A Permutation Model for the Self-Supervised Stereo Matching Problem

Pierre-André Brousseau
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, Canada
pierre-andre.brousseau@umontreal.ca

Sébastien Roy
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, Canada
roys@iro.umontreal.ca

***Abstract*—This paper proposes a novel permutation formulation to the stereo matching problem. Our proposed approach introduces a permutation volume which provides a natural representation of stereo constraints and disentangles stereo matching from monocular disparity estimation. It also has the benefit of simultaneously computing disparity and a confidence measure which provides explainability and a simple confidence heuristic for occlusions. In the context of self-supervised learning, the stereo performance is validated for standard testing datasets and the confidence maps are validated through stereo-visibility. Results show that the permutation volume increases stereo performance and features good generalization behaviour. We believe that measuring confidence is a key part of explainability which is instrumental to adoption of deep methods in critical stereo applications such as autonomous navigation.**

***Keywords*-Permutation; Stereo; Self-Supervised; Occlusions;**

I. Introduction

Stereo matching and stereo depth estimation is the process by which two simultaneously acquired images are put in correspondence to estimate the depth of the captured scene. This problem has been thoroughly investigated as it relates to the human vision and has practical applications. The depth estimation task has surged in popularity in recent years especially for autonomous navigation tasks and their associated datasets.

For many years, while they held the top positions on leaderboards, deep network approaches to depth estimation have had explainablity concerns. As vision systems are becoming more accessible and better integrated in everyday applications, stronger guarantees of their behaviour under various scene conditions are expected. While failures in the vision system of a smart vacuum cleaner pose little risks, some devices such as autonomous vehicles require accurate, explainable and generalizable stereo depth estimation.

As a reciprocal to explainability, we wish to explore uncertainty in the context of disparity estimation. Der Kiureghian et al. [11] identify aleatoric and epistemic uncertainties as the two main types which are further outlined by Kendall et al. [22] in the context of deep learning models. While aleatoric uncertainty relates to uncertainty in the observations, epistemic uncertainty captures situations not encountered in the training set. In the disparity estimation, this uncertainty accounts matching ambiguity, transparency, out-of-image matches and occlusions as well as out-of-domain image distributions resulting from generalization. Kendall et al. state that "this uncertainty is particularly important for safety-critical systems."

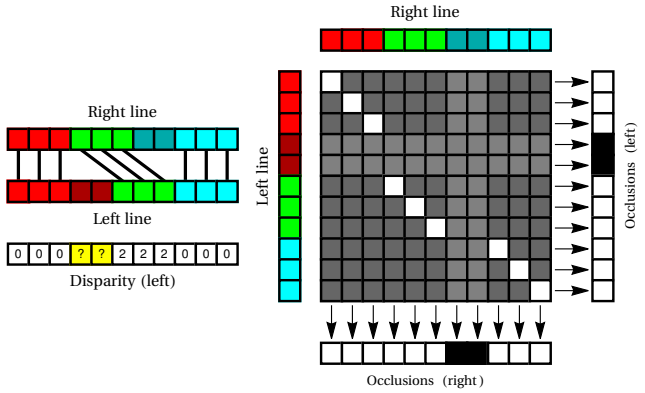

Figure 1. Permutation matrix. Stereo correspondence between left and right epipolar lines, with left disparity. Green object has disparity 2 and is closer than both red and cyan objects. Darker pixels are left-occluded (red) or right-occluded (cyan). The correspondence is represented as a permutation matrix, and occlusions are derived from the sums along rows or columns.

The contributions of this work are in the context of self-supervised approaches, where we propose a new natural representation of stereo constraints. By complementing the usual cost volume with a permutation approach for matching, a confidence measure can be modelled in a principled fashion as part of the stereoscopic process, not as a secondary step or as an explicit loss. As will be described, our proposed permutation volume and its normalization will simultaneously enforce occlusion modelling, disparity smoothness, and sharpness of depth discontinuities. This process disentangles stereo matching from monocular disparity estimation and extracts the stereo-visibility information present in a cost volume.

II. Previous Work

Estimating depth from color images is a long-standing problem which has been studied extensively

with deep neural network approaches[26]. Most traditional stereo matching algorithms boast two very interesting characteristics which do not appear naturally in deep stereo methods, namely that they are general and do not require ground truth depth maps.

Most algorithms establish a matching cost from one image to the other. This cost is very often based on the difference in pixel colors [5] or on image gradients [25]. These costs are aggregated within a certain neighborhood. Some methods use the sum of a fixed sized window at constant disparity [25] while others add a weight to each pixel within the window according to color similarity [45], [43]. Local algorithms select the disparity with the lowest matching cost [45], semi-global algorithms (SGM) solve for disparity by combining multiple 1D dynamic programming [19], global methods minimize a global energy function [5], [37], [25], [43]. These key steps work without ground truth and for most scenes.

A. Deep Disparity Estimation

Supervised deep methods tend to follow two general frameworks [26]. The first is an encoder-decoder structure such as FlowNet[12] and DispNet[32] which retrieve the disparity as a regression task and the second is to mimic traditional stereo with differentiable blocks. GC-Net [23] uses a 4D feature volume and a soft-argmin process for the disparity computation. GA-Net [46] proposes a semi-global aggregation layer and a local guided aggregation layer as ways to mimic the traditional stereo steps. The latest state-of-the-art work by Cheng et al. [9] uses a neural architecture search framework to model the general stereo matching problem as a neural network architecture. It is possible to improve the performance of models by simultaneously solving for analogous concepts such as occlusions or optical flow[48], [7], [31]. Many methods such OASMnet[27] and OASM-DDS [28] model occlusions as a separate neural network while others derive it from the disparity maps [35]. Another way to improve performances is to smooth the learning process with noise [30]. Combinations of these methods can be done such as in [38] where the trained LEAStereo[9] model is used in conjunction with data augmentation.

B. Self-Supervised Disparity Estimation

In the absence of ground truth training data, self-supervised (or unsupervised methods) are now proposed as alternatives. Self-supervised deep stereo matching approaches mainly rely on warping the left image from the image pair with the recovered disparity map such that the photometric difference between the warped left image and the right image is minimized [49]. These methods tend to introduce more geometric constraints in the image reconstruction losses such as taking advantage of the projective geometry and the spatial coherence [26]. This is further refined by adding a left-right image consistency over the recovered disparity map [16]. Some works generalize the mapping function through a generative model and recover the disparity map at an intermediate level [14]. Recently, the warping function has been shown to be defined as parallax attention maps [39], [40] or as an optimal transport problem of the latent space between images [20].

III. Permutation Stereo

The stereo matching problem, as depicted in Fig. 1, estimates a pairing between two epipolar lines (here "Left" and "Right"), and results in disparity values that express the displacement between corresponding pixels (Fig. 1, left). Notice that some pixels have no correspondence, as they are occluded. In Fig. 1, dark-red pixels are "left-occluded" and dark-cyan pixels are "right-occluded". This phenomenon occurs naturally in stereo because closer objects, with their larger disparities, are hiding farther objects. This pairing, or permutation, expresses stereoscopic disparities and occlusions simultaneously for both the left and right direction. This work proposes to express the traditional cost between epipolar lines in a permutation form.

A. Disparity from a Permutation Volume

The stereo correspondence between two epipolar lines can be expressed as a permutation matrix $\in \mathbb{R}^{w \times w}$, where w is image width, as illustrated on the right of Fig. 1. Assigning to each horizontal line i its own permutation P_i results in a permutation volume $P \in \mathbb{R}^{h \times w \times w}$, where h is the image height.

A permutation allows a pixel to match *any* pixel on the other line, thereby violating the stereo constraint stating that only a range of disparity is allowed, and this range is dictated by the camera geometry and the disparity d_{max} of the closest expected object. For classical stereo geometry (KITTI, SceneFlow, etc.), all valid disparities have the same sign, representing the fact that objects should appear to move left when the camera moves from left to right. It is easy to enforce the disparity range by assigning 0 probability to all invalid correspondences. Note that the permutation matrix cannot represent the full disparity range for the leftmost pixels of the left image, rows 1 to d_{max}, since some potential matches are outside the right image.

Using permutations, stereo matching can be considered as an optimal transport problem. In practice, this is accomplished by, firstly, estimating a traditional cost volume C and transforming it into an unnormalized permutation volume $\bar{C}$

$$C_{y,x,d} \rightarrow \bar{C}_{y,x,x-d} \tag{1}$$

and by, secondly, applying symmetric normalization which yields P, the permutation volume.

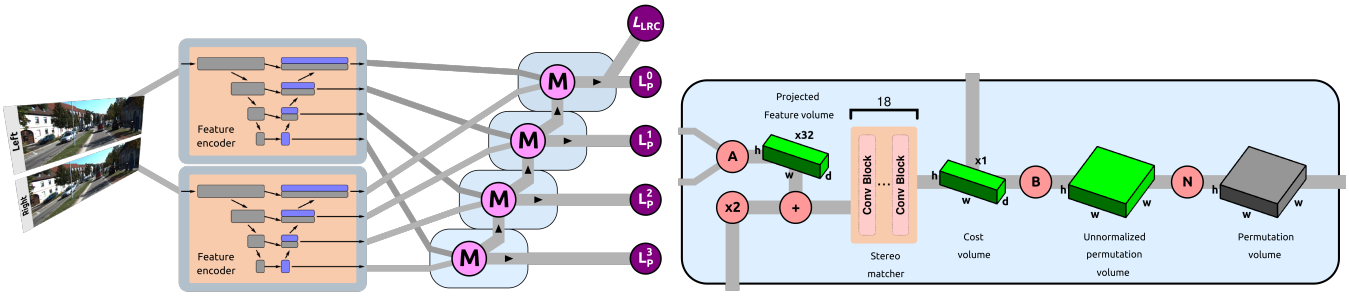

Figure 2. (LEFT) The training architecture for permutation-based stereo matching. M) The stereo matching network. The losses are L_P at various scales and L_{LRC}. (RIGHT) The stereo matching network in detail. A) The feature volume and projection operation, x2) upscaling of lower resolution cost volume, +) concatenation, B) cost volume reshaping and padding, N) symmetric normalization

Doubly Stochastic Normalization. Traditional cost volumes are usually normalized using a soft-argmax or soft-argmin and then projected to the disparity map by a weighted product over the disparity range. Quite similarly, the permutation volume allows this with a normalization which enforces that each permutation P_i is doubly stochastic and orthogonal.

As stated by Emami et al. [13] the problem of learning permutations is quite challenging since the number of permutations grows factorially with the size of the problem. Also, the permutation matrices prevent learning algorithms from directly using backpropagation as they are not differentiable. The Sinkhorn-Knopp algorithm is known to map a square matrix to a doubly-stochastic matrix where all rows and columns sum to one via iterative re-scaling. Adams et al. [3] proposed the iterative projection operator known as Sinkhorn normalization for learning doubly stochastic matrices based ranking functions. They propose an algorithm called *Sinkhorn propagation* which allows for gradients to be computed on the Sinkhorn normalization operator via backpropagation [3]. We propose to adapt this matrix normalization such that it becomes symmetric, by simultaneously normalizing columns and rows. We refer to this iterative process as symmetric normalization, Fig. 2(N), which yields the permutation volume P. Iterative applications of this normalization ensures P gradually becomes doubly stochastic.

$$
\begin{aligned}
P_{i,j,k}^{(t=0)} &= exp(\bar{C}_{i,j,k}) \\
P_{i,j,k}^{t} &= \frac{P_{i,j,k}^{t-1}}{\sqrt{\sum_m P_{i,m,j}^{t-1} \sum_m P_{i,j,m}^{t-1}}}
\end{aligned}
\tag{2}
$$

where t indicates the iteration step.

Orthogonality and Confidence. Orthogonality is an essential property of permutations, as it guarantees that a correspondence is bijective, which is equivalent in the context of stereo to the left-right consistency constraint [16], i.e. if non occluded pixel a matches pixel b, then pixel b matches pixel a. Therefore, we can enforce and measure if correspondences satisfy a left-right consistency constraint with orthogonality. The orthogonality of permutation P_i is defined as $P_i \cdot P_i^\top = \mathbb{1}$ where $\mathbb{1} \in \mathbb{R}^{w \times w}$ is the identity matrix. Section IV-B will show how orthogonality can be defined as a loss while it is naturally enforced through the normalization.

An interesting property of orthogonality is that each diagonal element of $P_i \cdot P_i^\top$ is the sum of a squared row in P_i. These values are 1 when matches are bijective, but closer to 0 when matches are occlusions or ambiguous. We can therefore measure the confidence of the stereo matching process as the sum of squared rows for the left confidence and as the sum of squared columns for the right confidence, as illustrated in Fig. 1.

This confidence map can be used as a measure that a given pixel will be correctly matched to its target in the matching image. As an example, a pixel in a textureless region usually presents a lot of matching ambiguity, indicated in the permutation by numerous low values. It will not satisfy left-right consistency and will have a low squared norm of its row. Similarly, the presence of low contrast or high noise will also result in a low confidence. This means that in the context of a permutation matrix, confidence encodes matching ambiguity resulting from occlusions as well as texture ambiguity. This is further detailed in Section. V. The confidence map expresses uncertainties induced by the stereo problem by providing information about the matching process, not the uncertainties of the algorithm itself.

B. Smoothness and Sharpness

Many stereo matching algorithms struggle with the representation of smoothness and sharp edges. It seems that one concept can only be favoured at the expense of the other. While smoothness wants adjacent pixels to bear the same disparity, sharp edges require disparity to abruptly change across depth discontinuities. A permutation matrix seems to naturally encourage both properties simultaneously.

Symmetric normalization strongly constrains the solution. The normalization for a pixel affects a whole row of the

image rather than individual pixels. This implies that a strong match will propagate itself horizontally across the disparity range to discourage incoherent or noisy matches. Disparity map smoothing is also encouraged by promoting orthogonality ($P_i \cdot P_i^\top = \mathbb{1}$). Any depth discontinuity, even small, induces an occlusion, so the constraint is best satisfied (i.e. absence of occlusion) when all pixels share a common disparity. When a discontinuity has to occur, there is no incentive to spread the change across multiple pixels, so the break will be clean, and objects will remain smooth.

IV. Self-Supervision on the Permutation Volume

The main contribution of this work is the introduction of a new formulation of the cost volume, the permutation volume, and its associated normalization. In a single volume, it naturally expresses both left and right disparity as well as both left and right confidence maps.

Because of the simple relationship between the permutation volume and the cost volume as defined by eq. 1-2, P and its losses can be implemented without allocating $h \times w \times w$ space in memory by accessing C as if it were P by using modified indices. Furthermore, false matches outside the disparity range are prohibited by setting probability to 0.

A. Training Architecture

Given an input stereo pair $\{I^L,\ I^R\} \in \mathbb{R}^{3\times h\times w}$, the images are fed into a siamese feature encoder network with a U-net architecture as displayed in Fig 2. This encourages strong feature representation at multiple resolutions.

This feature encoder yields latent features at various scales $F^{Ls},\ F^{Rs} \in \mathbb{R}^{(f*s)\times(h/s)\times(w/s)}$ where f is the number of features and $s \in \{2^0, 2^1, 2^2, 2^3\}$ is the scale. These latent features are then combined into a 4D feature volume $V^s \in \mathbb{R}^{(2f*s)\times(h/s)\times(w/s)\times(d_{max}+1)/s}$ by concatenating and where d_{max} is the maximum number of integer disparities with an initial value of 0. This step is shown as Fig. 2(A). The feature volume is defined as

$$V^s_{i,j,k,d} = Concat(F^{Ls}_{i,j,k}, F^{Rs}_{i,j,k+d}) \tag{3}$$

The feature volume is then projected down to f features with a 3D convolution block. This allows for the concatenation of the prior cost volume obtained at lower resolution. Fig. 2 displays this multiscale process where the stereo matching serves as a module that refines the disparity estimate with richer feature pairs at the corresponding scale of the same resolution.

The siamese stereo matching network is a Resnet with eighteen 3D convolution blocks. It yields a cost volume $C^s \in \mathbb{R}^{(h/s)\times(w/s)\times(d_{max}+1)/s}$. This cost volume is the traditional cost volume as used in many volumetric stereo approaches. It is reshaped and padded, Fig. 2(B), into an unnormalized permutation volume $\bar{C}^s \in \mathbb{R}^{(h/s)\times(w/s)\times(w/s)}$ according to eq. 1. Symmetric normalization, Fig. 2(N), is applied and results in the permutation volume $P^s \in \mathbb{R}^{(h/s)\times(w/s)\times(w/s)}$.

B. Losses

The total loss for the training is defined as:

$$\mathcal{L} = \frac{\sum \mathcal{L}^s_{\text{P}}}{s^2} + \lambda \mathcal{L}_{\text{LRC}} \tag{4}$$

where $\mathcal{L}_{\text{P}}$ and $\lambda\mathcal{L}_{\text{LRC}}$ are respectively the photometric loss and the left-right consistency loss. The s^2 imposes that pixels contribute equally regardless of scale.

Photometric Loss. This loss refers to the similarity between a reconstructed image and the input image from the stereo pair while respecting the occlusion information. This function is a combination of the structural similarity index (SSIM)[41] and the L_1-norm as defined in [7]. This loss is furthermore adapted to the context of permutations which allows to account for occluded and ambiguous pixels. The occlusion handling scheme is introduced in [8] but our novel model allows for continuous confidence maps and does not require to duplicate the stereo matching process.

$$\begin{aligned}
\mathcal{L}_{\text{P}} &= \frac{\sum \bar{\mathcal{L}}_{\text{P}} \odot O}{\sum O} \\
\bar{\mathcal{L}}_{\text{P}} &= \frac{\alpha}{2}(1 - \text{SSIM}(I^R_{i,j}, I^{R*}_{i,j}) + (1-\alpha)\|I^R_{i,j} - I^{R*}_{i,j}\|_1 \\
I^{R*}_i &= P_i \cdot I^L_i
\end{aligned} \tag{5}$$

where the $\odot$ operator indicates element-wise multiplication and the $*$ operator indicates a permuted image. For ease of reading, only the right-to-left photometric loss has been detailed but both left-to-right and right-to-left are computed and averaged. Moreover, we only define this loss for a single scale but it is applied at every scale. The confidence map computation is detailed in eq. 7.

Left-Right Consistency Loss. The left-right image consistency is a staple loss in self-supervised stereo matching [16]. With the permutation volume, it is possible to model this round trip as $P_i \cdot P_i^\top$, computed from a single stereo matching. In this way, the permutation volume can be regularized.

$$\mathcal{L}_{\text{LRC}} = \|P_i \cdot P_i^\top - \mathbb{1}\|_1 \tag{6}$$

C. Disparity and Confidence Maps

From the permutation volume P, we use the weighted product over the disparity range to solve for the left disparity $D_{i,j}$ and compute the sum of squared rows to recover the left confidence map $O_{i,j}$.

$$\begin{aligned}
D^L_{i,j} &= \sum_d P_{i,j,j-d} \times d \\
O^L_{i,j} &= \|P_{i,j,:}\|_2^2
\end{aligned} \tag{7}$$

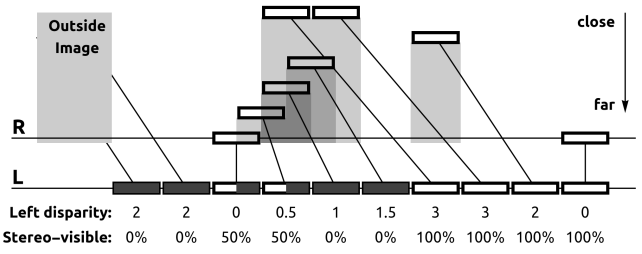

Figure 3. Stereo-visibility. Left image pixels, shifted by their disparity, hide farther pixels, fully or partially. Out-of-boundary pixels are also considered stereo-occluded.

This relationship also allows to trivially recover the right image disparity and the right occlusion map.

Confidence Heuristic: To manage pixels which are identified as part of an occlusion or have texture ambiguity, the confidence map is necessary. Although it is impossible to estimate the exact depth of an occluded pixel via stereo matching, we propose a simple heuristic at low confidence pixels. In the left image, pixels which have a confidence value under the threshold τ are attributed a disparity which comes from the closest high confidence disparity on their left. This heuristic is inspired by the observation that an occluded pixel tends to have the same disparity as its neighbors on the left if it is occluded in the left image. There are known monocular cues that could be used to fill occlusions such as texture similarity, but this work chooses the simplest heuristic. As a special case, left image pixels with out-of-image matches are instead propagated from the right.

Our heuristic is similar to GOAPP in [35] but instead uses the confidence map which is not derived from the disparity map but rather from the permutation volume. Also, some work has previously been done to express entropy of the stereo probabilities in [47] as a "matchability" score. Our confidence map and our heuristic leverage new information about the matching process rather than simply correcting incompatibilities in the disparity map.

In this paper, the usefulness of the confidence maps is validated by using the simplest heuristic possible. It is inspired by the simplest post-processing heuristics employed in classical methods. We consider that using the most naive occlusion handling approach makes it easier to highlight the usefulness of the confidence maps recovered from the permutation volume.

V. Stereo-visibility

We define a pixel as stereo-visible if it belongs to a correspondence in the stereo image pair. It is well known that objects that are closer hide objects that are further, thereby creating an occlusion. An occluded pixel has no correspondence in the other image, so it is not stereo-visible, it is stereo-occluded. Furthermore, in the context of stereo matching, we consider pixels that have their match outside the other image boundaries as stereo-occluded. It is possible to recover the map of stereo-visible pixels from the groundtruth disparity map and it will be referred as a stereo-visibility map. The KITTI dataset explicitly exposes out-of-boundary correspondences in the "noc" disparity maps, but does not provide a true occlusion map. Other datasets, such as SceneFlow, provide true occlusions but not out-of-boundary matches. Computing a stereo-visibility map from a disparity map is straightforward, and is depicted in Fig. 3.

We consider that pixels that are not stereo-visible can only be solved by monocular depth estimation. Since common datasets, such as KITTI or SceneFlow, do not limit error computations to stereo-visible pixels, they not only measure stereo performance, but also implicitly include monocular disparity estimation performance. We thus believe that stereo-visible regions can lead to a good performance estimator of the stereo algorithm itself, while the contribution of monocular depth estimation capabilities can be best evaluated in stereo-occluded regions.

The confidence map aims to provide explainability for the matching process. The permutation volume naturally provides this information as a result of symmetric normalization. Since high confidence pixels should be stereo-visible, and low confidence should be either stereo-occluded or feature an ambiguous texture, the stereo-visibility map computed from ground truth can help assess the usefulness of the confidence measure, as shown in Tab. II.

VI. Experiments and Discussion

The current work is interested in stereo matching capabilities for real-world images such as the KITTI 2012 [15] and KITTI 2015 [33], [34] as well as learning and generalization capabilities using very different stereo settings such as SINTEL-final[6] and FlyingThings3D[32]. The KITTI 2012 dataset has 194 training image pairs with 195 test image pairs with a pixel size of 1226×370. The KITTI 2015 dataset has 200 training image pairs with 200 test image pairs with a pixel size of 1242×375. The SINTEL-final dataset has 1064 training image pairs with a pixel size of 1024×436 with strong motion blur. It will be referred as SINTEL for brevity. SceneFlow FlyingThings3D contains 21818 training image pairs with a pixel size of 960×540. Section VI-B reports standard errors for KITTI 2012 testing images and for KITTI 2015 testing images. Sections VI-C and VI-D report the D1-All on the KITTI 2015 and SINTEL training datasets. Section VI-E presents the D1-All for the training data of KITTI 2012 and SINTEL.

A. Implementation Detail

Our model is trained on the datasets at half resolution on random image crops of size of 192×32 pixels with a batch size of 2. No other data augmentation is applied

Method	KITTI 2012		KITTI 2015			
	Out-Noc (%)	Out-All (%)	D1-bg (%)	D1-fg (%)	D1-Noc (%)	D1-All (%)
Hirschmuller et al. [19]	7.64	9.13	8.92	20.59	9.47	10.86
Hamzah et al. [17]	-	-	8.64	21.85	10.28	10.84
Ahmadi et al.[4]	-	-	-	-	11.17	16.55
Zhou et al. [48]	-	-	-	-	8.61	9.91
SegStereo [42]	-	-	-	-	7.70	8.79
OASM-Net [27]	6.39	8.60	6.89	19.42	7.39	8.98
PASMnet_192 [39]	-	-	5.41	16.36	6.69	7.23
Flow2Stereo [31]	4.58	5.11	5.01	14.62	6.29	6.61
Permutation Stereo	7.39	8.48	5.53	15.47	6.72	7.18

Table I
QUANTITATIVE RESULTS. COMPARISON OF PERMUTATION STEREO MATCHING WITH POPULAR UNSUPERVISED AND TRADITIONAL METHODS ON THE KITTI 2012 AND KITTI 2015 TESTING DATASETS.

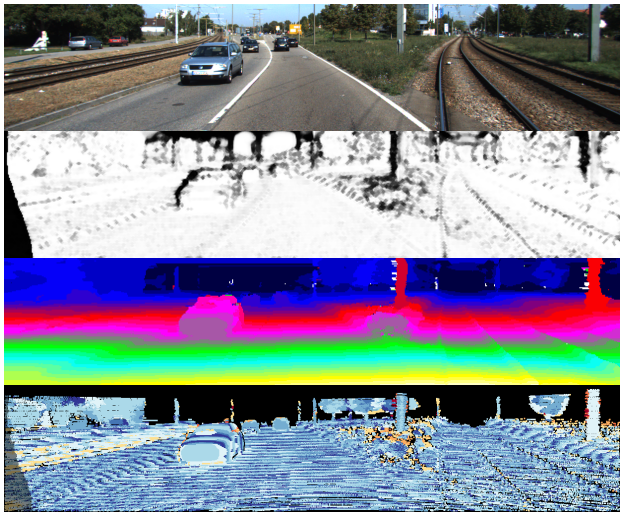

Figure 4. KITTI 2015 testing results. From top to bottom: Image 17, Confidence map, Disparity map, Error image.

Method	C	K2012 D1-All (%)		SINTEL D1-All (%)	
		Stereo-occluded	Stereo-visible	Stereo-occluded	Stereo-visible
PASMnet_192		-	-	61.7	21.7
P-Stereo		44.07	7.34	64.90	20.09
P-Stereo	✓	28.56	7.05	48.30	19.02

Table II
STEREO MATCHING ERROR FOR STEREO-OCCLUDED AND STEREO-VISIBLE PIXELS FOR KITTI 2012 AND SINTEL TRAINING DATASETS. P-STEREO IS PERMUTATION STEREO. C REPRESENTS THE CONFIDENCE HEURISTIC APPLIED TO THE DISPARITIES.

apart from random crops. The conv blocks are as defined in [18] and their Fig. 5 (right) with $f = 32$. The convolution layers apply fixed padding, have batch normalization and have a Relu activation function. The implementation is made with Mathematica[21] 12.3. The λ is set to 10 and the symmetric normalization has $t = 8$ iterations. Networks are trained until convergence with the Adam Optimizer[24] and a learning rate of 1×10^{-3}. The constant α is set to 0.85 as is customary [7] and τ is set to 0.1. The models are trained on an RTX3090. We like to highlight that comparisons will use the state-of-the-art PASMnet [39] as they have made their code available.

B. Evaluation

Quantitative results on the testing datasets KITTI 2012 and KITTI 2015 are shown in Tab. I. These results are available online [1], [2]. Our neural network was pretrained on SceneFlow and finetuned on KITTI 2012 and KITTI 2015 respectively. Tab. I indicates that our method does not achieve state-of-the-art on the datasets for self-supervised stereo matching methods falling behind method Flow2Stereo which optimizes multiple training objectives. With a straightforward volumetric stereo backbone, our method outperforms Semi-global matching from Hirschmuller et al. [19] which is a gold standard in traditional stereo in both KITTI 2012 and KITTI 2015 and outperforms, albeit slightly, a method such as PASMnet [39]. The permutation results unequivocally enforce stereo matching and yields very high quality disparity maps. This level of performance is further detailed in Fig 4 where one can see smooth and sharp disparities without any smoothness or sharpness losses. The qualitative results show streaking which is due to our confidence heuristic. This very simple heuristic allows our model to perform competitively, therefore the results can interpreted as consequence of the permutation model not a strong monocular completion heuristic.

C. Cost and Permutation Comparison

The introduction of the permutation volume formulation to volumetric stereo methods is studied and comparisons to cost volumes are presented in Tab. III. We argue that a better representation of true stereo constraints in the network architecture will lead to better performances in the D1-Noc and D1-All error metrics for the various losses. Results are provided for KITTI 2015 and SINTEL training datasets while all networks are trained on FlyingThings3D. The difference between cost volume and permutation volume is better displayed in Fig. 5. Results demonstrate that a permutation volume and with symmetric normalization

Volume for normalization	Normalization function	Losses $\mathcal{L}_P$	Losses $\mathcal{L}_{LRC}$	Conf	KITTI 2015 D1-Noc (%)	KITTI 2015 D1-All (%)	SINTEL D1-Noc (%)	SINTEL D1-All (%)	$\overline{\Delta}$
Cost	Softmax	✓			33.41	34.55	35.06	36.17	0.0
Permutation	Symmetric	✓			21.31	22.65	24.67	25.98	-11.1
Permutation	Symmetric	✓		✓	36.15	37.03	25.23	26.37	-3.6
Cost	Softmax	✓	✓		27.26	28.49	28.93	30.13	-6.1
Permutation	Symmetric	✓	✓		15.77	17.21	18.36	19.78	-17.0
Permutation	Symmetric	✓	✓	✓	16.61	17.20	17.67	18.99	-17.2

Table III
COMPARATIVE EVALUATION OF STEREO MATCHING ERROR FOR COST VOLUME AND PERMUTATION VOLUME AND THEIR RESPECTIVE NORMALIZATION. CONF REPRESENTS THE CONFIDENCE HEURISTIC APPLIED TO THE DISPARITIES. THE ERRORS ARE PRESENTED FOR THE KITTI 2015 AND SINTEL DATASETS.

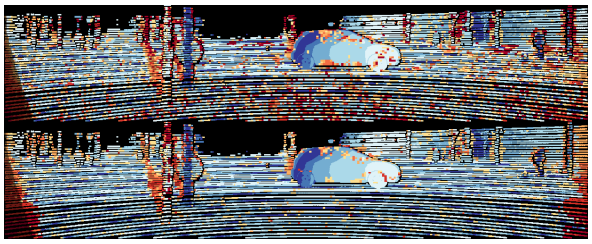

Figure 5. Stereo matching error for image 46 of the KITTI 2015 training dataset. Top) Error image (31.1%) for a cost volume and softmax normalization. Bottom) Error image (17.9%) for a permutation volume and symmetric normalization.

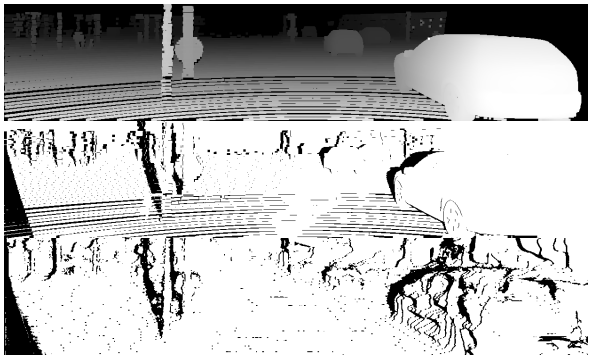

Figure 6. Stereo-visibility. Top to bottom: Groundtruth disparity, Stereo-visibility from groundtruth, Confidence map.

improve performances regardless of losses. Fig. 5 shows the expected behaviour of the normalization which is a row-wise consistency between matches where strong matches propagate and naturally correct incompatibilities. Tab. III shows that accurate confidence maps are tied to orthogonality. In the absence of left-right consistency, the confidence map yields very poor information and the confidence heuristic may increase errors significantly.

D. Confidence Evaluation

Fig. 6 displays the binarized stereo visibility maps that can be computed from the ground truth disparities (top) and the binarized confidence map recovered from our trained neural network (bottom). Using the computed stereo-visibility maps (middle), it is possible to separate pixels into stereo-visible regions and stereo-occluded regions during evaluation. Tab. II presents D1-ALL errors in the stereo-occluded and stereo-visible regions. Our Permutation Stereo method is trained on SceneFlow and finetuned on KITTI 2015 which is also the case for PASMnet_192. Fig. 7 shows how the distributions of errors are affected by the confidence heuristic. These results demonstrate that the recovered confidence maps and the heuristic lower the error primarily in the stereo-occluded regions. This indicates that the confidence maps correctly model for occlusion and out-of-image pixels and our simple heuristic has very limited effect on stereo-visible pixels. Self-supervised stereo is naturally unable to estimate disparities in occluded regions and this is specifically what the permutation formulation addresses. The photometric loss does not guide the network to yield accurate disparity estimate in these regions during training. However, the permutation formulation model accurately recovers where this uncertainty lies. Tab. II shows that PASMnet has poorer performance in stereo-occluded regions than our simple heuristic.

Model	Conf	KITTI 2012 D1-All (%)	SINTEL D1-All (%)
Permutation Stereo	✓	8.70	21.09
PASMnet		6.32	20.32
PASMnet	✓	6.15	19.62
PASMnet_192		5.74	23.17
PASMnet_192	✓	5.57	21.95

Table IV
GENERALIZATION RESULTS. EACH ROW REPRESENTS A DIFFERENT TRAINING SET. **CONF** REPRESENTS THE CONFIDENCE HEURISTIC APPLIED TO THE DISPARITIES. THE ERRORS ARE PRESENTED FOR THE KITTI 2012 AND SINTEL TRAINING DATASETS.

E. Generalization

The permutation volume allows for new constraints on the stereo matching problem and models for uncertainty through the confidence map. This should lead to better generalization performances. Tab. IV shows the performance in KITTI 2012 for models trained on SceneFlow and finetuned on KITTI 2015. Our model has similar performance to

Volume for normalization	Normalization function	Conf	KITTI 2015 D1-All (%)			Middleburry 2014 D1-All (%)		
			Stereo-occluded	Stereo-visible	All	Stereo-occluded	Stereo-visible	All
Cost	Softmax		54.2	15.0	17.3	77.4	28.6	35.0
Permutation	Symmetric		56.5	15.6	17.7	81.3	27.3	34.2
Permutation	Symmetric	✓	30.1	14.8	15.6	47.1	29.8	32.1

Table V
GENERALIZATION RESULTS. MODEL TRAINED ON SCENEFLOW FLYINGTHINGS3D AND TESTED ON KITTI 2015 AND MIDDLEBURRY 2014 TRAINING DATA. THE PERMUTATION FORMULATION AND ITS CONFIDENCE MAP HELP IN A GENERALIZATION SETTING.

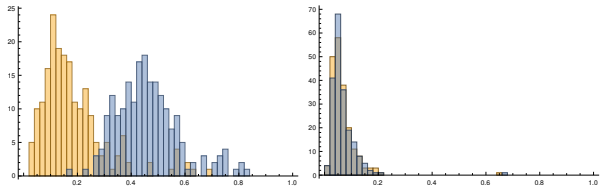

Figure 7. Performance of confidence heuristic, KITTI 2015, Distribution of D1-All (%). Left, impact on stereo-occluded pixels. Right, impact on stereo-visible pixels. Blue, before the confidence heuristic. Orange, after applying the confidence heuristic. We observe a reduction of error for stereo-occluded pixels and no impact on stereo-visible pixels as expected.

PASMnet_192 in Tab. I yet outperforms its results in an out-of-domain dataset such as SINTEL. Although it is unorthodox, we apply the confidence heuristic to the recovered disparity maps obtained from the PASMnet networks. This leads to consistent improvements in the results. For the KITTI 2012 results, PASMnet outperforms our model yet the confidence heuristic still improves its results. This demonstrates that the recovered confidence map is a good measure of uncertainty in stereo matching.

Tab. V presents the generalization improvement that results from the introduction of the permutation model. All models are trained on SceneFlow and are used to generalize to KITTI 2015 and Middleburry 2014 [36]. The introduction of the permutation volume by itself does not automatically improve results. Rather, the use of the uncertainty information contained in the confidence map allows for gains with even the simplest heuristic.

VII. LIMITATIONS

The permutation volume adds strong constraints to the stereo matching process. However, it does not directly help with monocular disparity estimation. The current paper did not explore how to best manage the stereo-occluded regions nor did it explore the use of a dedicated monocular disparity network for these regions. These monocular networks could be learned in an unsupervised manner [10], [29], [8], [44]. This paper presents the permutation model and its confidence map as a tool to identify matching uncertainty and allow for comparable results on popular datasets. Using the most naive occlusion handling approach makes it easier to highlight the usefulness of the confidence maps recovered from the permutation volume.

This paper explores the use of a permutation volume only for the self-supervised learning setting. As it has been stated, a goal was to establish a confidence measure during the stereo matching process. The self-supervised setting where models naturally cannot solve for occlusion and out-of-image pixels is the best setting for the evaluation of such a model. It would be interesting to study the effects of the permutation model in a supervised learning context, and its impact on explainability.

VIII. CONCLUSION

This paper proposed a novel permutation formulation to the stereo matching problem. The permutation volume allows for a more natural representation of constraints that were previously managed mostly through losses, such as smoothness, sharpness, and occlusions. Relying on symmetric normalization, it also provides a confidence measure which is closely related to the concept of stereo-visibility. Validation is accomplished with stereo performances measured on standard testing datasets, while the usefulness of confidence map is assessed with stereo-visibility. Backed by experimental results, the proposed confidence heuristic adequately resolved stereo-occluded pixels, which are considered monocular. We consider that the permutation volume formulation is a good addition to stereo matching algorithms that not only helps satisfy stereo constraints but also provides explainability, which is becoming important in practical applications of stereo algorithms.

REFERENCES

[1] http://www.cvlibs.net/datasets/kitti/eval_stereo_flow_detail.php?benchmark=stereo&error=3&eval=all&result=c3e639d5ab83f9018bd6e92ac553b33ea7edcdb0.

[2] http://www.cvlibs.net/datasets/kitti/eval_scene_flow_detail.php?benchmark=stereo&result=78dc427d034c849ebb9794ebb3fa8d5b204b8238.

[3] Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

[4] Aria Ahmadi and Ioannis Patras. Unsupervised convolutional neural networks for motion estimation. In *2016 IEEE international conference on image processing (ICIP)*, pages 1629–1633. IEEE, 2016.

[5] Stan Birchfield and Carlo Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

[6] 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, Oct. 2012.

[7] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7063–7072, 2019.

[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, 2020.

[10] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset. *Expert Systems with Applications*, 178:114877, 2021.

[11] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[12] 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.

[13] Patrick Emami and Sanjay Ranka. Learning permutations with sinkhorn policy gradient. *arXiv preprint arXiv:1805.07010*, 2018.

[14] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016.

[15] 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.

[16] 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.

[17] Rostam Affendi Hamzah, Haidi Ibrahim, and Anwar Hasni Abu Hassan. Stereo matching algorithm based on per pixel difference adjustment, iterative guided filter and graph segmentation. *Journal of Visual Communication and Image Representation*, 42:145–160, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[20] Baoru Huang, Jian-Qing Zheng, Stamatia Giannarou, and Daniel S Elson. H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry. *arXiv preprint arXiv:2104.11288*, 2021.

[21] Wolfram Research, Inc. Mathematica, Version 12.2.3. Champaign, IL, 2021.

[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[23] 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.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 15–18. IEEE, 2006.

[26] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *arXiv preprint arXiv:2006.02535*, 2020.

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

[28] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Shenghao Zhang, and Chong Zhang. Unsupervised occlusion-aware stereo matching with directed disparity smoothing. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[29] Huan Liu, Junsong Yuan, Chen Wang, and Jun Chen. Pseudo supervised monocular depth estimation with teacher-student network. *arXiv preprint arXiv:2110.11545*, 2021.

[30] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2020.

[31] 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.

[32] 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.

[33] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[34] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018.

[35] Liang Peng, Dan Deng, and Deng Cai. Geometry-based occlusion-aware unsupervised stereo matching for autonomous driving. *arXiv preprint arXiv:2010.10700*, 2020.

[36] 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.

[37] Geert Van Meerbergen, Maarten Vergauwen, Marc Pollefeys, and Luc Van Gool. A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision*, 47(1):275–285, 2002.

[38] Hengli Wang, Rui Fan, and Ming Liu. Co-teaching: An ark to unsupervised stereo matching. *arXiv preprint arXiv:2107.08186*, 2021.

[39] 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.

[40] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021.

[41] 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.

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

[43] Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewénius, and David Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):492–504, 2008.

[44] Xinchen Ye, Xin Fan, Mingliang Zhang, Rui Xu, and Wei Zhong. Unsupervised monocular depth estimation via recursive stereo distillation. *IEEE Transactions on Image Processing*, 30:4492–4504, 2021.

[45] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):650–656, 2006.

[46] 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.

[47] Jingyang Zhang, Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Learning stereo matchability in disparity regression networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1611–1618. IEEE, 2021.

[48] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1567–1575, 2017.

[49] Kun Zhou, Xiangxi Meng, and Bo Cheng. Review of stereo matching algorithms based on deep learning. *Computational Intelligence and Neuroscience*, 2020, 2020.