Strong Stereo Features for Self-Supervised Practical Stereo Matching

Pierre-André Brousseau
pierre-andre.brousseau@umontreal.ca
Sébastien Roy
roys@iro.umontreal.ca

Lab V3D
Département d'informatique et
de recherche opérationnelle
Université de Montréal
Montréal, Canada

Abstract

This paper proposes a new approach to the problem of self-supervised dense stereo correspondence. Disparity estimation is an important problem of computer vision but in many situations, correspondence-based reconstruction cannot be accompanied by a ground truth to train supervised methods. This paper proposes to train a siamese feature encoder in a self-supervised permutation framework and then build a cost volume which is fed to a classical stereo algorithm to compute the disparity. In the absence of ground truth disparity, rather than handcrafting features, we suggest a novel and straightforward way to leverage input images to train for features. A key aspect of this method is that all the trainable weights are located inside a feature computation step, which is followed by strong non-trainable constraints that enforce bidirectional correspondence through cross-attention. Validated on real and synthetic datasets and compared to various methods, our proposed approach yields competitive results. Given its high performance, simplicity, and direct integration with current stereo algorithms, we expect this method to further the adoption of deep methods in real life stereo applications.

1 Introduction

Stereo depth estimation is the task of recovering depth by establishing a correspondence in a stereopscopic image pair. Once disparity is estimated along the epipolar lines between the left and the right image features, it is possible to compute depth by triangulating using the camera parameters. Disparity estimation is an important problem of computer vision and is used in autonomous driving, medical 3D tools such as stereo endoscopy and intra-oral scanners, robotics and many more applications.

Stereo matching algorithms based on deep neural networks have shown strong performances with respect to the traditional computer vision algorithms. This shift from physics-model-driven to data-driven[47] should have been followed by industrial adoption, especially when considering the large amount of footage being collected on a regular basis by 3D reconstruction devices. However, leveraging such data is not simple, as most state-of-the-art methods require ground truth depth information, which is not readily available. Many applications such as medical *in situ* scans of patients cannot rely on LIDAR or laser for ground truth acquisition. The problem at hand is that stereo depth estimation with deep networks has mostly progressed for applications where correspondence-based reconstruction is subordinate to more powerful modalities. When stereo disparity is the only source of depth

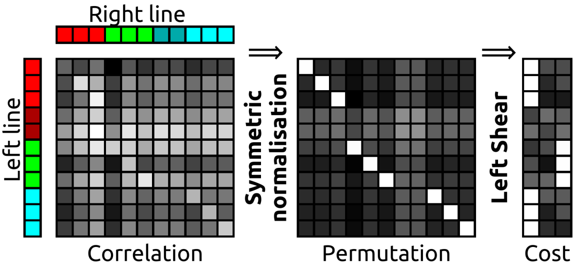

Figure 1: From Correlation to Cost. A correspondence volume is obtained from the dot product of left and right features. It is then normalized, and then sheared to provide a standard cost volume.

information, ground truth is rarely available for training supervised deep methods. Self-supervised learning-based stereo methods have been shown to yield strong performances, since accurate pretext tasks and constraints can compensate for the lack of ground truth. Self-supervised methods are immune to the problem of ground truth modality mismatch, which arises when non stereo sensors, such as time-of-flight, are used as ground truth disparity estimates. Moreover, we consider self-supervised methods to provide better explainability, given that they can't memorize a direct relationship between an image pattern and its disparity.

In this paper, we propose a hybrid method; i.e., a self-supervised feature encoder working with a classical matching algorithm. The contributions are as follow:

- 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.
- 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.

In this paper, a feature representation is deemed *strong* for stereo if it enables not only itself but other dense stereo algorithms to compute accurate disparity results. Inversely, a representation is deemed *weak* if it is not well suited for stereo estimation. At inference time, the output of our proposed method is not disparity but rather a *matching cost volume* which allows direct integration with industry standard classical stereo algorithms, such as the *OpenCV stereoSGBM*[3, 19] and leads to strong performances on natural image datasets. The implementation and supplementary content are available at:

`https://gitlab.com/labv3d/strong-stereo-features`.

2 Previous Work

Supervised Deep Stereo Matching. Stereo matching can be expressed as a direct regression task in an end-to-end deep learning pipeline. This can be done with an encoder-decoder style network such as DispNet[30]. Using a more explicit cost volume representation such as in GC-Net[21] and a softArgMax to mimick winner-take-all in an end-to-end fashion leads to improved performances. PSM-Net[6] proposes spatial pyramid pooling and stacked hourglass networks to aggregate costs at different scales. GwcNet[16] extends this by proposing group-wise correlations, a more efficient representation for measuring feature similarities. GA-Net[43] proposes two novel aggregation layers as approximation to semi-global matching. LEAStereo[9] explores this problem using neural architecture search to optimize the stereo architecture. RAFT-Stereo[26] adapts RAFT[35] for stereo where the network iteratively refines the disparity estimates using multi-level GRU units. STTR[25] tackles disparity estimation with a transformer architecture by relating stereo matches to attention. PCW-Net [33] model for two cost volumes, a multi-scale combination volume which yields domain-invariant features and a warping volume which refines the disparity estimates. Undeniably, deep stereo matching has shown strong results on benchmarks. However, they have not yet reached deployment level in the industry [23], which still heavily relies on traditional semi-global matching.

Supervised Learning-based SGM Methods. Semi-global matching (SGM)[18] is a powerful method for estimating stereo correspondence and it has been added to the standard CV library *OpenCV stereoSGBM*[19]. Handcrafting features for specific scenes such as the census transform for reflective surfaces significantly increases performance and robustness but requires expert hands-on knowledge of scene content. MC-CNN[42] proposes to create a dataset of image patches using the ground truth disparities and train a siamese convolutional neural network to predict a matching cost which is then fed to SGM. Such a pipeline has been used extensively[29]. Luo et al.[28] propose to define the matching cost as the dot product of learned feature vectors and then solve it using SGM. While removing the need for handcrafting, these supervised learning methods for strong features can be impractical as they require ground truth signal.

Self-Supervised Methods. Many stereo-vision applications, such as stereo-endoscopy, cannot rely on other 3D reconstruction modalities, such as time-of-flight sensors, to provide ground truth depth. For deep networks, the stereo task is redefined as an image warping pretext task [45] with a left-right consistency constraint[15]. Ye et al.[41] use autoencoders to jointly predict left and right disparities. OASM-Net[24] integrates a cost volume representation and models occlusion explicitly. PASMnet[37] adds the epipolar constraint as attention between features in a cascaded parallax-attention module. Permutation Stereo[4] explicitly models disparity as a permutation, a constrained version of attention. Chen et al.[8] revisit the encoder-decoder structure but with the PWC-Net[34] backbone in a knowledge distillation framework. Flow2Stereo [27] uses a teacher-student framework to jointly learn optical flow and stereo. CRD_Fusion[11] distills knowledge from the SGM algorithm, it uses the raw predicted disparity maps as a supervision term. More recently, StereoCRL[36] performs contrastive representation learning to obtain features which are then integrated in a standard deep self-supervised network. These methods do not require ground truth supervision signal but depart significantly from industry standards.

Our proposed approach is a novel self-supervised learning-based SGM method. This is a promising hybrid[47] pathway to more explainable deep stereo and industry adoption.

3 Self-Supervised Features for Stereo

Given a rectified stereo pair of I^L, I^R images, our approach aims to recover a strong feature representation f^L, f^R of both images for traditional stereo vision algorithms. Because computing the cost volume on a latent feature space leads to better disparity scenes[42], this work proposes a self-supervised learning method, illustrated in Fig. 2, that encourages a feature encoder to accurately represent images for the purpose of stereo matching.

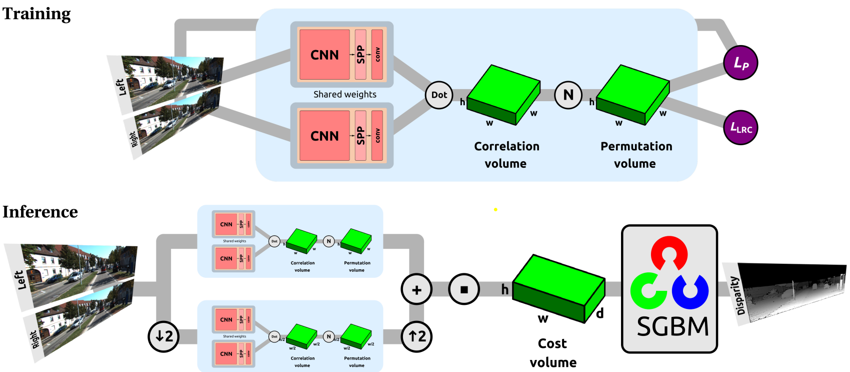

Figure 2: Training architecture and inference pipeline. (TOP) The stereo pair is processed by a siamese feature encoder F. The **Dot** operation compares features into a Correlation Volume. The **N** operation applies symmetric normalization. The losses are $\widetilde{\mathcal{L}}_{\text{P}}$ and $\mathcal{L}_{\text{LRC}}$. (BOTTOM) The stereo pair is converted into permutation volumes at two different scales which are summed, sheared into a matching cost and fed into OpenCV stereoSGBM. The ↓ **2** and ↑ **2** indicate downsampling and upsampling by a factor of 2. The ■ applies left-shear.

Correlation Volume. We use a siamese feature encoder F where the left and right images are processed independently. It projects both images in the learned feature space f^L and f^R

$$f^L = F(I^L), \qquad f^R = F(I^R). \tag{1}$$

As was introduced in Luo et al.[28], we use the inner dot product between the two representations but in our case it is done at full image scale rather than image patches. This leads to the concept of *Correlation Volume*, also designated as such in Lipson et al.[26], defined as

$$C_{i,j,k} = \sum_g f^L_{i,j,g} \cdot f^R_{i,k,g} \quad \text{where} \quad C \in \mathbb{R}^{H \times W \times W} \tag{2}$$

and i denotes a horizontal line. The images are of size $H \times W$.

The comparison of features is formulated as a simple dot product and has the benefit of not imposing a maximum disparity value. We consider the dot product as a natural choice since it yields high positive values for good matches and low or negative values for mismatches, and alleviates the need for further processing before normalization. The correlation volume closely relates to feature cross-attention.

Permutation Model. The permutation formulation introduced in Brousseau et al.[4] defines a permutation volume P between the left image I^L and the right image I^R such that

$$I^R_i = P_i \cdot I^L_i, \quad I^L_i = P^\top_i \cdot I^R_i \quad \text{where} \quad P \in \mathbb{R}^{H \times W \times W}. \tag{3}$$

A permutation P allows a pixel to match any stereo-visible pixel, which is defined as a pixel visible in both the left and right images, along the corresponding horizontal line in the correlation volume C. In this work, we compute the permutation volume by directly normalizing the correlation volume using symmetric normalization,

$$P^{t=0}_{i,j,k} = exp(C_{i,j,k}), \quad P^{t+1}_{i,j,k} = \frac{P^t_{i,j,k}}{\sqrt{\sum_m P^t_{i,m,j} \sum_m P^t_{i,j,m}}}. \tag{4}$$

By formulating stereo matching as an optimal transport problem, the authors in [4] adapt the Sinkhorn-Knopp[22] algorithm such that it simultaneously normalizes columns and rows. The iterative application of this normalization ensures that the correlation volume becomes doubly stochastic, where all rows and columns sum to one. The permutation volume is therefore a matching cost that explicitly describes cross-attention along the image rows.

Modelling occlusions as part of the matching process has been shown to improve stereo performances[8, 24]. Permutation implicitly encodes matching ambiguity such as textureless regions and occlusions through higher entropy in rows or columns. Beside computing entropy, this occlusion information can easily be approximated by the squared norms

$$O^L_{i,j} = \|P_{i,:,j}\|_2^2 \quad \text{and} \quad O^R_{i,j} = \|P_{i,j,:}\|_2^2. \tag{5}$$

Training the Permutation Task. Training for self-supervised permutation involves a photometric loss $\mathcal{L}_\mathrm{P}$. It relies on the structural similarity index (SSIM)[38] and the L_1-norm applied to the input image pair as

$$\begin{aligned} \mathcal{L}^R_\mathrm{P} &= \frac{\alpha}{2}(1-\mathrm{SSIM}(I^R, I^{R*})) + (1-\alpha)\|I^R - I^{R*}\|_1, \qquad & I^{R*}_i = P_i \cdot I^L_i \\ \mathcal{L}^L_\mathrm{P} &= \frac{\alpha}{2}(1-\mathrm{SSIM}(I^L, I^{L*})) + (1-\alpha)\|I^L - I^{L*}\|_1, \qquad & I^{L*}_i = P_i^\top \cdot I^R_i \end{aligned} \tag{6}$$

The permutation formulation further enables for an occlusion-aware photometric loss $\widetilde{\mathcal{L}}_\mathrm{P}$ as

$$\widetilde{\mathcal{L}}_\mathrm{P} = \frac{1}{2}\left(\frac{\sum \mathcal{L}^L_\mathrm{P} \odot O^L}{\sum O^L} + \frac{\sum \mathcal{L}^R_\mathrm{P} \odot O^R}{\sum O^R}\right) \tag{7}$$

where the $\odot$ operator indicates element-wise multiplication.

The left-right consistency loss $\mathcal{L}_\mathrm{LRC}$ enforces the uniqueness constraint as well as bijective matches. In the permutation formulation, this *round-trip* constraint is equivalent to orthogonality, which can be readily measured as

$$\mathcal{L}_\mathrm{LRC} = \sum_i \|P_i \cdot P_i^\top - \mathbb{1}\|_1 \tag{8}$$

where $\mathbb{1}$ is the identity matrix. The training loss $\mathcal{L}$ is the weighted sum of the occlusion-aware photometric loss and the left-right consistency loss $\mathcal{L} = \widetilde{\mathcal{L}}_\mathrm{P} + \lambda \mathcal{L}_\mathrm{LRC}$.

The permutation formulation proposed in [4] is used in this paper, however it serves as a pretext task to train for features rather than as the disparity inference objective. During training, the permutation model makes it possible to define an image cross-attention estimation task which is entirely self-supervised. As shown in Fig. 2, an interesting characteristic of our proposed model is the absence of learnable weights or layers which would act on the correlation volume or the permutation volume. This compels the siamese feature encoder weights to encode the stereo matching process as completely as possible.

4 Inference Pipeline

During training, our approach is specifically learning strong features for stereo. Beside the feature encoder, modern stereo methods also use a variety of approaches to propagate matching costs, manage occlusion and select the best disparity for each pixel. In deep stereo methods, this is often achieved with 3D convolutions. Instead, this paper proposes to solve for disparity by providing the cost volume to a classical stereo method, such as the popular and publicly available SGM method: stereoSGBM from the OpenCV library[19]. By separating the training and inference steps, and focusing the learning on the feature encoder, the training can be done in isolation, without ground truth depth on any dataset, and then easily integrated into a functioning stereoSGBM pipeline.

Matching Cost Computation. At inference time, the input stereo pair is processed at full resolution as well as half resolution to yield two permutation volumes which are then added (see Fig. 2). By feeding the input images at a lower resolution, the recovered features can better express lower frequency signals present in the images. The summed permutation volume can then be converted to a cost volume $\bar{C}$ using the shear operation (see Fig. 1)

$$\text{Left Shear}: \; P_{i,j,j-d} \rightarrow \bar{C}^{L}_{i,j,d} \quad \text{and} \quad \text{Right Shear}: \; P_{i,j+d,j} \rightarrow \bar{C}^{R}_{i,j,d}. \tag{9}$$

The left-shear operation results in a cost volume $\bar{C}^L \in \mathbb{R}^{H\times W\times D}$ where D covers the range of allowed disparities, which are positive upto a selected maximum disparity value.

SGM Aggregation & Disparity Computation. The Semi-global Matching (SGM) [18] stereo method uses a matching cost based on mutual information, which is then aggregated along eight directions over two passes. The disparity computation step returns the minimum aggregated cost located with subpixel accuracy, and then proceeds to refine it by removing speckle. This method is a gold standard in stereo matching and is very much used today in a variety of devices. The OpenCV[3, 19] library offers an implementation of the SGM method called "stereoSGBM" which uses the Birchfield et al.'s[2] matching cost function.

This current work makes use of the stereoSGBM method by providing the computed left matching cost volume $\bar{C}^L$ directly to the cost aggregation step[18].

Monocular Disparity Completion. The monocular depth estimation task is relevant when depth must be estimated from a single image. This task is important in many stereo testing datasets since stereo disparity benchmarks often require disparity estimates at each pixel even when they are occluded or match an out-of-image pixel. This situation can be caused by modality mismatch (i.e. using LIDAR for evaluating stereo), or synthetic datasets which provide disparities for all pixels, stereo-visible or not. Therefore, self-supervised stereo methods must be accompanied during inference by a heuristic to estimate disparity in non stereo-visible regions.

As presented in [18], the consistency check and the speckle filtering in SGM may declare some pixels as invalid for disparity estimation. In this work, we adopt the most naive disparity completion strategy, which is the left-disparity propagation. For pixels that are identified as inconsistent, we propagate the first valid left-disparity. For pixels declared invalid because they would match out-of-image pixels, we instead propagate the first valid right-disparity. This very simple completion heuristic is chosen because it highlights the performance of the stereo matching method as it does not introduce any additional knowledge to the disparity maps. More precisely, this chosen heuristic works adequately because our method recovers disparity where it is available and accurately identifies the pixels where it cannot.

5 Experiments and Discussion

For detailed information on the datasets and implementation details, see supplementary material. Our feature encoder architecture is PSMnet's, detailed in [6]-Tab. 1. This architec-

	Method	Kitti 2015 (D1)			
		bg	fg	Noc	All
SGM	SGM [18]	8.92	20.59	9.47	10.86
	SGM_RVC	5.06	13.00	5.62	6.38
Self-Supervised	Zhou et al. [46]	-	-	8.61	9.91
	SegStereo [40]	-	-	7.70	8.79
	OASM-Net [24]	6.89	19.42	7.39	8.98
	PASMnet [37]	5.41	16.36	6.69	7.23
	Perm. Stereo [4]	5.53	15.47	6.72	7.18
	Flow2Stereo [27]	5.01	14.62	6.29	6.61
	CRD_Fusion [11]	4.59	13.68	5.69	6.11
	ssf-SGBM(Ours)	4.93	13.81	5.77	6.41

Table 1: Quantitative results. Comparison of ssf-SGBM with SGM and self-supervised methods on the Kitti 2015[14] testing datasets. Results are available online. Lower is better.

Methods	Mean SSIM	std. SSIM
ELAS [12]	47.3	0.08
SPS [39]	54.7	0.09
V-Siamese [41]	60.4	0.07
StereoCRL [36]	83.7	0.02
OpenCV SGBM	79.0	0.07
LEAStereo [9]	83.9	0.05
ssf-SGBM(Ours)	84.4	0.05

Table 2: Quantitative results. Comparison of ssf-SGBM with self-supervised methods and LEAStereo trained on SceneFlow[30] on the DaVinci Si Hamlyn[41] testing set. Higher is better for Mean SSIM.

ture was chosen as it is well detailed, well-known, and most importantly to demonstrate that the contribution of this paper does not lie in the feature architecture by itself.

5.1 Kitti Benchmark

Quantitative results on the testing dataset Kitti 2015 are in Tab. 1. Our results are available on the leaderboard online under the name *ssf-SGBM*. Training is done exclusively on the training set for 1000 epochs. Tab. 1 splits competing methods into two categories, SGM and self-supervised.

Very importantly, ssf-SGBM outperforms semi-global matching from Hirschmuller et al.[18] which uses the mutual information matching cost function and performs similarly to SGM_RVC (2020 Robust Vision Challenge) which uses the census matching cost. This result supports that our self-supervised pretext task results in features for stereo that contend to expertly selected ones and significantly improves upon standard mutual information.

In the comparison to learning-based self-supervised methods, ssf-SGBM performs favorably with respect to PASMnet[37], which models attention, and Permutation Stereo[4], which relies on the permutation cross-attention model. Although CRD_Fusion[11] outperforms our method, both it and Flow2Stereo[27] use knowledge distillation to boost performances which amounts to implicit ensembling [1, 17]. Ssf-SGBM improves upon Flow2Stereo on all metrics. This indicates that the strong constraints of our approach perform as well as a state-of-the art distillation method. Our method yields competitive performance on a robotic vision task with respect to state-of-the-art self-supervised methods or expertly handcrafted features.

5.2 Surgical Applications

Depth estimation serves in a variety of applications of which many depend on stereo-vision because it is the most reliable passive way of depth acquisition. Sometimes, it is impossible to use active sensors, such as LIDAR technologies or laser scanners, and therefore ground truth depth can never be measured. In such situations, it is hard to quantify how effective self-supervised disparity estimation is with respect to supervised disparity estimation. Obviously, supervised methods outperform self-supervised methods on benchmarks such as Kitti 2015 where the LIDAR ground truth is available for training. For the DaVinci Si Hamlyn dataset, however, no ground truth is available as the image pairs are captured during robotic surgery.

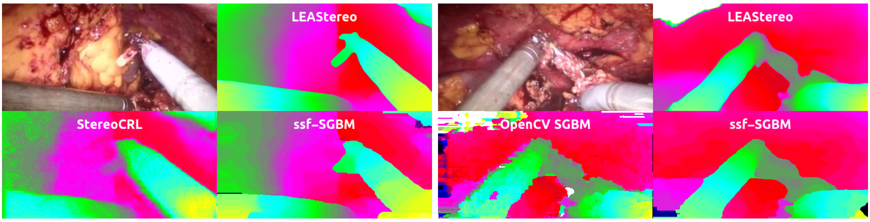

Figure 3: Qualitative results for stereo-endoscopic surgery (LEFT), for image pair 2680 of the DaVinci Si Hamlyn testing set. Examples of failure modes (RIGHT), for image pair 2341.

Method	Sintel (D1) Stereo	Sintel (D1) Mono	Kitti 2012 (D1) Stereo	Kitti 2012 (D1) Mono	Middlebury 2014 (D1) Stereo	Middlebury 2014 (D1) Mono
OpenCV SGBM	14.63	72.76	14.89	69.61	17.57	70.12
PASMnet_192 [37]	8.71	72.31	6.55	41.83	16.45	65.54
CRD-Fusion [11]	8.25	52.81	5.73	45.98	12.60	71.06
ssf-SGBM(Ours)	10.30	49.80	6.32	37.53	14.74	47.88

Table 3: Generalization results. Trained on SceneFlow and tested on Sintel[5]. Trained on Kitti 2015 and tested on Kitti 2012[13] and Middlebury 2014[32]. The D1 error is for stereo-visible pixels (D1-Stereo) and stereo-occluded pixels (D1-Mono). Lower is better.

Results are in Tab. 2 and in Fig. 3 for the DaVinci Si Hamlyn testing set. Comparison is made with StereoCRL[36], the recent state-of-the-art method on this dataset, and with LEAStereo[9] a gold standard in supervised deep stereo matching trained on SceneFlow. This experiment illustrates a real world situation where synthetic data with ground truth is available for training but no ground truth for inference can be obtained. In this situation, one must choose to either use a supervised method in cross-domain generalization or use a self-supervised method.

In Tab. 2, ssf-SGBM slightly outperforms the state-of-the-art method StereoCRL and LEAStereo on the Mean SSIM metric. This illustrates the advantage of our self-supervised approach in some real-world applications over much stronger supervised methods. In Fig. 3 (LEFT), ssf-SGBM sharply identifies the tools edges while StereoCRL yields a blurrier result. In Fig. 3 (RIGHT), LEAStereo hallucinates false matches, the white mass at top left, while ssf-SGBM tends to smooth over occluded discontinuities. LEAStereo generally results in good looking disparity estimates but it can hallucinate false disparities. This is why LEAStereo qualitatively appears better but does not outperform ssf-SGBM. The SSIM error metric is relevant to optical metrology and thus favors robustness over hallucinations and outliers. In this setting, our method enforces explainable matches and significantly improves OpenCV SGBM.

5.3 Generalization Performance

During training, our feature encoder treats images from a stereo pair independently. The correlation volume with the symmetric normalization prevents our feature encoder from

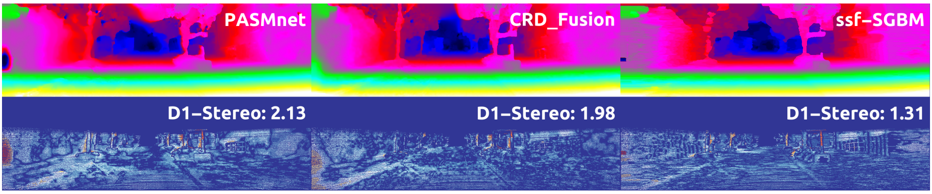

Figure 4: Kitti 2012 generalization results, trained on Kitti 2015. (LEFT) Output disparity and (RIGHT) D1-Error map for PASMnet, CRD_Fusion and our method.

Method	Aggregation & Disparity Computation	SceneFlow D1-Stereo	SceneFlow EPE-Stereo	Kitti 2015 D1-Stereo	Kitti 2015 EPE-Stereo
ssf-WTA	Winner-Take-All	6.82	2.97	8.21	2.34
ssf-SoftArgMax	× Disparity Range	5.87	2.79	9.84	1.89
ssf-MFMC	MaxFlow MinCut [31]	6.08	2.18	6.57	1.46
ssf-DP	Dynamic Programming	5.99	2.17	6.10	1.38
ssf-SGBM	openCV SGBM [20]	5.58	2.00	5.39	1.23

Table 4: Ablation study for the aggregation and disparity computation. The errors are for stereo-visible pixels (D1-Stereo and EPE-Stereo). Lower is better.

Capacity	Nbr. of Params.	Kitti 2015 D1 All	Kitti 2015 D1 Stereo	Kitti 2015 EPE All	Kitti 2015 EPE Stereo	Middlebury 2014 D1 All	Middlebury 2014 D1 Stereo	Middlebury 2014 EPE All	Middlebury 2014 EPE Stereo
Full Channels	3.49M	5.51	5.39	1.26	1.23	18.60	15.38	3.28	2.69
Half Channels	0.88M	5.76	5.65	1.27	1.25	18.29	15.13	3.25	2.70
Quarter Channels	0.22M	6.40	6.27	1.33	1.30	19.69	16.32	3.53	2.95

Table 5: Ablation study for the capacity of the feature encoder. The errors are for all pixels (D1-All and EPE-All) and stereo-visible pixels (D1-Stereo and EPE-Stereo). Lower is better. Note: Training is 500 epochs.

resolving this task as a single image depth estimator and consequently mitigates its shortcomings [7]. A reasonable way of confirming that our method indeed learns features tuned for generic stereo is to measure generalization performance.

Quantitative generalization results are presented in Tab. 3 and Fig. 4 shows a qualitative result with the visual depiction of the error maps. Sintel results are obtained using the networks trained on SceneFlow while Kitti 2012 and Middlebury 2014 results are obtained using networks trained on Kitti 2015 training set. Results are presented separately for stereo-visible (Stereo) and stereo-occluded (Mono). These regions are computed from the ground truth disparity[4, 44]. Comparison is made with the publicly available methods PASMnet[37] and CRD_Fusion[11] both state-of-the-art self-supervised stereo matching methods which have made t,heir code and weights available. Parallax attention[37] has been well adopted as a strong stereo representation while Confidence Guided Raw Disparity Fusion[11] is the current self-supervised best performing method. Comparison with openCV SGBM's implementation of SGM is done as an important performance guideline.

On real-world images, ssf-SGBM achieves lower errors when compared to SGBM and PASMnet, suggesting that its performance translates to generalization. Regarding results for monocular predictions, the chosen left-fill heuristic is very simple, yet it performs better in stereo-occluded regions than both competing self-supervised methods. Our explanation is that self-supervised neural networks must predict disparity everywhere, even where it is stereo-occluded. In our case, ssf-SGBM will usually identify a pixel as invalid rather than propose an incorrect disparity. Our heuristic provides disparity in stereo-occluded regions and naturally integrates with the ssf-SGBM's occlusion detection process. This is visible in the the leftmost areas of Fig. 4, which are stereo-occluded, where the competing methods provide poor monocular estimates. These results indicate that strong stereo constraints allow to identify where stereo cannot be solved and represent, in our opinion, a step forward in explainability.

5.4 Ablation Studies

Cost aggregation & Disparity selection. Ablation results in Tab. 4 aim to evaluate if our strong features yield viable disparity maps for various cost aggregation and disparity computation methods and to confirm the choice of OpenCV SGBM for cost aggregation. The Winner-Take-All (WTA) and SoftArgMax methods for disparity computation do not have any mechanism for propagation or disparity smoothness. Consequently, these methods recover the smallest details but obtain the worst overall performances, especially on Kitti 2015, as they mostly neglect speckle and are unable to disambiguate difficult matches. The MaxFlow MinCut (MFMC)[31] method and Dynamic Programming (DP) perform strong cost aggregation which leads to results that have smooth disparities, which explain the better overall performances. Most notably, MFMC propagates very strongly in both the horizontal and vertical direction as it searches for a global solution whereas DP solves the problem for individual horizontal lines. OpenCV SGBM leads to overall best performance on both sets for both metrics as it strikes a good balance between preserving details and overall smoothness.

Feature Encoder Capacity. Ablation results in Tab. 5 aim to evaluate if the strong stereo constraints promote the desired behaviour in the downstream disparity task when integrated with openCV SGBM. In Tab. 5, *Half Channels* has a small performance loss on the Kitti 2015 D1 metric but not for EPE. However, there is no loss in Middlebury generalization performance as the D1 metrics are actually lowered and the EPE errors are similar. This indicates that a higher capacity leads to learned features that are more tuned to the dataset but slightly lose generalization performance. *Quarter Channels* presents a performance degradation of less than 1% in D1-All and D1-Stereo, in both Kitti and Middlebury, with a network capacity reduced by a factor of 16. This illustrates that our method's performance does not significantly depend on the number of parameters, but mostly relies on the strong constraints of the permutation pretext task combined with openCV SGBM.

6 Limitations

This paper trains a strong stereo feature architecture by leveraging the constraints of the permutation model. However, it does not explore which deep neural network architecture leads to the best performing features, as the aim of this paper is to investigate how stereo features can be learned in a self-supervised context and how they can be integrated into a functioning standard stereo pipeline. For stereo-occluded pixels, this paper uses a very naive left-fill completion heuristic. It does not explore other stronger means of completion such as knowledge distillation [8, 11] or a convolutional network [37]. This choice is on purpose to highlight the performance of the features with respect to occlusions and SGM's aggregation. Finally, this paper does not explore stronger aggregation methods such as MGM[10] or the other SGM derivatives since OpenCV SGBM is the most widely accepted standard.

7 Conclusion

This paper revisits the key idea that learned features can bolster performance of classical stereo matching. The proposed approach demonstrates how these features can be learned effectively with a self-supervised permutation formulation on a variety of datasets. Rather than handcrafting, we suggest a novel and straightforward way to leverage only input images to train for strong features and results show that our method compares favorably with current state-of-the-art self-supervised methods. It is further demonstrated that in circumstances where no ground truth is available, our proposed method is on par with state-of-the-art supervised methods. Ssf-SGBM can be readily integrated in industry standard stereo matching pipelines and increases performance while providing a much higher level of adaptability to the application-specific stereo capture conditions, such as biomedical stereo-endoscopy.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[2] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (4):401–406, 1998.

[3] Gary Bradski and Adrian Kaehler. Opencv. *Dr. Dobb's journal of software tools*, 3: 120, 2000.

[4] Pierre-André Brousseau and Sébastien Roy. A permutation model for the self-supervised stereo matching problem. In *2022 19th Conference on Robots and Vision (CRV)*, pages 122–131. IEEE, 2022.

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[6] 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.

[7] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016.

[8] Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, and Liusheng Huang. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15529–15538, 2021.

[9] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33: 22158–22169, 2020.

[10] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt. Mgm: A significantly more global matching for stereovision. In *BMVC 2015*, 2015.

[11] Xiule Fan, Soo Jeon, and Baris Fidan. Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion. In *2022 19th Conference on Robots and Vision (CRV)*, pages 132–139. IEEE, 2022.

[12] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. The kitti vision benchmark suite. *URL http://www.cvlibs.net/datasets/kitti*, 2(5), 2015.

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[16] 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.

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[18] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[19] *The OpenCV Reference Manual*. Itseez, 2.4.9.0 edition, April 2014.

[20] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

[21] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017.

[22] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[23] Alexander Lavin, Ciarán M Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, Atílím Güneş Baydin, Amit Sharma, Adam Gibson, et al. Technology readiness levels for machine learning systems. *Nature Communications*, 13(1):6039, 2022.

[24] Ang Li and Zejian Yuan. Occlusion aware stereo matching via cooperative unsupervised learning. In *Asian Conference on Computer Vision*, pages 197–213. Springer, 2018.

[25] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021.

[26] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[27] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6648–6657, 2020.

[28] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016.

[29] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

[30] 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.

[31] S. Roy and I.J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Sixth International Conference on Computer Vision (ICCV)*, pages 492–499, 1998. doi: 10.1109/ICCV.1998.710763.

[32] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.

[33] Zhelun Shen, Yuchao Dai21, Xibin Song11, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching.

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[36] Samyakh Tukra and Stamatia Giannarou. Stereo depth estimation via self-supervised contrastive representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 604–614. Springer, 2022.

[37] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[39] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014.

[40] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, 2018.

[41] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.

[42] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015.

[43] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[44] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020.

[45] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

[46] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[47] Chao Zuo, Jiaming Qian, Shijie Feng, Wei Yin, Yixuan Li, Pengfei Fan, Jing Han, Kemao Qian, and Qian Chen. Deep learning in optical metrology: a review. *Light: Science & Applications*, 11(1):39, 2022.