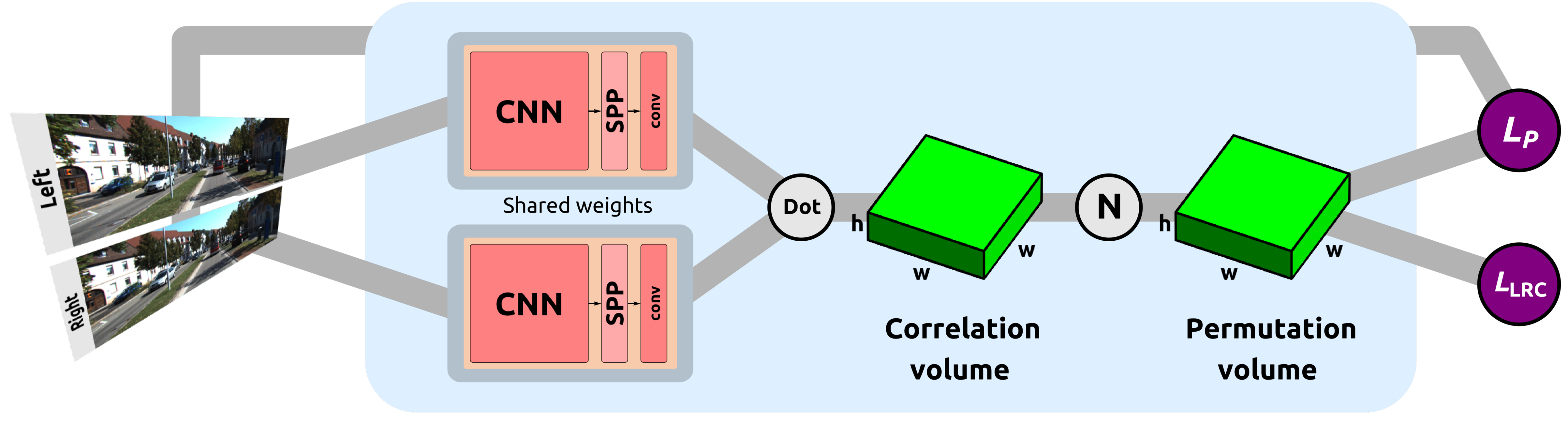
1. explicit **cross-attention** in left-right stereo pairs,

2. matching ambiguities such as **occlusions**, out-of-image pixels or textureless regions.



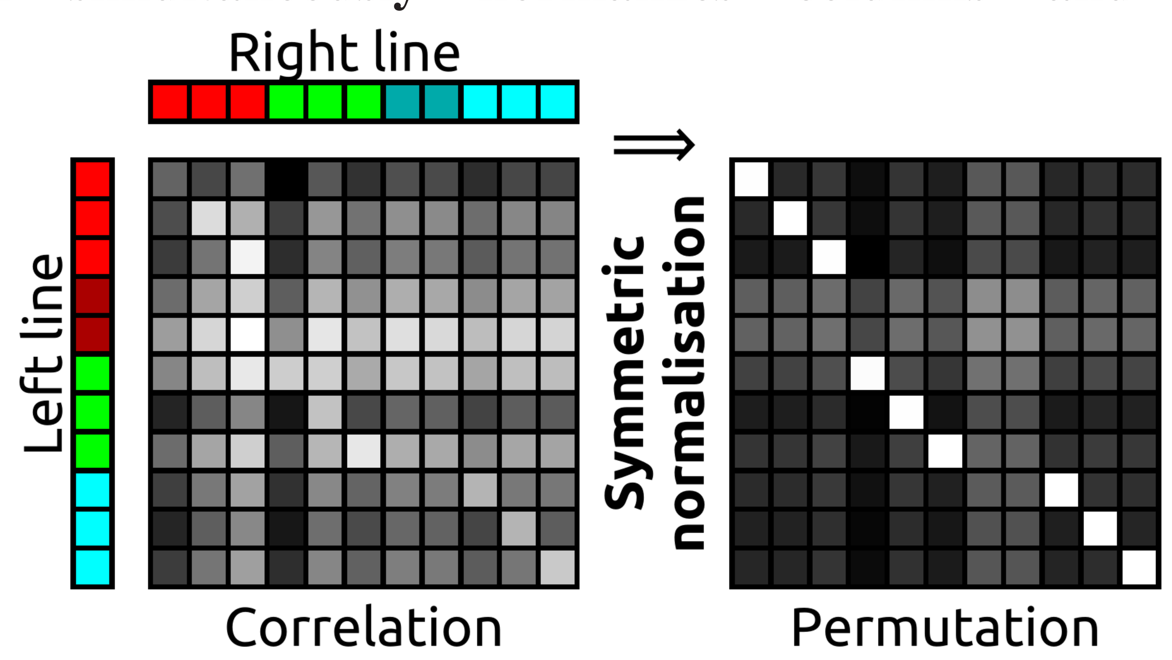
SELF-SUPERVISED MODEL FOR FEATURE LEARNING

Neural architecture that encourages a feature encoder to accurately represent images for the purpose of stereo matching.



TRAINING ON THE PERMUTATION PRETEXT TASK

By formulating stereo matching as an **optimal transport** problem, the iterative application of symmetric normalization **simultaneously normalizes columns and rows**.



Occlusions

$$O^L_{i,j} = \|P_{i,:,j}\|_2^2 \quad \text{and} \quad O^R_{i,j} = \|P_{i,j,:}\|_2^2$$

Photometric Loss

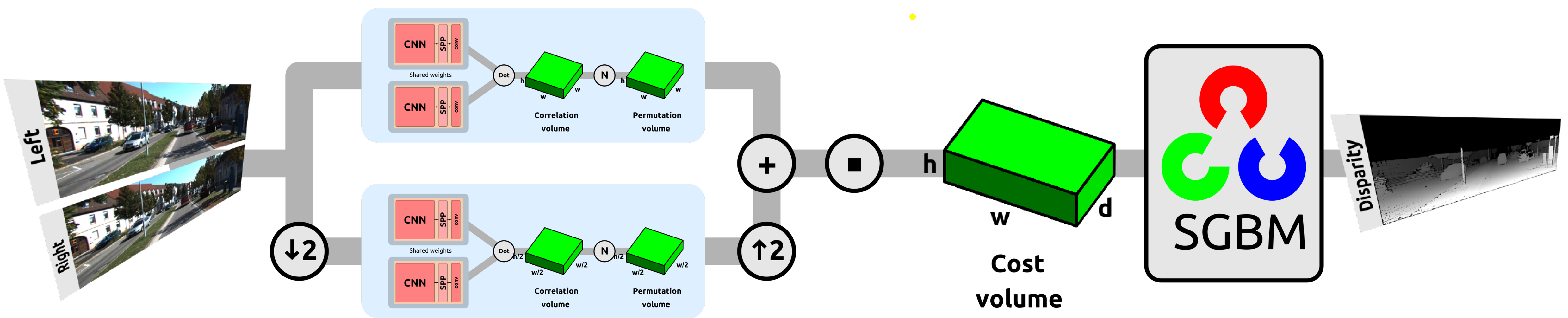
$$\widetilde{\mathcal{L}}_{\mathrm{P}} = \frac{1}{2}\left(\frac{\sum \mathcal{L}^L_{\mathrm{P}} \odot O^L}{\sum O^L} + \frac{\sum \mathcal{L}^R_{\mathrm{P}} \odot O^R}{\sum O^R}\right)$$

Left-Right Consistency Loss

$$\mathcal{L}_{\mathrm{LRC}} = \sum_i \|P_i \cdot P_i^\top - \mathbb{1}\|_1$$

PRACTICAL STEREO INFERENCE

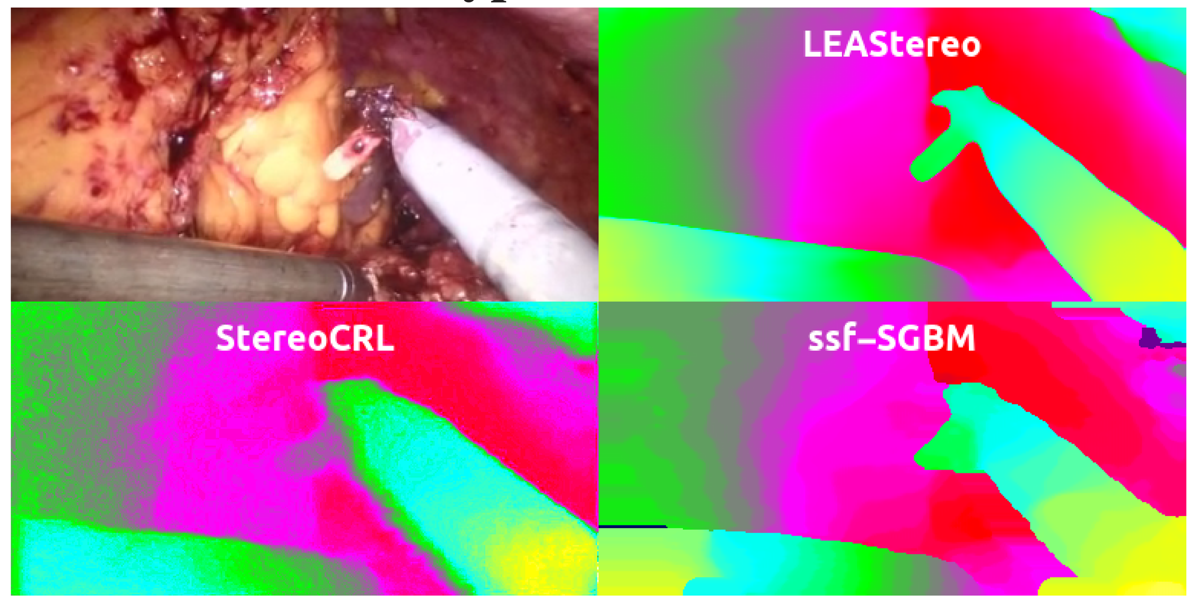
Pipeline to solve for disparity by providing the cost volume to a classical stereo method, such as the popular and publicly available stereoSGBM from OpenCV.



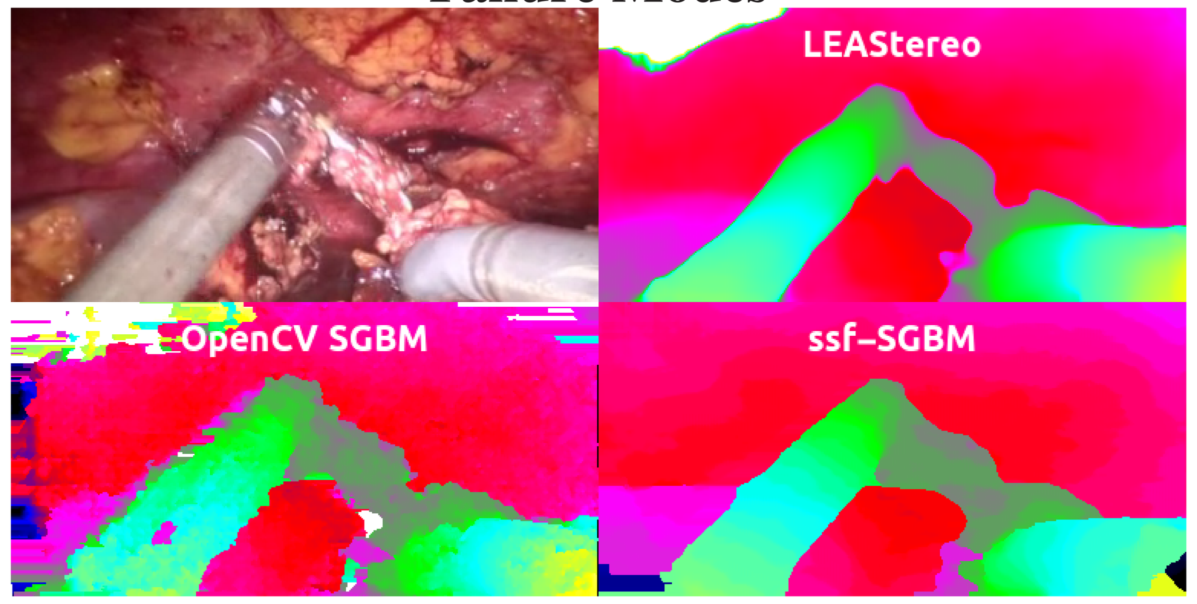
• **Monocular Disparity Completion.** Left-disparity propagation, the most naive disparity completion strategy. Is chosen because it does not introduce any additional knowledge to the disparity.

ENDOSCOPIC SCENES

Typical Result



Failure Modes



Comparison to State-of-the-Art

Methods	Mean SSIM	std. SSIM
ELAS	47.3	0.08
SPS	54.7	0.09
V-Siamese	60.4	0.07
StereoCRL	83.7	0.02
OpenCV SGBM	79.0	0.07
LEAStereo	83.9	0.05
ssf-SGBM(Ours)	84.4	0.05

DRIVING SCENES

Comparison to State-of-the-Art

	Method	Kitti 2015 (D1) fg	Noc	All
SGM	SGM	20.59	9.47	10.86
SGM	SGM_RVC	13.00	5.62	6.38
Self-Supervised	Zhou et al.	-	8.61	9.91
Self-Supervised	SegStereo	-	7.70	8.79
Self-Supervised	OASM-Net	19.42	7.39	8.98
Self-Supervised	PASMnet	16.36	6.69	7.23
Self-Supervised	Perm. Stereo	15.47	6.72	7.18
Self-Supervised	Flow2Stereo	14.62	6.29	6.61
Self-Supervised	CRD_Fusion	13.68	5.69	6.11
	ssf-SGBM(Ours)	13.81	5.77	6.41