Supplementary Material: EpO-Net: Exploiting Geometric Constraints on Dense Trajectories for Motion Saliency

Muhammad Faisal[1], Ijaz Akhter[2], Mohsen Ali[1], and Richard Hartley[3]

[1]Information Technology University of the Punjab Lahore, Pakistan
[2]KeepTruckin, Inc, [3]Australian National University, Australia
{m.faisal, mohsen.ali}@itu.edu.pk, ijaz.akhter@keeptruckin.com,
richard.hartley@anu.edu.au

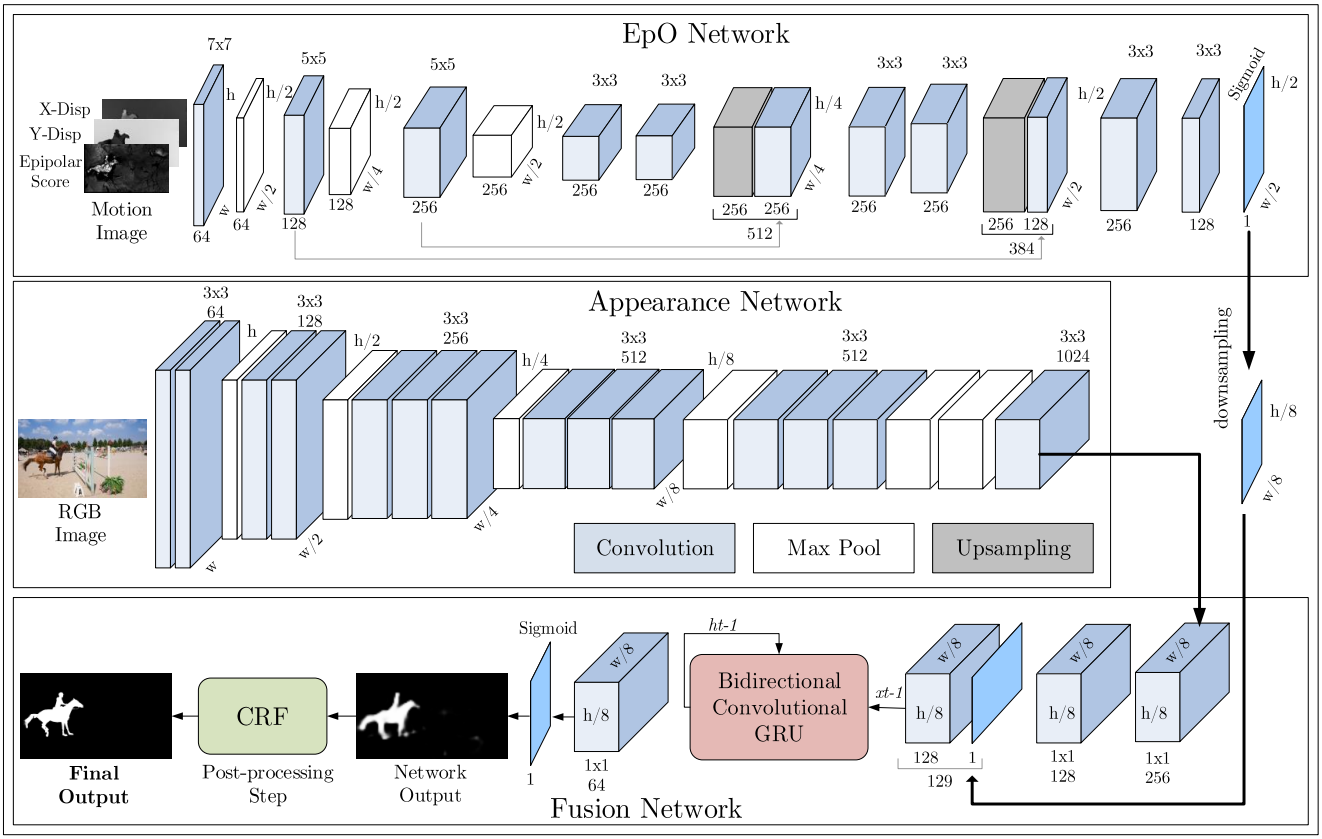

Figure 1: Network Architecture depicting different parts & information transition between different modules. Top Row: network parameters of EpO-Net, that takes motion frame (Optical Flow + Epipolar Distance) as input and output a motion saliency map. Middle Row: appearance network takes RGB video frame and produces an intermediate representation. Bottom Row: Fusion Network combines the two-stream representation (Motion Saliency + Appearance Features) using the Convolutional GRU module and CRF is applied to produce the final segmentation result. The height (h), width (w) of the feature maps, and size of convolutional filters are mentioned at each layer.

Figure 2: Qualitative Comparison of our our motion network (EpO-Net in Red) with Mp-Net [6] (Blue) against the ground truth (Green). EpO-Net only relies on optical flow and dense trajectories based epipolar distances. Unlike MpNet, we do not require any objectness score for motion segmentation.

Figure 3: Qualitative Comparison of our EpO+ with STP [1], LSMO [8] (an improved version of LVO [7]), MOA [4], and AGS [9] (an improved version of PDB [5]) on validation sequences of DAVIS-2016, all the networks are trained on DAVIS-2016. Green is GroundTruth, Blue is STP, LSMO, MOA, AGS and Red is EpO+ (our fusion network). Most of the errors in the other methods are over-segmenting, and are due to over-exploitation of appearance information. While the proposed method, due to more informative proposed motion features (based on geometric constraints) and input-dropout training procedure, is being able to learn, how to balance appearance and motion cues.

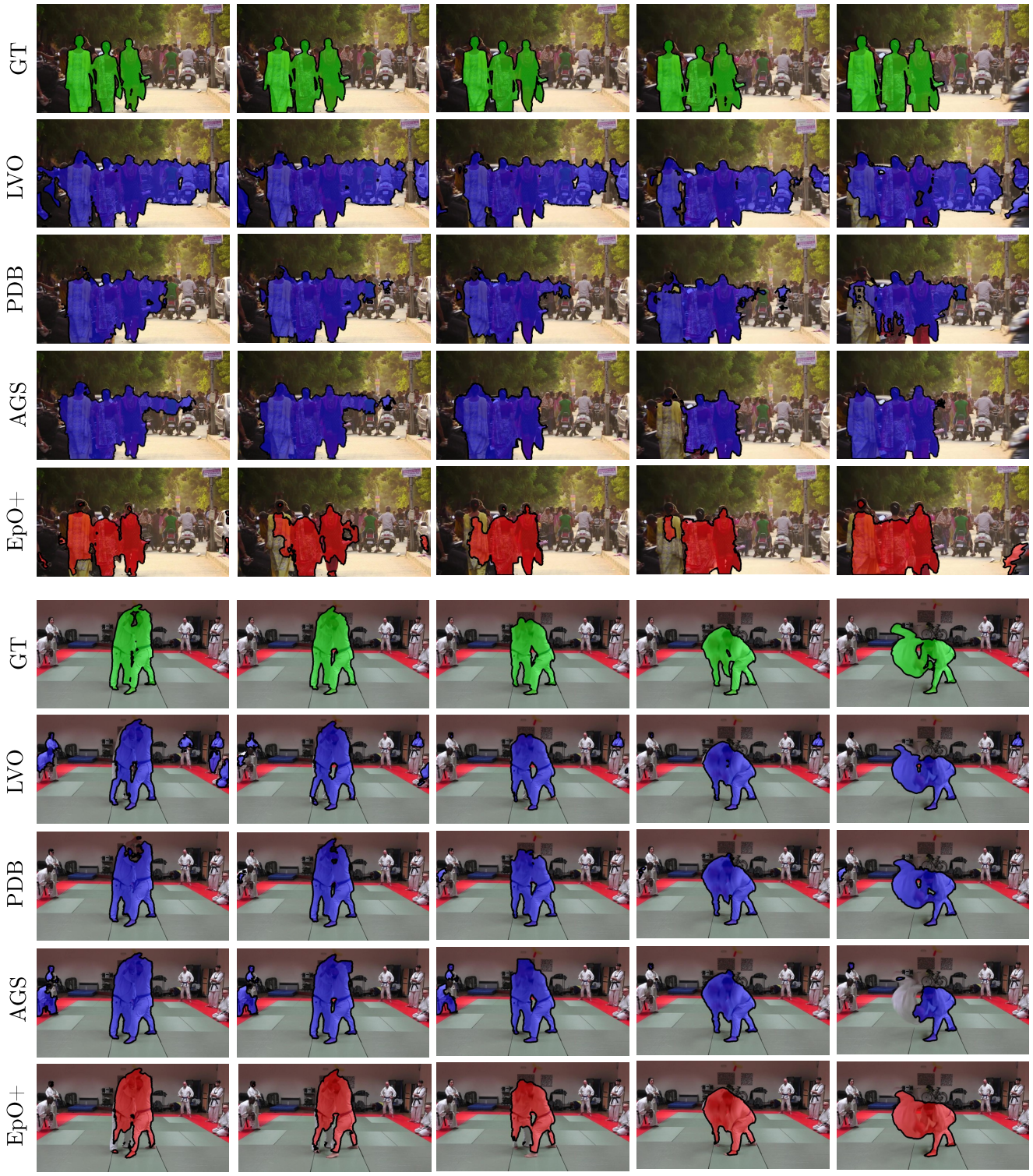

Figure 4: Qualitative Comparison of our EpO+ with LVO [7], PDB [5], and AGS [9] on DAVIS-2017, all the networks are trained on DAVIS-2016 and tested on validation set of DAVIS-2017. Green is GroundTruth, Blue is LVO, PDB, AGS and Red is EpO+ (our fusion network). Most of the errors in the other methods are over-segmenting, and are due to over-exploitation of appearance information. While the proposed method, due to more informative proposed motion features (based on geometric constraints) and input-dropout training procedure, is being able to learn, how to balance appearance and motion cues.

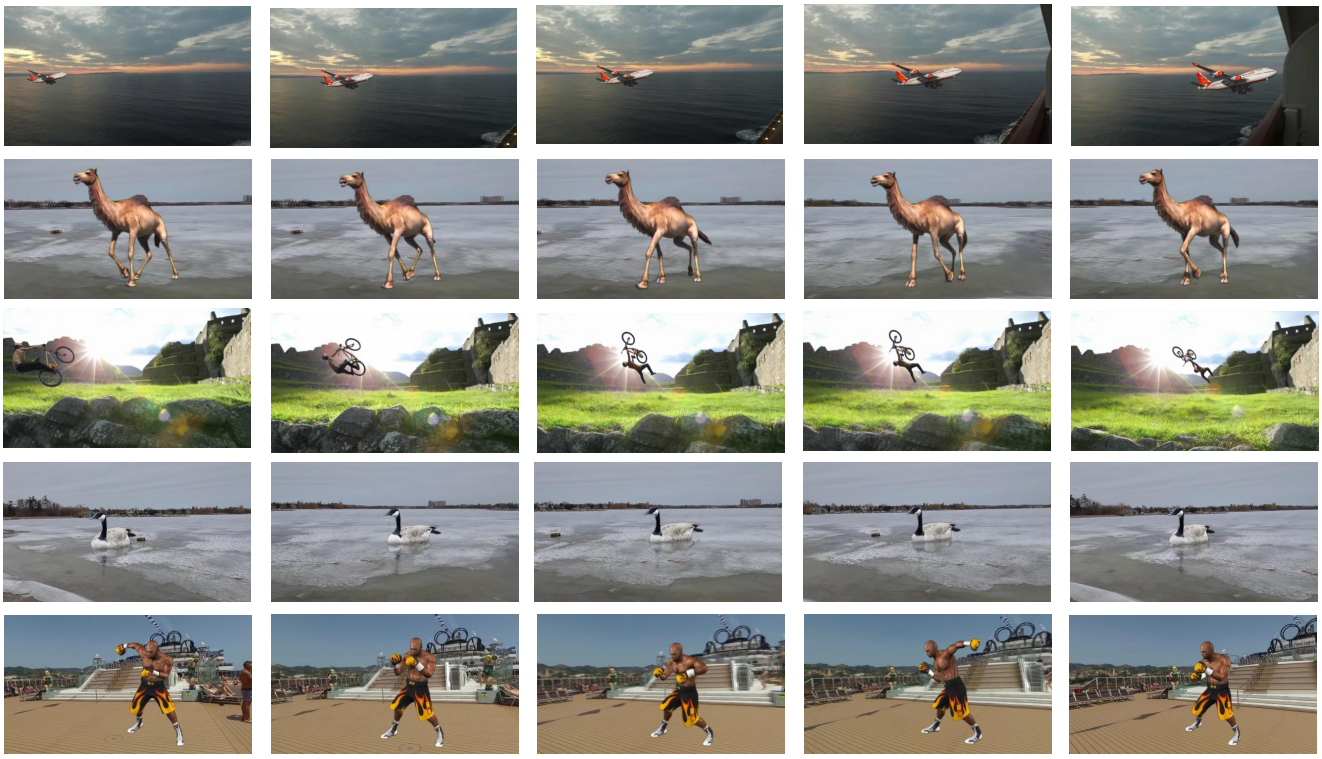

Figure 5: Frames from RBSF dataset sequences are shown.

RBSF Dataset: We create our own synthetic dataset, called **RBSF** (Real Background, Synthetic foreground), by overlaying 20 different foreground objects performing various movements with 5 different real background videos. We downloaded both the foreground and background videos from YouTube and mixed them using Video Editing Tool. The foreground objects include running dog, cat, camel, jumping human, football player, dancing girl, etc. The background videos include scenes capturing riversides, shipyard, house-outdoors, houses-indoors, and playgrounds. The videos are stored in 720p (720x1280) resolution and foreground objects are fairly large size (30% to 50% of the frame). The reasonable fast motion of foreground objects allows us to compute accurate optical flow and long trajectories.

We use this dataset to only train the EpO (motion) network, using the optical flow and the epipolar distances. No appearance information was used, therefore, repeating background shall not affect the learning process. It is easy to scale this dataset up. However, for our requirements, the dataset is large enough, containing 100 videos and 19,797 frames., We observe that generating more data does not improve results, thanks to the well-constrained epipolar distances (ED). Therefore the backgrounds do not affect the training. Frames from few sequences from RBSF are shown in Fig 5. The dataset is released at https://github.com/mfaisal59/RBSF.

Attribute	EpO+	EpO	AGS[9]	MOA[4]	LSMO[8]	STP[1]	PDB[5]	ARP[3]	LVO[7]	Mp-Net[6]	FSeg[2]
AC	**0.83** ***-0.04***	0.77 *-0.03*	0.80 *-0.01*	0.78 *-0.01*	0.78 *+0.00*	0.72 *+0.07*	0.78 *-0.01*	0.79 *-0.04*	0.75 *+0.01*	0.71 *-0.02*	0.71 *-0.01*
DB	**0.72** ***+0.10***	0.63 *+0.14*	0.66 *+0.16*	0.61 *+0.20*	0.55 *+0.27*	0.66 *+0.15*	0.62 *+0.17*	0.72 *+0.05*	0.55 *+0.24*	0.58 *+0.14*	0.50 *+0.24*
FM	**0.78** ***+0.04***	0.72 *+0.06*	0.77 *+0.04*	0.74 *+0.05*	0.73 *+0.08*	0.75 *+0.04*	0.74 *+0.04*	0.73 *+0.04*	0.70 *+0.09*	0.68 *+0.04*	0.68 *+0.04*
MB	**0.78** ***+0.06***	0.67 *+0.14*	0.74 *+0.10*	0.71 *+0.10*	0.73 *+0.10*	0.74 *+0.06*	0.72 *+0.09*	0.71 *+0.09*	0.71 *+0.09*	0.65 *+0.10*	0.64 *+0.13*
OCC	0.75 *+0.08*	0.67 *+0.11*	0.76 *+0.05*	0.78 *-0.02*	0.74 *+0.06*	**0.81** ***-0.05***	0.76 *+0.02*	0.72 *+0.06*	0.73 *+0.03*	0.69 *+0.01*	0.61 *+0.13*

Table 2: The Attribute Analysis comparing state-of-the-art methods on DAVIS-2016 dataset. Mean IoU for a specific attribute: Appearance change (AC), Dynamic Background (DB), Fast Motion (FM), Motion Blur (MB) and Occulusion (OCC) is presented. The values in small font size show the increase or decrease in performance without the sequences corresponding to that specific attribute. Our method outperforms in AC, FM, DB, and MB. The best scores are highlighted in bold and second best are underlined.

Sequence	EpO	EpO+	AGS [9]	MOA [4]	LSMO [8]	PDB [5]	STP [1]	ARP [3]	LVO [7]	MpNet [6]	FSeg [2]
blackswan	0.7816	0.8418	0.7958	0.9198	**0.9296**	0.9080	0.6741	0.8815	0.7426	0.5059	0.8109
bmx-trees	0.4129	0.4898	0.51905	0.4644	0.5023	0.4634	0.6446	0.5017	0.5011	**0.5224**	0.4353
breakdance	0.6678	**0.8221**	0.6078	0.3638	0.4595	0.5912	0.7783	0.7628	0.3715	0.5276	0.5109
camel	0.8824	**0.9207**	0.8578	0.8290	0.8863	0.8240	0.8328	0.9016	0.8816	0.7847	0.8355
car-roundabout	**0.9251**	0.8903	0.8439	0.7685	0.8595	0.8522	0.8523	0.8164	0.8836	0.7978	0.9015
car-shadow	0.8893	0.8855	0.9141	**0.9209**	0.8817	0.9126	0.7316	0.7287	0.9199	0.8362	0.8960
cows	0.8904	0.9109	0.9217	**0.9438**	0.9099	0.9181	0.9049	0.9081	0.9023	0.8382	0.8681
dance-twirl	0.6392	0.8267	0.7894	0.6737	**0.8309**	0.6603	0.8154	0.7988	0.8089	0.5974	0.7042
dog	0.9008	0.8971	0.9352	**0.9393**	0.9292	0.9232	0.8380	0.7169	0.8870	0.8188	0.8891
drift-chicane	0.7038	0.6780	0.6910	0.7543	0.6899	0.6014	0.4538	**0.7932**	0.6289	0.6751	0.5985
drift-straight	0.7318	0.8253	0.8935	**0.9105**	0.8233	0.8571	0.7078	0.7046	0.8472	0.7145	0.8106
goat	0.8325	0.8407	0.8474	**0.8832**	0.8446	0.8374	0.8515	0.7770	0.8226	0.7517	0.8308
horsejump-high	0.8120	0.8400	0.8398	0.8776	0.8621	0.8574	**0.8888**	0.8358	0.8235	0.8319	0.6493
kite-surf	0.5202	0.6549	0.6880	**0.6907**	0.5005	0.6745	0.4361	0.5931	0.6461	0.5381	0.3897
libby	0.6035	0.7338	0.6583	**0.8307**	0.7812	0.7307	0.8233	0.6573	0.6932	0.6500	0.5847
motocross-jump	0.7714	0.8535	0.8213	0.8541	0.8228	**0.8547**	0.5040	0.8257	0.8052	0.7036	0.7749
paragliding-launch	0.6010	0.6238	0.6295	0.6404	0.6352	0.6337	**0.6423**	0.6011	0.6246	0.6380	0.5699
parkour	0.8151	0.8735	**0.9063**	0.9053	0.8926	0.9007	0.8907	0.8248	0.8489	0.7720	0.7580
scooter-black	0.8288	**0.8301**	0.7509	0.5457	0.7083	0.6893	0.7880	0.7436	0.7182	0.7206	0.6847
soapbox	0.8224	0.8825	0.7556	0.7170	0.8796	0.7312	**0.9071**	0.8433	0.8114	0.7636	0.6256
Mean IoU	0.7516	**0.8061**	0.7970	0.7720	0.7820	0.7711	0.7483	0.7608	0.7585	0.6995	0.7065

Table 1: Comparison of our motion (EpO) and fusion network (EpO+), with state-of-the-art on validation set of DAVIS-2016 with intersection over union $\mathcal{J}$ on each sequence. The mean IoU of ARP and STP differs from the paper, because here we are reporting the results on validation set only. The best scores are highlighted in bold and second best are underlined.

Figure 6: Failure Cases of our EpO Network. In the Left sequence, the trajectories are short due to occlusion that results in low epipolar distance. The optical flow in the right sequence was too noisy because of the smoke in some of the frames. This results in an under-segmentation of the foreground object.

References

[1] Y.-T. Hu, J.-B. Huang, and A. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proc. ECCV*, 2018. 3, 5, 6

[2] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. *arXiv preprint arXiv:1701.05384*, 2(3):6, 2017. 5, 6

[3] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2017. 5, 6

[4] M. Siam, C. Jiang, S. W. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jägersand. Video segmentation using teacher-student adaptation in a human robot interaction (HRI) setting. *CoRR*, abs/1810.07733, 2018. 3, 5, 6

[5] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018. 3, 4, 5, 6

[6] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 531–539. IEEE, 2017. 2, 5, 6

[7] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. pages 4491–4500, 2017. 3, 4, 5, 6

[8] P. Tokmakov, C. Schmid, and K. Alahari. Learning to segment moving objects. *International Journal of Computer Vision*, 127(3):282–301, 2019. 3, 5, 6

[9] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. Hoi, and H. Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3064–3074, 2019. 3, 4, 5, 6