EXOCENTRIC TO EGOCENTRIC IMAGE GENERATION VIA PARALLEL GENERATIVE ADVERSARIAL NETWORK

Gaowen Liu[1], *Hao Tang*[1,2], *Hugo Latapie*[3], *Yan Yan*[1*]

[1]Department of Computer Science, Texas State University, USA
[2]DISI, University of Trento, Italy
[3]Chief Technology & Architecture Office, Cisco, USA

ABSTRACT

Cross-view image generation has been recently proposed to generate images of one view from another dramatically different view. In this paper we investigate exocentric (third-person) view to egocentric (first-person) view image generation. This is a challenging task since egocentric view sometimes is remarkably different from exocentric view. Thus, transforming the appearances across the two views is a non-trivial task. Particularly, we propose a novel Parallel Generative Adversarial Network (P-GAN) with a novel cross-cycle loss to learn the shared information for generating egocentric images from exocentric view. We also incorporate a novel contextual feature loss in the learning procedure to capture the contextual information in images. Extensive experiments on Exo-Ego datasets [1] show that our model outperforms the state-of-the-art approaches.

Index Terms— Egocentric, Exocentric, Cross-View Image Generation, Parallel GAN

1. INTRODUCTION

Wearable cameras, also known as first-person cameras, nowadays are widely used in our daily lives since the appearance of low price but high quality wearable products such as GoPro cameras. Meanwhile, egocentric (first-person) vision is also becoming a critical research topic in the field. As we know, egocentric view have some unique properties other than exocentric (third-person) view. Traditional exocentric cameras usually give a wide and global view of the high-level appearances happened in a video. However, egocentric cameras can capture the objects and people at a much finer level of granularity. In the early egocentric vision studies, researchers [2] found that people perform different activities or interacting with objects from a first-person egocentric perspective and seamlessly transfer knowledge between egocentric and exocentric perspective. Therefore, analyzing the relationship between egocentric and exocentric perspectives is an extremely useful and interesting topic for image and video understanding. However, there is few research to address this important problem in literature.

GANs [3] have been shown effectively in image generation tasks. Isola *et al.* [4] propose Pix2Pix adversarial learning framework on paired image generation, which is a supervised model and uses a conditional GAN framework to learn a translation function from input to output image domain. Zhu *et al.* [5] introduce cycle-GAN which develops cycle-consistency constraint to deal with unpaired image generation. However, these existing works consider an application scenario in which the objects and the scenes have a large degree of overlapping in appearance and view. Recently, some works investigate cross-view image generation problems to generate a novel scene which is drastically different from a given scene image. This is a more challenging task since different views share little overlap information. To tackle this problem, Krishna *et al.* [6] propose X-Fork and X-Seq GAN-based architecture using an extra semantic segmentation map to facilitate the generation. Hao *et al.* [7] propose a multi-channel attention selection module within a GAN framework for cross-view image generation. However, these methods are not able to generate satisfactory results due to the drastically differences between exocentric and egocentric views.

To bridge egocentric and exocentric analaysis, in this paper we propose a novel Parallel GAN (P-GAN) to generate exocentric images from egocentric view. P-GAN framework is able to automatically learn the shared information between two parallel generation tasks via a novel cross-cycle loss and hard-sharing of network layers. We also utilize a novel contextual loss in our objective function to capture texture information over the entire images. To the best of our knowledge, we are the first to attempt to incorporate a parallel generative network for exocentric to egocentric image translation. Our proposed P-GAN is related to CoGAN [8] and DualGAN [9]. However, CoGAN and DualGAN have limited ability in generating image pairs with dramatically different viewpoints. As shown in Fig. 1, our architecture is designed in a bi-directional parallel fashion to discover the shared information between egocentric and exocentric images. Two parallel GANs are trained simultaneously with hard-sharing

[*] Corresponding author.

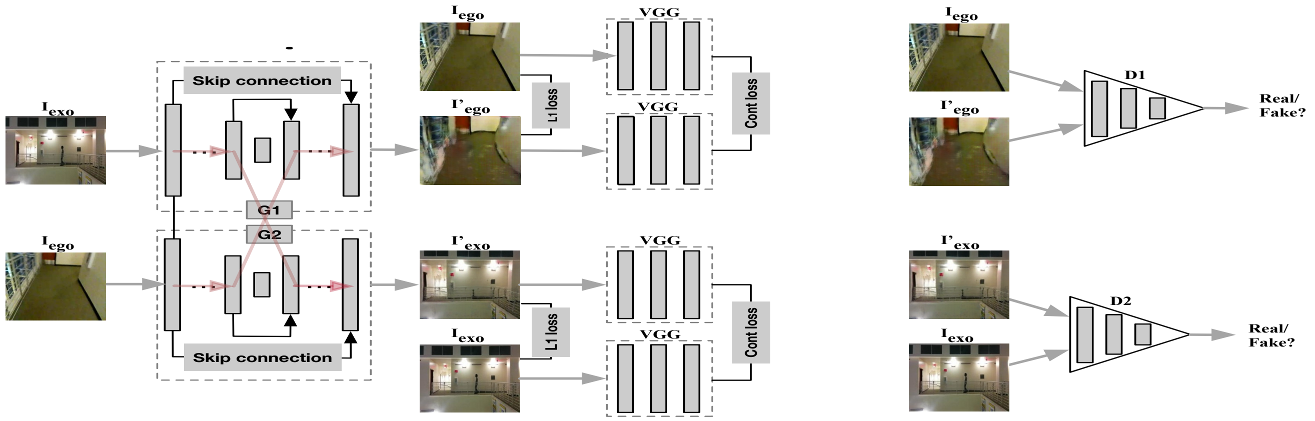

Fig. 1. The pipeline of our P-GAN model. It consists of two parallel generators G_1, G_2, and two discriminators D_1, D_2. The total loss contains pairs of L_1 loss, contextual loss and adversarial loss.

of certain layers.

In summary, our contributions can be highlighted as follows. (i) A novel P-GAN is proposed to learn the shared information between different views simultaneously via a novel cross-cycle loss. (ii) A novel contextual feature loss is incorporated in the training to capture the contextual information. (iii) Experiments on Exo-Ego dataset show the effectiveness of our hard-sharing of network layers in multi-directional parallel generative models.

2. PARALLEL GENERATIVE ADVERSARIAL NETWORK

2.1. Network Architecture

Cross-view exocentric to egocentric image synthesis is a challenging task, because these two views have little overlapping in image appearance. Most existing works on cross-view image synthesis are based on GANs. A traditional GAN consists of a generative model and a discriminative model. The objective of the generative model is to synthesize images resembling real images, while the objective of the discriminative model is to distinguish real images from synthesized ones. Both the generative and discriminative models are realized as multi-layer perceptrons. Since there will be some shared high-level concept information in a pair of corresponding images between exocentric and egocentric views, we propose a P-GAN with two GANs in parallel which is able to learn the shared high-level semantic information among different views. Fig. 1 shows our framework which contains two generators and two discriminators. A set number of layers from two generators are shared across P-GAN. We force the first three layers of two generators to have the identical structure and share the weights, and the rest layers are task-specific. The experiments show that sharing three layers of generators yield the best performance.

Particularly, we employ U-Net [10] as the architecture of our generators G_1 and G_2. We impose skip connection strategy from down-sampling path to up-sampling path to avoid vanishing gradient problem. To learn the shared information between exocentric and egocentric view, we perform hard-sharing in the first three layers of down-sampling path. We adopt PatchGAN [4] for the discriminator D_1 and D_2. The feature maps for contextual loss are extracted by the VGG-19 network pretrained on ImageNet.

2.2. Overall Optimization Objective

The training objective can be decomposed into four main components which are contextual loss, adversarial loss, cross-cycle loss and reconstruction loss.

Contextual loss. Different from the commonly used $\mathcal{L}1$ loss function which compares pixels at the same spatial coordinates between the generated image and the target image, we incorporate contextual loss in our P-GAN learning framework. The key idea is to measure similarity between images at the feature level.

Given a generated fake image I'_{ego} and a real image I_{ego} in egocentric view, we obtain a list of VGG-19 [11] features as $I_{ego} = \{I_i\}$ and $I'_{ego} = \{I'_j\}$, where $I_i = \psi^i(I_{ego})$, $I'_j = \psi^j(I_{ego})$, ψ means VGG-19 feature. i, j are i-th and j-th layer in the network ψ. The similarity between the generated image I'_{ego} and the real image I_{ego} in egocentric view can be defined as follows,

$$\mathcal{S}_{I_i,I'_j} = exp\left(1 - \frac{1-d_{ij}}{min_k d_{ik} + \zeta}\right)/h \qquad (1)$$

where d_{ij} is the cosine distance between I_{ego} and I'_{ego}. We define $\zeta = 1e^{-5}$, h = 0.5 in our experiments. The similarity can be normalized as,

$$\bar{S}_{ij} = \frac{\mathcal{S}_{I_i,I'_j}}{\sum_k \mathcal{S}_{I_i,I'_k}} \qquad (2)$$

Then the contextual loss is formulated as follows,

$$\mathcal{L}_{cont}(I_i, I'_j) = \frac{1}{max(\mid I_{ego} \mid, \mid I_{exo} \mid)} \sum_j max\bar{S}_{ij} \qquad (3)$$

where $|\cdot|$ denotes the numbers of feature maps.

Cross-cycle loss. As shown in Fig. 1, we employ U-Net [10] as our generators G_1 and G_2. Each U-Net contains a down-sampling encoder EN which is a feature contracting path,

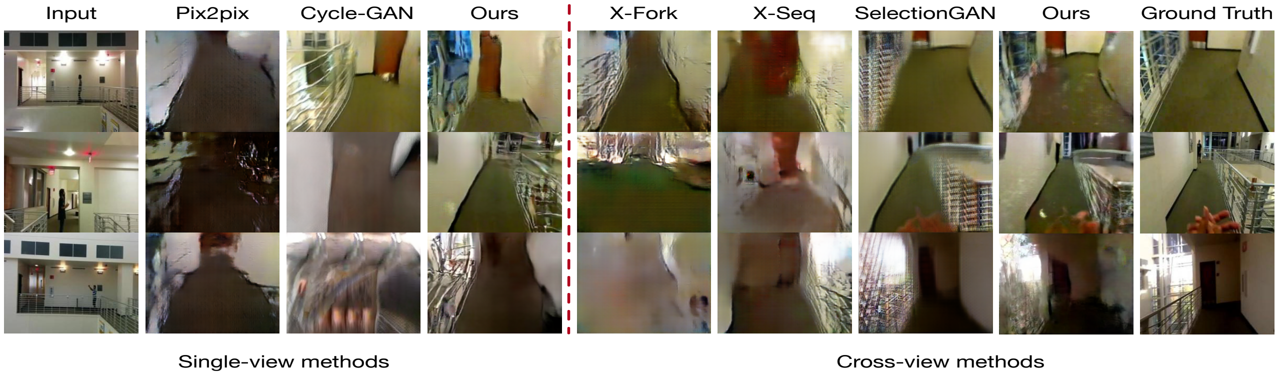

Fig. 2. Results generated by different methods on Side2Ego dataset. These samples were randomly selected for visualization purposes. Columns from left to right are: Input, Pix2pix [4], Cycle-GAN [5], Ours, X-fork [6], X-Seq [6], Selection-GAN [7], Ours + Segmentation map, Ground Truth.

and an up-sampling decoder DE which is a feature expanding path. Inspired by the U-net properties, we design a novel cross-cycle loss as follows,

$$\begin{aligned}\mathcal{L}_X(G_1,G_2) =& \mathbb{E}_{I_{exo},I'_{exo}}\left[\|I_{exo} - DE_2(EN_1(I_{exo}))\|_1\right] + \\ & \lambda_1 \mathbb{E}_{I_{ego},I'_{ego}}\left[\|I_{ego} - DE_1(EN_2(I_{ego}))\|_1\right]\end{aligned} \tag{4}$$

Adversarial loss. Recent works [3, 12, 13, 14] have shown that one can learn a mapping function by tuning a generator and a discriminator in an adversarial way. Assuming we target to learn a mapping $G\colon I_{exo} \rightarrow I_{ego}$ from input exocentric image I_{exo} to output egocentric image I_{ego}. The generator G is trained to produce outputs to fool the discriminator D. The adversarial loss can be expressed as,

$$\begin{aligned}\mathcal{L}_{GAN_1}(G_1,D_1) =& \mathbb{E}_{I_{exo},I_{ego}}\left[\log D_1(I_{exo},I_{ego})\right] + \\ & \mathbb{E}_{I_{exo},I'_{ego}}\left[\log(1 - D_1(I_{exo},G_1(I_{exo})))\right]\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{GAN_2}(G_2,D_2) =& \mathbb{E}_{I_{ego},I_{exo}}\left[\log D_2(I_{ego},I_{exo})\right] + \\ & \mathbb{E}_{I_{ego},I'_{exo}}\left[\log(1 - D_2(I_{ego},G_2(I_{ego})))\right]\end{aligned} \tag{6}$$

The adversarial loss is the sum of Eqn. 5 and Eqn. 6.

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN_1}(G_1,D_1) + \lambda_2\mathcal{L}_{GAN_2}(G_2,D_2) \tag{7}$$

Reconstruction loss. The task of the generator is to reconstruct an image as close as the target image. We use $\mathcal{L}1$ distance in the reconstruction loss,

$$\begin{aligned}\mathcal{L}_{re}(G_1,G_2) =& \mathbb{E}_{I_{exo},I'_{ego}}\left[\|I_{ego} - DE_1(EN_1(I_{exo}))\|_1\right] + \\ & \lambda_3 \mathbb{E}_{I_{ego},I'_{exo}}\left[\|I_{exo} - DE_2(EN_2(I_{ego}))\|_1\right]\end{aligned} \tag{8}$$

Overall loss. The total optimization loss is a weighted sum of the above losses. Generators G_1, G_2 and discriminators D_1, D_2 are trained in an end-to-end fashion to optimize the following objective function,

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda_4\mathcal{L}_X + \lambda_5\mathcal{L}_{re} + \lambda_6\mathcal{L}_{cont} \tag{9}$$

where λ_i's are the regularization parameters.

3. EXPERIMENTAL RESULTS

Datasets. To explore the effectiveness of our proposed P-GAN model, we compare our model with the state-of-the-art methods on Exo-Ego dataset [1] which contains two different viewpoint subsets (Side2Ego and Top2Ego). This dataset is challenging due to two reasons. First, it contains dramatically different indoor and outdoor scenes. Second, the dataset is collected simultaneously by an exocentric camera (side and top view) and an egocentric body-worn wearable camera. It includes a huge amount of blurred images for egocentric view. For Side2Ego subset, there are 26764 pairs of images for training and 13788 pairs for testing. For Top2Ego subset, there are 28408 pairs for training and 14064 pairs for testing. All images are in high-resolution 1280 × 720 pixels.

Experimental Setup. We compare our P-GAN with both single-view image generation methods [4, 5] and cross-view image generation methods [6, 7]. We adopt the same experimental setup as in [4, 6, 7]. All images are scaled to 256 × 256 pixels. We enable image flipping and random crops for data augmentation. To compute contextual loss, we use the VGG-19 network to extract image feature maps pretrained on ImageNet as the same as [15, 16]. We train 35 epochs with the batch size of 4. In our experiments, we set $\lambda_1 = 10$, $\lambda_2 = 10$, $\lambda_3 = 100$, $\lambda_4 = 10$, $\lambda_5 = 1$, $\lambda_6 = 1$ in Eqn. (4), (7), (8), (9) respectively. The proposed P-GAN is implemented using Pytorch. The state-of-the-art cross-view generation methods, *i.e.*, X-fork [6], X-Seq [6] and Selection-GAN [7] utilize segmentation map to facilitate target view image generation. To compare with these cross-view methods, we adopt RefineNet [17, 18, 19] to generate segmentation maps on Side2Ego and Top2Ego subsets as in [6, 7]. The generated segmentation maps are used as the conditional input of G_1 and G_2. To optimize the proposed P-GAN, we follow the optimization as in [3], we perform our experiments on Nvidia Geforce GTX 1080 Ti GPU with 11 GB memory.

Evaluation Metrics. We apply metrics such as top-k prediction accuracy and KL score for evaluations as in [7, 6]. We also employ pixel-level similarity metrics, *i.e.*, Structural-

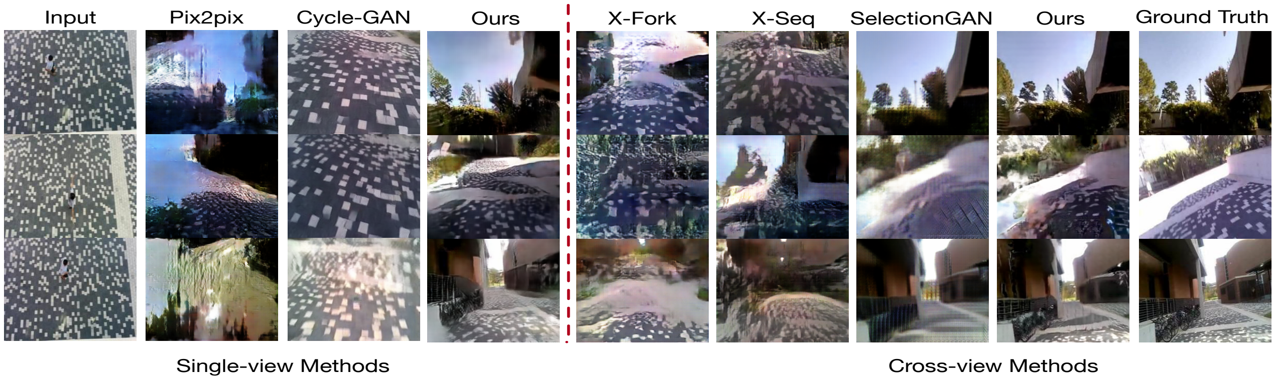

Fig. 3. Results generated by different methods on Top2Ego dataset. These samples were randomly selected for visualization purposes. Columns from left to right are: Input, Pix2pix [4], Cycle-GAN [5], Ours, X-fork [6], X-Seq [6], Selection-GAN [7], Ours + Segmentation map, Ground Truth.

Table 1. SSIM, PSNR, Sharpness Difference (SD), KL score (KL) and Accuracy of different single-view image generation methods. For these metrics except KL score, higher is better.

Dataset	Method	SSIM	PSNR	SD	KL	Top-1 Accuracy (%)		Top-5 Accuracy (%)	
Top2ego	Pix2pix[4]	0.2514	15.0532	18.1002	62.74 ± 1.78	1.24	1.22	4.21	4.35
	Cycle [5]	0.2806	15.5486	18.5678	52.09 ± 1.69	**2.10**	0.99	5.37	2.72
	Ours	**0.3098**	**17.0236**	**18.6043**	**31.46 ± 1.74**	1.81	**5.90**	**5.74**	**9.17**
Side2ego	Pix2pix [4]	0.3946	16.0716	19.8664	75.27 ± 2.01	3.20	5.18	8.41	13.30
	Cycle [5]	0.4017	15.9678	19.7533	62.41 ± 2.41	4.18	7.60	15.62	21.45
	Ours	**0.4908**	**17.995**1	**20.6521**	**13.92 ± 1.53**	**16.21**	**30.80**	**27.57**	**46.51**

Table 2. SSIM, PSNR, Sharpness Difference (SD), KL score (KL) and Accuracy of different cross-view image generation methods. For these metrics except KL score, higher is better.

Dataset	Method	SSIM	PSNR	SD	KL	Top-1 Accuracy (%)		Top-5 Accuracy (%)	
Top2ego	X-Fork [6]	0.2952	15.8849	18.7349	63.96±1.74	0.8	1.22	3.16	4.08
	X-Seq [6]	0.3522	16.9439	19.2733	54.91 ± 1.81	1.07	1.77	4.29	6.94
	SelectionGan [7]	0.5047	22.0244	19.1976	**10.07 ± 1.29**	8.85	16.55	24.32	33.90
	Ours	**0.5287**	**22.2891**	**19.2389**	12.07 ± 1.69	**9.76**	**29.67**	**24.80**	**51.79**
Side2ego	X-Fork [6]	0.4499	17.0743	20.4443	51.20 ± 1.94	4.49	9.76	11.63	19.44
	X-Seq [6]	0.4763	17.1462	20.7468	45.10 ± 1.95	6.51	12.70	11.97	19.36
	SelectionGan [7]	0.5128	18.3021	20.9426	**7.26 ± 1.27**	20.84	37.49	42.51	65.22
	Ours	**0.5205**	**19.4521**	**20.9684**	25.25 ± 1.88	**20.96**	**39.08**	**42.58**	**66.00**

Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD). These metrics evaluate the generated images in a high-level feature space.

Quantitative Results. The quantitative results are presented in Table 1 and Table 2. We observe that our P-GAN network achieves better results than state-of-the-art methods in most cases. Compared with single-view image generation methods, our P-GAN outperforms Pix2pix [4] and Cycle-GAN [5]. On the other hand, we also achieve better results than other cross-view image generation methods in most metrics while incorporating semantic segmentation map as in the Selection-GAN [7].

Qualitative Results. Qualitative results are shown in Fig. 2 and Fig. 3. The results confirm that the proposed P-GAN network has the ability to transfer the image representations from exocentric to egocentric view, *i.e.,* objects are in the correct positions for generated egocentric images. Results show that egocentric images generated by P-GAN are visually much better compared with other baselines.

4. CONCLUSIONS

In this paper we introduce a novel P-GAN which is able to learn shared information between cross-view images via a novel cross-cycle loss for a challenging exocentric to egocentric view image generation task. The proposed method utilizes both pixel level and contextual level information. Moreover, we incorporate a novel contextual feature loss to capture the contextual information in images. Experimental results demonstrate that the hard-sharing of network layers in multi-directional parallel generative models can be used to increase the performance of cross-view image generation.

5. ACKNOWLEDGEMENT

This research was partially supported by a gift donation from Cisco Inc. and NSF NeTS-1909185 and NSF CSR-1908658. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

6. REFERENCES

[1] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji, "From third person to first person: Dataset and baselines for synthesis and retrieval," in *CVPR*, 2019.

[2] T. Kanade and M. Hebert, "First-person vision," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2442–2453, 2012.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *ICCV*, 2017.

[6] Krishna Regmi and Ali Borji, "Cross-view image synthesis using conditional gans," in *CVPR*, 2018.

[7] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J. Corso, and Yan Yan, "Multi-channel attention selection gan with cascaded semantic guidancefor cross-view image translation," in *CVPR*, 2019.

[8] Ming-Yu Liu and Oncel Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016.

[9] Yi Zili, Zhang Hao, Tan Ping, and Gong Minglun, "Dualgan: Unsupervised dual learning for image-to-image translation," in *ICCV*.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *CVPR*, 2015.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen, "Improved techniques for training gans," in *NIPS*, 2016.

[13] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou, "An improved evaluation framework for generative adversarial networks," *arXiv preprint arXiv:1803.07474*, 2018.

[14] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li, "Mode regularized generative adversarial networks," in *ICLR*, 2017.

[15] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," *arXiv preprint arXiv:1803.02077*, 2018.

[16] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor, "Learning to maintain natural image statistics," *arXiv preprint arXiv:1803.04626*, 2018.

[17] Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[18] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2017.

[19] Jifeng Dai, Kaiming He, and Jian Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *CVPR*, 2016.