Large-scale Unbiased Neuroimage Indexing via 3D GPU-SIFT Filtering and Keypoint Masking

Étienne Pepin[1], Jean-Baptiste Carluer[2], Laurent Chauvin[1], Matthew Toews[1][0000−0002−7567−4283], and Rola Harmouche[3]

[1] École de Techologie Supérieure
{etienne.pepin.1@ens.etsmtl.ca,
laurent.chauvin0@gmail.com,matt.toews@gmail.com}
[2] Université de Nantes, Nantes, France
{carluer.jean-baptiste@orange.fr}
[3] Canadian National Research Council
{rolaharmouche@gmail.com}

Abstract. We propose a feature extraction method via a novel description and a scalable GPU implementation (the first to our knowledge) of the 3D scale-invariant feature transform (SIFT). The feature extraction is first represented as a shallow convolutional neural network with pre-computed filters, followed by a masked keypoint analysis. We use the implementation in order to investigate feature extraction for specific instance identification on natural non-skull-stripped magnetic resonance image (MRI) neuroimaging data. The proposed implementation is invariant to 3D similarity transforms and aims to improve robustness by reducing noise and bias for image processing convolution operations. We show interpretable feature visualizations, which help explain the obtained results. We demonstrate state-of-the-art results in large-scale neuroimage family indexing experiments on 3D data from the Human Connectome Project repository, and show significant speed gains compared to a CPU implementation. The results imply that using feature extraction using SIFT for neuroimaging analysis can lead to less noisy results without the need for hard masking during preprocessing. The proposed algorithm can be applied on arbitrary image modalities and anatomical structures.

1 Introduction

Convolutional neural networks are considered state-of-the-art for medical image classification problems. Recent advances in computational power and parallel processing with GPUs have lead to quick and accurate classifiers, particularly in the presence of a large amount of training data and a small number of classes [9]. However, this does not hold when identifying image pairs in a large medical imaging cohort, where there is a large number of subjects and a limited number of samples per subject. In addition, CNNs introduce bias in their filters towards local object textures in contrast to global object shapes [18], whereas it has been shown that networks that learn shape-based representations can improve

robustness, detection performance, and generalization [7]. Third, there is no standard method for the interpretability of CNNs, resulting in a lack of trust in these systems from end users [8]. Finally, while segmentation masks are commonly used to extract structures of interest prior to processing, e.g. brain masking or 'skull-stripping' [4, 5, 14, 20, 21], hard masking produces an irregular and abrupt boundary in the image similar to zero-padding, which introduces artifacts such as ringing at mask borders [15]. An early study emphasized the need for reducing noise and bias for convolution operations in the context of image processing, and demonstrated the benefit of GPU implementations in that context [12]. Various GPU implementations of 2D SIFT have been proposed [2, 19]. However, extending these to 3D medical image volumes is non-trivial and has not been done to date, particularly 3D keypoint description, and no work has described invariant keypoint descriptor correspondence as a massive scale convolution operation.

This paper addresses the limitations mentioned above by proposing an efficient GPU Gaussian scale-space feature extraction method for medical image analysis. A novel representation of 3D SIFT as a shallow convolutional neural network with pre-computed filters (SIFT-CNN) is shown for extracting visualizable and interpretable keypoints with powerful shallow information in image data. We show that the backbone of the SIFT algorithm is a single channel Gaussian CNN, i.e. the Gaussian scale-space generated via recursive Gaussian, and can be generated efficiently via separable filters. These features are designed to be rotation invariant [11] and thus are robust against rotation bias. We use these features in a keypoint masking process on medical image data. Keypoints are first extracted on an entire image, and those outside the masks of the structures of interest are then discarded for future analysis. This helps avoid linear filtering artifacts due to sharp boundaries that would normally affect any linear filtering system including CNN or SIFT.

SIFT features [13] have shown to be efficient at image matching applications. Toews and Wells [22] developed a 3D-SIFT-Rank keypoint method and demonstrated its usefulness in identifying MRI pairs of siblings using the Jaccard measure of overlap on skull-stripped images. The resulting signatures require little memory usage and can thus be used on a large dataset. A recent study [3] used the 3D-SIFT-Rank keypoints in order to extract signatures of individuals and subsequently identify MRIs of siblings. This method was the first to detect subject duplication errors in the ADNI and OASIS cohorts. Keypoint extraction requires on the order of seconds per image, the keypoint data are approximately 100x smaller than the original image, and highly efficient nearest neighbor keypoint indexing may be computed in $O(logN)$ complexity in the number of keypoints N via approximate nearest neighbor search [16].

Similarly to the work in [3], we show the effectiveness of our proposed keypoint masking algorithm in identifying similarities between image pairs, particularly due to family relationships, and compare it to previously obtained results on skull-stripped brain volumes. In order to allow for further speedups, the feature extraction is performed using a 3D graphics processing unit (GPU) SIFT imple-

mentation which results in an approximate 7x speedup compared to the CPU implementation. The contributions of our work can be summarized as follows:

- A robust method for Gaussian feature extraction in scale-space followed by keypoint masking via shallow CNN with precomputed filters
- A scalable GPU implementation of 3D SIFT which offers an approximate 7x speedup compared to the CPU implementation, which to our knowledge has not been done to date.
- Application of interpretable keypoint masking in order to classify all brain MRIs of the same family from a large dataset using with state of the art results for brain indexing.
- The first application keypoint transfer segmentation on brain MRI data.

2 Method

Methods here are two-fold. First, we present a solution to convolution filter bias and noise via an adaptation of Gaussian scale-space theory [11] to the GPU architecture widely used in deep CNN processing. Second, we propose an ROI analysis strategy, where neuroimage features are extracted from natural images without the limits of hard boundaries, followed by selection of feature points that lie within a ROI for further analysis.

2.1 Gaussian Scale-Space Filtering on the GPU

Let $x \in R^3$ represent a 3D coordinate system, and let $I_x : R^3 \rightarrow R^1$ represent a scalar image. Scale-space theory seeks to model the image in a manner independent of the image resolution, with defined continuous Gaussian convolutions: $I_{\sigma,x} : R^4 \rightarrow R^1 = I_x * G_\sigma$, where G_σ is a Gaussian filter defined by pixel scale σ. The Gaussian filter is shown to be the only filter satisfying a number of axioms, including non-creation and non-enhancement of spurious local maxima and providing both an unbiased visual front end due rotational symmetry, and a means of computing scale-normalized derivative operators. In contrast, filters resulting from typical CNN training via stochastic backpropagation [10] are not invariant to image scaling or rotation and highly biased towards training data.

We propose integrating the Gaussian scale-space (GSS) directly into deep CNN filtering on the GPU, thereby limiting bias or image artifacts. Figure 1 shows our GPU implementation of the widely-used scale-invariant feature transform (SIFT) algorithm [13] based on 3D SIFT-Rank [22]. A detailed description is beyond the scope of this paper and is provided along with full source code [4]. Several notable details are as follows. The GSS is shown in Figure 1 a), where scale is sampled in constant multiplicative increments $k : \sigma_{i+1} = k\sigma_i$ in order to remain invariant to scale change. Each sample $I_{\sigma,x}$ may be generated via a single convolution of the input I_x (and thus remains a 'shallow' filtering operation), a more computationally efficient strategy is recursive filtering $I_{\sigma_{i+1},x} = I_{\sigma_i,x} * G_{\sigma'}$, in which case the GSS can be viewed as a 'deep' CNN with pre-defined filters.

[4] anonymous link to code

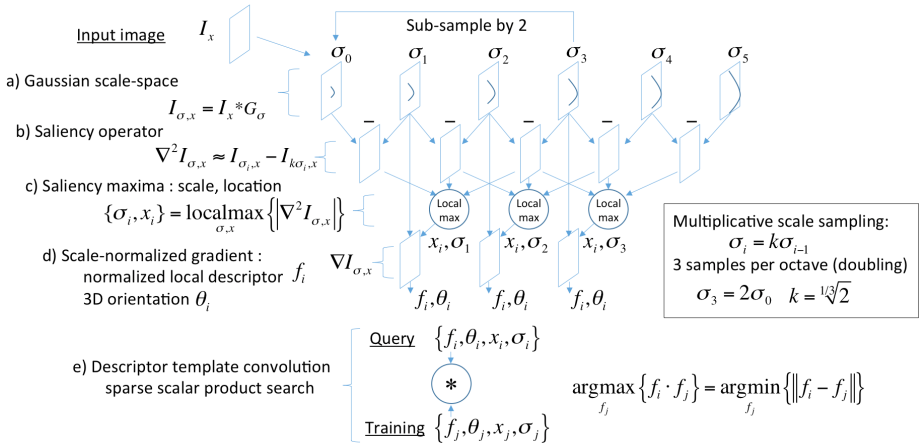


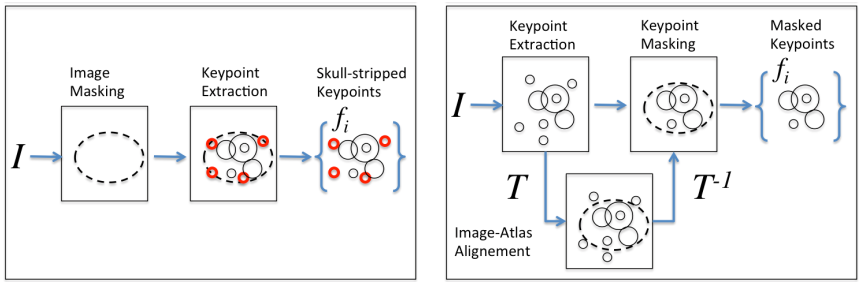
Fig. 1. The SIFT algorithm as a deep CNN. The Gaussian scale-space a) may be viewed as a deep CNN filtering process. Parallel networks approximate b) a Laplacian-of-Gaussian saliency operator as a difference-of-Gaussian (DoG) operation, where local saliency maxima $\{x_i, \sigma_i\}$ c) define the locations and scales of discrete scale-invariant keypoints representing informative, localizable image patches. Local scale-normalized image gradients d) are used to determine local keypoint orientation $\theta_i \in R^3$ and are sampled and normalized local keypoint descriptor templates $\{f_i\}$. Finally, e) peaks of convolution between the query image an a large bank of training descriptor templates $\{f_j\}$ can be efficiently detected via nearest neighbor search, as the Euclidean distance between normalised descriptors $\|f_j - f_j\|$ is a monotonically decreasing function of the scalar product $f_j \cdot f_i$. Scale space is sampled here at 3 equal multiplicative increments, similar to tones of an augmented triad in twelve-tone equal tempered musical scale.

Convolution via descriptor templates can be viewed as an evaluation of a scalar product evaluated at points on the geometrical lattice, for example the $\{x, \Theta, \sigma\} \in R^7$ coordinate space of 3D similarity transforms. Minimizing the distance between the normalized descriptors is equivalent to maximizing the convolution.

2.2 Masked Keypoint Analysis

Neuroimage datasets are often skull-stripped prior to processing, applying brain ROI masks in order to restrict analysis to data arising from neuroanatomical parenchyma as opposed to extraneous tissues such as skull, etc. Nevertheless, segmentation algorithms may produce variable or noisy results along the segmentation boundary, even for different images of the same subject. These hard, irregular boundaries generally exhibit unfavorable signal processing characteristics, e.g. Gibbs phenomenon, that may impact convolution responses, both in the cases of shallow Gaussian filters and via deep CNN filters. Recent solutions have investigated difference-of-Gaussian filtering to counter input bias [1], or conditional random field regularization of output noise [25].

We thus propose a new approach to neuroimage processing, particularly useful in the case of local keypoint analysis, as illustrated in Figure 2 a) which illustrates typical processing of skull-stripping pipeline, where image features extracted from a skull-stripped neuroimage may represent spurious, noisy content. We propose instead extracting features in natural image space and then use brain masks to separate keypoints present in the brain from others.



(a) Skull-stripped Keypoints (b) Masked Keypoints (proposed method)

Fig. 2. Illustrating a) Skull-stripped keypoint analysis, where skull-masking prior to filtering may lead to artifacts. b) Keypoint masking, where keypoints are extracted in natural image data, then filtered according to a mask. Transform T is a robust image-to-atlas similarity transform determined via feature-based alignment [22], and T^{-1} is used to transform an existing atlas brain mask to image space in a manner equivalent to keypoint transfer segmentation [24], except here the mask applied to filter keypoints rather than to mask image intensity data.

3 Experiments

3.1 Data

The dataset used in this experiment is a subset of 1010 subjects from the Human Connectome Project [23] Q4 release containing 439 unique families, including some unrelated subjects (see table 3.1). T1-weighted MR images have been acquired between 2012-2015 on a 3T MR scanner, at a 0.7mm isotropic resolution. Through the Freesurfer pipeline [6], images have been registered to the MNI space, brain masks have been generated, and images have been resampled to a 1.25mm isotropic resolution, as well as corrected for image artefacts such as eddy-currents and head-motion. Keypoints are extracted from individual images, numbers of keypoints per method is shown in table 3.1 .

3.2 Processing

We compare methods of keypoint extraction (see figure 1). Skull-stripped keypoints are generated by masking a brain volume with its mask, then extracting

Table 1. HCP demographic information

Image number	1010
age	29 ± 13
male	468
female	542
Full siblings (FS)	607
Dizygotic Twins (DZ)	71
Monozygotic Twins (MZ)	134

Table 2. Average number of keypoints extracted and pairwise correspondence counts.

methods		# keypoints	corres.
0.7 mm	skull-stripped	1468 ± 189	233.8
	masked	1662 ± 241	264.8
	original	2102 ± 277	335.4
1.25 mm	skull-stripped	180 ± 34	28.9
	masked	253 ± 54	40.8
	original	334 ± 60	53.8

keypoints from the resulting image. Masked keypoints are generated by extracting keypoints from the original brain volume, then using a brain mask to filter non-brain keypoints outside of the mask. We hypothesize masked keypoints will result in improved performance in indexing experiments, as they are not subject to irregular segmentation boundaries.

Our evaluation replicates the methodology of Chauvin et al. [3], measuring the effectiveness of each method at classifying relationships between subject pairs. We will use 1010 subjects from the HCP dataset instead of the 7536 originally used. A pairwise comparison is done measuring the Jaccard overlap (eq. 1) of each of the $\binom{N}{2} = N(N-1)/2$ image pairs. It is a measure of the proportion of the keypoint correspondences shared an image pair [3]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \tag{1}$$

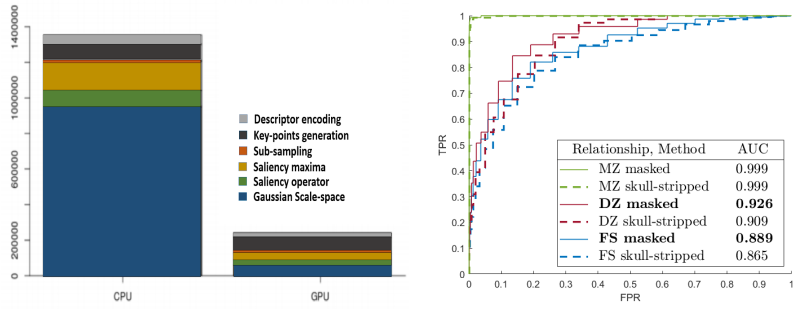
where $|A \cap B|$ represents the number of keypoint correspondences between image pair (A, B). Each class of relationships between pairs has a distinct Jaccard coefficient distribution that enables us to classify the relationship with a Jaccard coefficient threshold.

3.3 Keypoint extraction performance

Figure 3 a) shows a comparison of average computation times between a CPU and our GPU implementation for each processing step in the keypoint extraction process using a set of 1010 brain volumes. Overall, the GPU implementation results in an approximate 7x speedup. The biggest speedups observed were for the Gaussian scale-space (20X), saliency maxima (3X), sub-sampling (3X), and saliency operator (2X).

3.4 Keypoint visualisation

Figure 4 shows a saggittal brain MRI slice with original and skull-stripped keypoints. Matches between keypoints can be visualized and subsequently interpreted, and validated: Matches and non-matches between the keypoints are represented by different colored circles. Most matches between the two sets of



(a) CPU vs. GPU time (b) ROC curves

Fig. 3. Illustrating a) Comparison between average CPU and GPU processing times (in microseconds) for each processing step in the keypoint extraction process. b) ROC curves for relationship classification between pairs using a Jaccard score threshold (0.7mm resolution).

keypoints are further inwards from the mask edge, and most unmatched keypoints are closer to the edge of the brain. There is significantly more keypoints on the non-skull-stripped image, most of which can be found on the cortex.

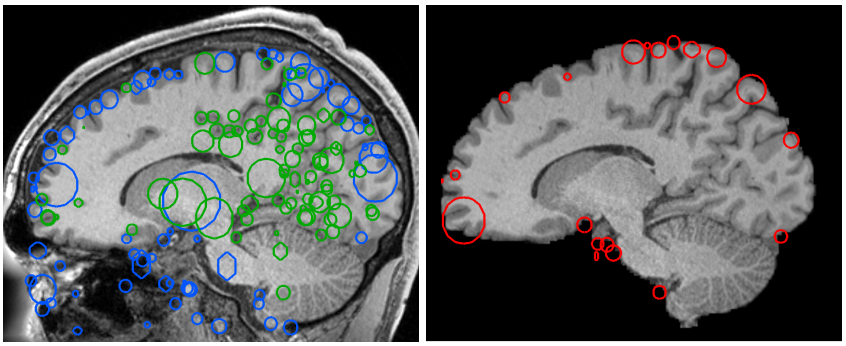


Fig. 4. Visualizing keypoints (circles) in original (left) and skull-stripped (right) images. Keypoints present in both images are shown as green (left), unique to original image as blue (left) and unique to skull-stripped image as red (right). Keypoint masking generally identifies additional keypoints located primarily on the cortex in regions affected by boundary artifacts.

Across all images at 0.7mm resolution, 85% of the skull-stripped keypoints matched with the original keypoints of the same image, while at a 1.25 mm resolution, 75% of the skull-stripped keypoints matched with the original keypoints. To test our hypothesis that the percentage of matches scales with the volume of the mask, we modeled the brain as a sphere of keypoints, with it's

volume being the total number of skull-stripped keypoints and the surface being the skull-stripped keypoints that did not match. Using this model, we predicted that 86% of the skull-stripped keypoints will match at 0.7mm resolution using the 1.25mm data. This is a simple model accurately representing our intuition that the border interference effect is dependent on the size of the mask.

3.5 HCP Family Relationship Classification

We performed experiments to measure the pairwise similarity between 1010 subjects from the HCP dataset using the Jaccard overlap score introduced in [3] following the same method. In the original article, skull-stripped keypoints at 0.7mm resolution were used. We compared the ability to find relationships between subjects pairs using original, masked, and skull-stripped keypoints at both 0.7mm and 1.25mm resolutions. The figure 3 b) shows ROC curves for the masked and skull-stripped representations at 0.7mm. Though the area under the curve (AUC) is similarly very high for MZ cases using masked and skull-stripped points, a higher AUC is observed in the case of FS and DZ using the masked points when compared to the skull-stripped points. Unlike previous work, the proposed masked method also leads to statistically significant differences between DZ and FS brain similarity at a 1.25mm resolution. This may be because we have never been able to observe cortical morphology in this amount of detail, due to skull-stripping noise.

Table 3. AUC values for different keypoint representations and resolutions

keypoints		FS	DZ	MZ
0.7 mm	skull-stripped	0.865	0.909	**0.999**
	masked (ours)	**0.889**	**0.926**	**0.999**
	original	0.931	0.970	**0.999**
1.25 mm	skull-stripped	0.824	0.851	0.991
	masked (ours)	**0.858**	**0.905**	**0.998**
	original	0.889	0.950	**0.998**

Table 3.5 compares relationship classification using original, masked, and skull-stripped keypoints at different resolutions. Using masked keypoints results in higher AUC than skull-stripped keypoints for any relationship at both resolutions. The increase in AUC is amplified at a lower resolution, because a higher fraction of the skull-stripped keypoints are affected by the brain mask.

4 Conclusion

We presented a fast GPU-based method for keypoint extraction based on a Gaussian scale space and ROI masking. The proposed method is the first GPU implementation of 3D SIFT, is invariant to 3D similarity transforms, offers a

solution to convolution filter bias and circumvents limitations introduced by hard boundaries typically present in neuroimaging analysis, and is interpretable. Analysis using this method led to improvements to the current state-of-art for family indexing on a large cohort of 3D MRI data, and to significant speedups compared to a CPU implementation. This method can be used in medical image analysis studies with a variety of image modalities and anatomical structures.

Our analysis opens various venues for future technological advancements. Keypoint networks can be trained for specific tasks [17, 26] but a challenge will be coping with bias towards training data. While the majority of CNN implementations are limited to translation invariant convolution via brute force convolution over the 3-parameter space of 3D translations, we demonstrate that nearest neighbor correspondences between normalized descriptors achieve convolution peaks across 7-parameter 3D similarity transforms, thus offering a mechanism to achieve invariance to orientation and scaling within the CNN framework.

References

1. Reza Azad, Abdur R Fayjie, Claude Kauffman, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. On the texture bias for few-shot cnn segmentation, 2020.
2. Mårten Björkman, Niklas Bergström, and Danica Kragic. Detecting, segmenting and tracking unknown objects using multi-label mrf inference. *Computer Vision and Image Understanding*, 118:111–127, 2014.
3. L. Chauvin, K. Kumar, C. Wachinger, M. Vangel, J. de Guise, C. Desrosiers, W. Wells, and M. Toews. Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. *NeuroImage*, 2019.
4. Jimit Doshi, Guray Erus, Yangming Ou, Bilwaj Gaonkar, and Christos Davatzikos. Multi-atlas skull-stripping. *Academic radiology*, 20(12):1566–1576, December 2013.
5. Simon F. Eskildsen, Pierrick Coupé, Vladimir Fonov, José V. Manjón, Kelvin K. Leung, Nicolas Guizard, Shafik N. Wassef, Lasse Riis Østergaard, and D. Louis Collins. Beast: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362 – 2373, 2012.
6. Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.
7. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
8. Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.
9. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
10. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

11. Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.
12. Stefan Lindholm and Joel Kronander. Accounting for uncertainty in medical data: A cuda implementation of normalized convolution. In *Proceedings of SIGRAD 2011. Evaluations of Graphics and Visualization—Efficiency; Usefulness; Accessibility; Usability; November 17-18; 2011; KTH; Stockholm; Sweden*, number 065, pages 35–42. Linköping University Electronic Press, 2011.
13. David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.
14. Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
15. Don P. Mitchell and Arun N. Netravali. Reconstruction filters in computer-graphics. *SIGGRAPH Comput. Graph.*, 22(4):221–228, June 1988.
16. Marius Muja and David G Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2227–2240, 2014.
17. Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6234–6244, 2018.
18. Samuel Ritter, David G. T. Barrett, Adam Santoro, and Matt M. Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2940–2949, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
19. Sudipta N Sinha, Jan-Michael Frahm, Marc Pollefeys, and Yakup Genc. Gpu-based video feature tracking and matching. In *EDGE, workshop on edge computing using new commodity architectures*, volume 278, page 4321, 2006.
20. Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.
21. F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060 – 1075, 2004.
22. Matthew Toews and William M Wells III. Efficient and robust model-to-image alignment using 3d scale-invariant features. *Med Image Anal*, 17(3):271–82, 2013 Apr 2013.
23. David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
24. C. Wachinger, M. Toews, G. Langs, W. Wells, and P. Golland. Keypoint transfer for fast whole-body segmentation. *IEEE Transactions on Medical Imaging*, 39(2):273–282, Feb 2020.
25. Christian Wachinger, Martin Reuter, and Tassilo Klein. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.
26. Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016.