HAPNet: Hierarchically aggregated pyramid network for real-time stereo matching

Patrick Brandao, Dimitris Psychogyios, Evangelos Mazomenos, Mirek Janatka
and Danail Stoyanov

Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS)
University Collge London, London, UK
{patrick.brandao.15,danail.stoyanov}@ucl.ac.uk

Abstract. Recovering the 3D shape of the surgical site is crucial for multiple computer assisted interventions. Stereo endoscopes can be used to compute 3D depth but computational stereo is a challenging, non-convex and inherently discontinuous optimization problem. In this paper, we propose a deep learning architecture which avoids the explicit construction of a cost volume of similarity which is one of the most computationally costly blocks of stereo algorithms. This makes training our network significantly more efficient and avoids the needs for large memory allocation. Our method performs well, especially around regions comprising multiple discontinuities around surgical instrumentation or around complex small structures and instruments. The method compares well to the state-of-the-art techniques while taking a different methodological angle to computational stereo problem in surgical video.

1 Introduction

Robot-assisted interventions rely on stereo endoscopes and this presents an opportunity to recover of the underlying 3D structure of the operating field *in vivo* using computational stereo. 3D information is essential for registering pre-operative data to the surgical field-of-view using augmented reality [1], enabling dynamic active constraints or motion compensation using robot control [2, 3]. However, computational stereo in intra-operative endoscopic images remains very challenging due to reflective surfaces, large instrument-tissue discontinuities and regions where the texture of organ surfaces is homogeneous [2, 3].

Recent efforts have shown convolutional neural networks (CNNs) can be used for surgical scene reconstruction by using a standard encoder-decoder network and relying on feature extraction, a feature correlation metric and maximizing a spatial consistency[1]. Notable performance boosts can be achieved designing architectures that avoid such explicit steps [4, 5]. Siamese networks can create a high level representation of the data where deep features are aggregated and used for disparity computation [6–8]. End-to-end architectures can efficiently perform cost volume regularization by using 3D convolutions [6]. The quality of the features extracted to build the cost volume can also be enhanced [9], however, both manual alignment of deep features and 3D convolutions are both

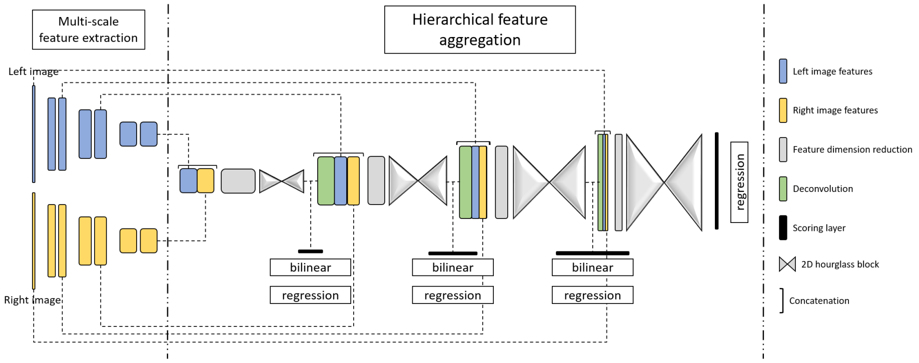

Fig. 1. Illustration of our Hierarchically aggregated pyramid network (HAPNet) architecture and building blocks. The legend on the right clarifies the function of different layer blocks after the input images which can be viewed in two phases, one focusing on individual image information and the second combining the stereo pair.

computationally and memory demanding operations, and correlation layers collapse the feature dimension, limiting the contextual information used in disparity regularization.

In this paper, we avoid the use of 3D convolution and explicit feature alignment operations to create a fast, memory efficient CNN, capable of accurate and real-time computational stereo. Our approach is fast and efficient yet it is able to produce comparable results to the state-of-the-art. We sequentially aggregate feature representations of the stereo pair from lower to higher resolutions, allowing point correspondences to be encoded by the network without the need of slow nested loop operations or feature replication. To use small encoder-decoder blocks to encode point correspondence at different scales, achieving an effective receptive field big enough for most disparities without requiring a massive amount of complexity. The main contributions of our work are:

- To sequentially aggregate feature representations of the stereo pair from the lower to the higher resolutions, allowing point correspondences to be encoded by the network without the need of slow nested loop operations or feature replication.
- To use small encode-decoder blocks to encode point correspondence at different scales, achieving an effective receptive field big enough for most disparities without requiring a massive amount of complexity.

2 Methods

We focus on creating an accurate, fast and memory efficient model for stereo matching. One of the fastest ways to create a cost volume is to simply concatenate the feature representation of both stereo images. In this case, if we solely focus on the problem of finding spatial displacements between corresponding points, the network would need an effective receptive field equal or larger than the biggest disparity considered. Due to computational complexity restrictions, the most common way to increase a network receptive field is the use of downsampling operations, which in turn, can cause some loss of detail during pixel-level matching. A similar argument for the feature extraction step has been made by Brandao *et al.* [10], showing that the best accuracy is achieved with a compromise between the size of the receptive field and the loss of fine detail. Our fast and memory efficient approach avoids the use of correlation layers or manually built cost volumes. Our hierarchically aggregated pyramid network (HAPNet), illustrated in Figure 1, allows a theoretical receptive field big enough to infer big disparities without losing information about fine details important for small objects. The detailed parameters of the proposed HAPNet are detailed in Table 1.

2.1 Multi-resolution feature extraction

The first part of the HAPNet model is responsible for extracting deep feature descriptors of the stereo image pair. High dimensional representations are more robust to appearance ambiguities and can incorporate local context [6].

Our feature extractor is a Siamese network built by stacking three sequential pairs of convolution layers. Each pair starts with a 2-strided convolution, halving the spatial resolution and doubling the feature dimensionality. This allows to extract a deep feature representation downsampled by three different factors: 8, 4 and 2. The weights of both branches are shared to more effectively learn corresponding features. A detailed description of this architecture is presented in Table 1 and the accompanying Figure 1.

2.2 Hierarchical feature aggregation

If the ability to find point correspondences and compute their distance can be encoded by a fixed number of stacked convolution layers then, in theory, this ability is only limited by the effective receptive field of those stacked layers. Considering the range of disparities expected in most real cases, the required receptive field can only be achieved by using several downsample operations through the network. As mentioned before, high level features tend to lose more detail, so most approaches choose to handle the receptive field requirement by using a correlation layer or by manually aligning features [8].

We avoid the use of correlation layers or manually built cost volumes by performing a coarse to fine concatenation of the feature representations of the stereo pair. Our pyramid network encodes point correspondences at multiple resolutions, progressing from coarse to finer prediction. This allow us to have a receptive field big enough to encode large displacements in the lower resolutions and to encode finer correspondences in the higher ones. The full architecture is illustrated in Figure 1 with a more detailed description of the parameters in Table 1.

Table 1. Summary of the proposed HAPNet. Each convolutional layer represents a block of convolution, batch normalization and ReLU nonlinearity except for the scoring layers

Siamease network for multi-scale feature extraction

Name	k	s	Input	Output Dimension
conv1_left	3	2	Left Image	4*H/2 x W/2 x 32
conv1_right	3	2	Right Image	
conv2_left	3	1	conv1_left	
conv2_right	3	1	conv1_right	
conv3_left	3	2	conv2_left	4*H/4 x W/4 x 64
conv3_right	3	2	conv2_right	
conv4_left	3	1	conv3_left	
conv4_right	3	1	conv3_right	
conv5_left	3	2	conv4_left	4*H/8 x W/8 x 128
conv5_right	3	2	conv4_right	
conv6_left	3	1	conv5_left	
conv6_right	3	1	conv5_right	

2D Hourglass network

Name	k	s	Input	Output Dimension
input	-	-	-	H x W x F
conv1	3	2	input	H/2 x W/2 x 2F
conv2	3	2	conv1	H/4 x W/4 x 4F
conv3	3	1	conv2	H/4 x W/4 x 4F
deconv	3	2	conv3	H/2 x W/2 x 2F
residual1	-	-	deconv conv2	H/2 x W/2 x 2F
deconv2	3	2	residual1	H x W x F
residual2	-	-	deconv2 conv1	H x W x F

Hierarchical Feature Aggregation

Name	k	s	Input	Output Dimension
concat_x8	-	-	conv6_left conv6_right	H/8 x W/8 x 256
conv7	3	1	concat_x8	H/8 x W/8 x 256
2D_hourglass_x8	-	-	conv7	H/8 x W/8 x 256
score_x8	3	1	2D_hourglass_x8	H/8 x W/8 x 1
deconv1	3	2	2D_hourglass_x8	H/4 x W/4 x 128
concat_x4	-	-	deconv1 conv4_left conv4_right	H/4 x W/4 x 256
conv8	3	1	concat_x4	H/4 x W/4 x 128
2D_hourglass_x4	-	-	conv8	H/4 x W/4 x 128
score_x4	3	1	2D_hourglass_x4	H/4 x W/4 x 1
deconv2	3	2	2D_hourglass_x4	H/2 x W/2 x 64
concat_x2	-	-	deconv2 conv2_left conv2_right	H/2 x W/2 x 128
conv9	3	1	concat_x2	H/2 x W/2 x 64
2D_hourglass_x2	-	-	conv9	H/2 x W/2 x 64
score_x2	3	1	2D_hourglass_x2	H/2 x W/2 x 1
deconv3	3	2	2D_hourglass_x2	H x W x 32
concat_x1	-	-	deconv3 Left Image Right Image	H x W x 38
conv10	3	1	concat_x1	H x W x 32
2D_hourglass_x1	-	-	conv10	H x W x 32
final_score	3	1	2D_hourglass_x1	H x W x 1

2.3 2D hourglass Network

While computationally efficient, simple feature concatenation does not implicitly encode spatial correspondences like a correlation layer or a manually built cost volume. Because of this, we introduce small encoder-decoder networks to encode stereo matches.

Our 2D hourglass network consists of single 3×3 convolution layers, with two levels of downsampling, followed by two deconvolution layers with residual connections [9]. One important aspect is that the network maintains the same feature dimensionality and resolution as the input. The full description of the 2D hourglass block is presented in Table 1.

2.4 Scale-aware disparity regression

In pixel-wise problems, such as semantic segmentation, it is common to add loss functions at different levels of the network. However, the stereo matching problem has another particularity given that the distance (in pixels) between two points varies when we rescale the stereo pair. For example, two corresponding points will be two times closer when the feature space is down-sampled by a factor of two. Because of this, for the output of the network for each pixel, d_n, we use a absolute difference loss where the labels, y_n, are scaled by the pyramid's level downsample factor, s.

Table 2. Evaluation with different settings on the scene flow test set. We computed the percentage of three-pixel-error, >3px, of five-pixel-error, >5px, and the mean average error (MAE). Results are comparative to metrics reported in [5, 7, 6, 8, 7]

HAPNet settings			Scene Flow test set			time (s)
2D Stacked Hourglass	Scale-aware loss	Negative mining	>3 px (%)	>5 px (%)	MAE (px)	
			10.65	7.17	2.92	0.05
✓			9.16	6.19	1.89	
✓	✓		7.69	5.09	1.69	
✓	✓	✓	6.62	4.24	1.40	

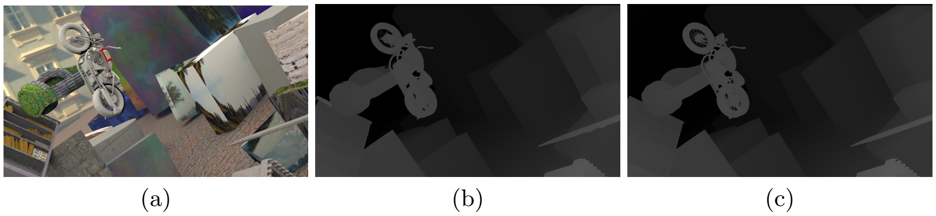
(a) (b) (c)

Fig. 2. Scene Flow test set qualitative results. (a) left stereo input image; (b) disparity prediction; (c) ground truth.

$$Loss = \frac{1}{N}\sum_{n=1}^{N}\left\| d_n - \frac{y_n}{s} \right\| \quad (1)$$

Apart from deriving a more accurate representation of the matching problem, scaling the labels by the downsample factor of the network's level also implicitly minimizes the importance of small displacements in the lower resolution levels.

3 Experimental setup

We train and quantitatively evaluate our method using a popular stereo dataset: Scene flow [5]. The Scene Flow dataset created from synthesized environments, containing 35,454 training image pairs and 4,370 testing image pairs. We use the Scene Flow dataset to investigate the effect of different aspects of our method. Evaluation is done in natural and medical environments. When public datasets are used, the recommended evaluation protocol and metrics where implemented.

3.1 Experimental details

All parameters are randomly initialized with a normalized Gaussian distribution and input images are color normalized to have zero mean and unit standard deviation. All models were end-to-end trained with Adam optimizer [11] and a batch size of 4. During training, we randomly sample smaller patches of size 320 × 640 to allow more diverse training batches while being memory efficient. The maximum disparity was set to 192.

Table 3. Comparative results on the Scene Flow testing set for other stereo CNNs. Four different metrics are presented: three-pixel-error,>3px, of five-pixel-error, >5px, the mean average error, MAE and total running time in seconds

Model	>3 px	>5px	MAE	Time (s)
DispnetC [5, 7]	9,67	-	1.84	0.06
GC-Net [6]	9.34	7.22	2.51	0.95
iResNet-i3 [8]	4.57	3.32	2.45	0.148
CRL [7]	6.20	-	1.32	0.47
HAPNet (ours)	6.62	4.24	1.40	0.05

We train our models from scratch using the scene flow dataset with a initial learning rate of 1×10^{-3} for 300K iterations. We also perform negative mining by training the model an additional 5K iterations with images that have a predicted 3-pixel error bigger than 10%.

All models were developed using Tensorflow [12] and trained on a single Nvidia Titan Xp GPU.

3.2 Scene flow

We use the scene flow dataset to evaluate the importance of key ideas in this paper. Scene flow is the only stereo dataset big enough to train deep networks without over-fitting and to provide dense ground truth without any discrepancies due to erroneous labels. Table 2 lists comparative results of the different variants of the proposed HAPNet.

We first evaluate the effect of the 2D stacked hourglass encoders. By replacing this block with the same number of convolution layers but without the encoder-decoder structure. We verified that the hourglass block results in big improvements for large disparity matches, resulting in a considerable decrease of the mean average error. This indicates that the bigger receptive field of the 2D stacked hourglass block is essential for the network to be able to encode large distance correspondences. The proposed scaled loss also results in an incremental improvement in all evaluation metrics. Finally, the negative mined images were mostly stereo pairs with large disparity objects, which also significantly improved large displacement predictions. Figure 2 shows qualitative results of our best model.

When comparing our work with other methods (Table 3), IResNet-i3 [7] achieves a slightly lower 3 and 5 pixel error but with a relatively high MAE. The CRL network [7] performs with a lower mean average error but requires much more computational power. Our performance comes slightly under CRL's, even beating the very deep and regularized GC-Net [6], but HAPNet is 10 times faster. The only network comparable in speed is the DispnetC [5, 7], which underperforms our model in every metric. Our results show that by simply adapting the architecture to the particularities of the stereo matching problem, significant improvements can be achieved. Because we avoid computational demanding op-

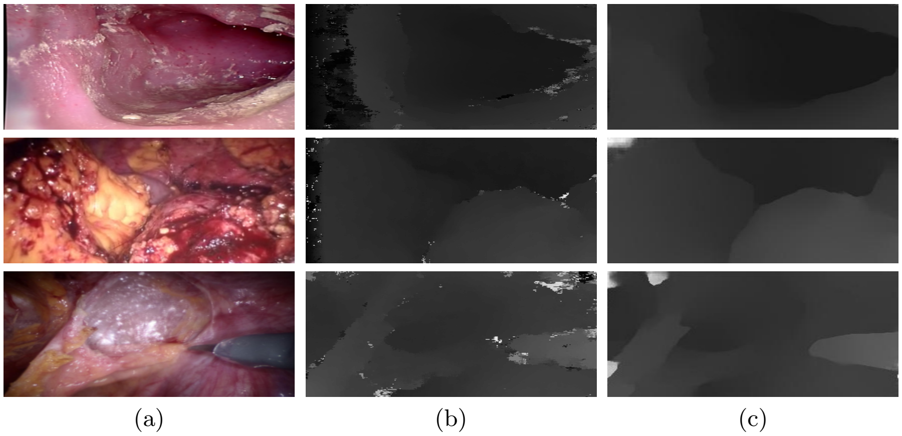

Fig. 3. Colon phantom, partial nephrectomy and prostatectomy qualitative results. (a) left stereo input image; (b) disparity prediction without spatial consistency [10]; (c) disparity prediction with the proposed method.Images re-sized to 192×384 and processed in 0.014s.

erations, such as building a manual cost volume, our model can run in real time with much smaller memory requirements.

3.3 Qualitative medical data

We qualitatively evaluate our method in several different medical environments. Figure 3 presents disparity estimations from a PVA-C colon phantom manufactured from a 3D model of a human colon, a partial nephrectomy and a prostatectomy procedures. Stereo data was acquired using a stereo camera from a da Vinci Surgical System.

The low amount of detail and repetitive patterns in the colon phantom makes it a particularly hard to find accurate stereo matches. Even when using deep features [10], pixel level matching tends to be noisy and unreliable. On the other hand, HAPNet produces visibly smother and more accurate disparity maps by guarantying the spatial consistency of the environment. Because the features are aggregated at different levels of the network, HAPNet is still able to handle sharp depth transitions. A similar analysis can be done in Figure 3, where the HAPNet tries to maintain depth consistency for the different tissue areas and tools of the image. Because the resolution of this images is substantially lower that the ones in the training data, some of the tools edges in the disparity maps are not as sharp as the ones in the scene flow dataset.

3.4 Quantitative medical data

We also quantitatively evaluate out method on a public surgical stereo dataset [13] depicting different real organs(liver, kidney, heart) captured from different

Table 4. 3D error statistics as reported in [13]

Method	Mean	SD	RMS	Median	Lower quartile	Upper quartile	Min	Max
MADNet [14]	13.32	14.02	19.34	5.76	2.80	21.68	0.87	50.32
DeepPruner [15]	22.83	18.41	29.33	13.58	7.01	38.28	1.50	62.06
DispNet [5, 7]	7.47	8.68	11.45	4.98	2.90	7.62	1.43	49.36
HAPNet (ours)	2.46	1.39	2.82	2.17	1.48	2.95	0.54	6.34

angles and distances. Each sample contains two stereo image pairs(distorted and stereo-rectified), a stereo calibration file, ground truth 3D reconstruction and validation masks to limit the evaluation of the outputs in a specified region. The 3D geometry of the tissue was captured using CT scans and the registration between the stereo images and the reconstructed scene was done using markers visible both in the CT scan and the images.

We compare our method with three other publicly available models trained on Scene Flow dataset [5]: DispNet [5, 7], MADNet [14] and DeepPruner [15]. Predictions from networks are used to create 3D point clouds and the resulting reconstructions are used to calculate all the error statistics metrics. The results are presented in Table 4.

The proposed method outperforms all the other models in every single metric without sacrificing computational efficiency. Table 4 shows that other models struggle to accurately reconstruct challenging surgical environments. On the other hand, even though it was trained with the same data, HAPNet is able to better generalize achieving a mean error of 2.46 mms.

4 Conclusion

In this paper, we have proposed a novel, fast and memory efficient end-to-end architecture for stereo vision and 3D surgical site reconstruction. We show that our architecture is able to learn to regress disparity without any additional post-processing or regularization which is appealing for a number of practical reasons and can cope with some of the challenges commonly faced in surgical video data. Experimentally, in the paper we have demonstrated that significant improvements in 3D reconstruction are possible by small, problem specific adaptations that simplify the learning problem.

Our approach achieves competitive performance on existing large vision datasets for non-surgical applications, like Scene Flow, while being substantially faster than all other architectures we compared against. For robotic surgery video, we show that our model encapsulates a wider receptive field which has a significant impact on dealing with high disparity discontinuities due to verged cameras. Our method also seems to perform better regularization and presents significantly more compelling quantitative and qualitative results than previously reported work.

Interesting directions for future work would be to combine our architecture with monocular depth estimation models [16] to potentially enable 3D estimation not only with stereo systems. It would also be compelling to link camera motion either through vision or through the robot kinematics into the system to allow wider field-of-view reconstructions and surgical site mapping [17, 18].

References

1. H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *CVPR*, 2018, pp. 340–349.
2. D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *MICCAI 2010*.
3. A. R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "3d reconstruction of whole stomach from endoscope video using structure-from-motion," in *EMBC*. IEEE, 2019, pp. 3900–3904.
4. M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
5. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conference on CVPR*, 2016, pp. 4040–4048.
6. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," *CoRR*, vol. abs/1703.04309, 2017. [Online]. Available: http://arxiv.org/abs/1703.04309
7. J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *International Conf. on Computer Vision-Workshop on Geometry Meets Deep Learning (ICCVW 2017)*, vol. 3, no. 9, 2017.
8. Z. Liang, Y. Feng, Y. Guo, H. Liu, L. Qiao, W. Chen, L. Zhou, and J. Zhang, "Learning deep correspondence through prior and posterior feature constancy," *arXiv preprint arXiv:1712.01039*, 2017.
9. J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," *arXiv preprint arXiv:1803.08669*, 2018.
10. P. Brandao, E. Mazomenos, and D. Stoyanov, "Widening siamese architectures for stereo matching," *arXiv preprint arXiv:1711.00499*, 2017.
11. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
12. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.
13. L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N. Clancy, D. S. Elson, S. Haase, E. Heim, *et al.*, "Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction," *IEEE transactions on medical imaging*, vol. 33, no. 10, pp. 1913–1930, 2014.
14. A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, "Real-time self-adaptive deep stereo," in *IEEE Conference on CVPR*, 2019, pp. 195–204.
15. S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "Deeppruner: Learning efficient stereo matching via differentiable patchmatch," in *IEEE ICCV*, 2019, pp. 4384–4393.

16. A. Rau, P. J. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 7, pp. 1167–1176, 2019.
17. C. da Costa Rocha, N. Padoy, and B. Rosa, "Self-supervised surgical tool segmentation using kinematic information," in *2019 ICRA*. IEEE, 2019, pp. 8720–8726.
18. F. Qin, Y. Li, Y.-H. Su, D. Xu, and B. Hannaford, "Surgical instrument segmentation for endoscopic vision with data fusion of rediction and kinematic pose," in *2019 ICRA*. IEEE, 2019, pp. 9821–9827.