MBA-RAINGAN: A MULTI-BRANCH ATTENTION GENERATIVE ADVERSARIAL NETWORK FOR MIXTURE OF RAIN REMOVAL

Yiyang Shen[⋆]*, Yidan Feng*[⋆]*, Weiming Wang*[†]*, Dong Liang*[⋆]*, Jing Qin*[‡]*, Haoran Xie*[§]*, Mingqiang Wei*[⋆]

[⋆]Nanjing University of Aeronautics and Astronautics [†]Hong Kong Metropolitan University
[‡]Hong Kong Polytechnic University [§]Lingnan University

ABSTRACT

Rain severely degrades the visibility of scene objects, especially when images are captured through the glass under rainy weather. We observe three intriguing phenomena: 1) rain is a mixture of *raindrops*, *rain streaks* and *rainy haze*; 2) the depth from the camera determines the degree of object visibility, where objects nearby and far away are visually blocked by rain streaks and rainy haze, respectively; and 3) raindrops on the glass randomly affect the object visibility of the whole image space. However, existing solutions and benchmark datasets lack full consideration of the mixture of rain (MOR). In this paper, we originally consider that the overall object visibility is determined by MOR, and enrich the RainCityscapes by considering real-world raindrops to construct the MOR dataset, named RainCityscapes++. To solve the practical rain removal problem arisen from MOR, we formulate a new rain imaging model and propose a multi-branch attention generative adversarial network (MBA-RainGAN). Extensive experiments show clear improvements of our approach over SOTAs on RainCityscapes++.

Index Terms— MBA-RainGAN, Image deraining, Bidirectional coordinate attention, Mixture of rain, GAN

1. INTRODUCTION

Rain is one of the most common dynamic weather phenomena, which significantly degrades the visibility and contrast of the image. Existing image deraining methods [7, 8, 9, 10, 11] mainly focus on a single type of rainwater artifacts, i.e., only rain streaks or raindrops, regardless of the fact that rainwater is transformable and can appear in various forms under different shooting conditions. Beyond previous image deraining wisdom, we investigate a more comprehensive rain removal problem by taking photographic conditions into account, which covers an outdoor camera lens without protection, indoor photographing through windows, and driver assistance systems behind the windshield. Images taken under these situations suffer from the mixture of rain (MOR), that is, the effect of rain from close to far appears as raindrops, rain streaks, and rainy haze. Based on this observation, we formulate a new rain imaging model by an additional consideration of raindrops [12], where we regard the raindrop location as a piece of important prior knowledge for dissolving the MOR problem. Accordingly, we construct an improved version of the RainCityscapes dataset, named RainCityscapes++, by composing real-world raindrop layers on the images affected by both rain streaks and rainy haze to reproduce realistic scenes affected by MOR.

To remove the entangled MOR effect, we develop a three-stage decomposition strategy: (1) We separate rain streaks and rainy haze by a low-pass filter based on the distinctness in the frequency domain. (2) We learn separate attention maps for each form of rainwater artifacts using a multi-branch structure. These attention maps are extracted by collateral recurrent networks in a coarse-to-fine manner, which progressively guides the final image decomposition in the contextual autoencoder. (3) We develop an attentive discriminator for the image-level constraint to ensure the fidelity of output.

2. FORMULATION AND DATASET

2.1. Rainy Image Formulation

Formulation of MOR. We consider that a rainy image is composed of a clean background image and a mixture of three layers, i.e., the raindrop layer, the rainy haze layer, and the rain streak layer. In contrast to existing rain imaging models, we formulate the rainy image $I(x)$ as:

$$\begin{aligned} I(x) =&(1 - M_d(x)) \cdot [B(x)(1 - S(x) - A(x)) + S(x) \\ &+ A_0 A(x)] + D(x) \end{aligned} \tag{1}$$

where $M_d(x) \in \{0, 1\}$ indicates whether the pixel x is corrupted by raindrops (1 is Yes, and 0 is No), $B(x)$ denotes the clean background image with the clear scene radiance, $S(x) \in [0, 1]$, $A(x) \in [0, 1]$ and $D(x) \in [0, 1]$ are the rain streak layer, the rainy haze layer and the raindrop layer re-

This work was supported by the National Natural Science Foundation of China (No. 62172218) and the HKMU 2020/2021 S&T School Research Fund (R5091), and the Direct Grant (DR22A2) and the Faculty Research Grants (DB22A5 and DB21A9) of Lingnan University, Hong Kong. Corresponding authors: H. Xie (hrxie2@gmail.com) and M. Wei (mqwei@nuaa.edu.cn).

(a) Input (b) MSPFN [1] (c) DGCN [2] (d) Eigen's [3] (e) SPA-Net [4] (f) AttentGAN [5] (g) DGNL-Net [6] (h) Ours

Fig. 1: Deraining results on a real-world MOR image.

spectively, and A_0 is a global constant following [13]. Based on the MOR model, we construct a new dataset called RainCityscapes++, which contains three forms of rainwater.

2.2. RainCityscapes++ Dataset

Dataset Generation. We have collected *14783* real outdoor photos of raindrops that are randomly distributed on the glass as the *cover layer*, and *8580* images from the training and validation sets of RainCityscapes [12] as our *background layer*. First, to simulate the scene of raindrops that randomly scatter on the glass, an overlay model is designed to superimpose the cover layer on the background layer:

$$C_1 = \begin{cases} \frac{A \times B}{128}, A \leq 128 \\ 255 - \frac{A_t \times B_t}{128}, A > 128 \end{cases} \tag{2}$$

where A is the background layer and B is the cover layer. A_t and B_t represent the anti-phase of A and B, respectively. Using Eq. 2, the composition layer C_1 is obtained but mingled with unreal occlusion brought by the cover layer. To solve visual occlusion, we evolve the overlay mode into the highlight mode to enhance the color contrast between the two layers:

$$C_2 = \begin{cases} \frac{A \times B}{128}, B \leq 128 \\ 255 - \frac{A_t \times B_t}{128}, B > 128 \end{cases} \tag{3}$$

Furthermore, to approximate the real scene, we further emphasize the background layer by increasing the transparency of the cover layer:

$$C_3 = t \times A + (1 - t) \times B \tag{4}$$

where t denotes the transparency. Based on Eq. 3 and Eq. 4, the final blending model can be expressed as:

$$C_4 = \begin{cases} \frac{t \times A \times (1-t) \times B}{128}, B \leq 128 \\ 255 - \frac{t \times A_t \times (1-t) \times B_t}{128}, B > 128 \end{cases} \tag{5}$$

After highlighting and adjusting the transparency, we obtain the rainy images C_4 that form our RainCityscapes++.

3. MBA-RAINGAN

Inspired by the success of the attention mechanism [5, 14, 4], we propose the multi-branch attention generative adversarial network (MBA-RainGAN) for the removal of an entangled mixture of different rainwater forms. The overall architecture is illustrated in Figure 2. In the following, we elaborate our generative network and discriminative network to demonstrate how we dissolve the intractable MOR problem.

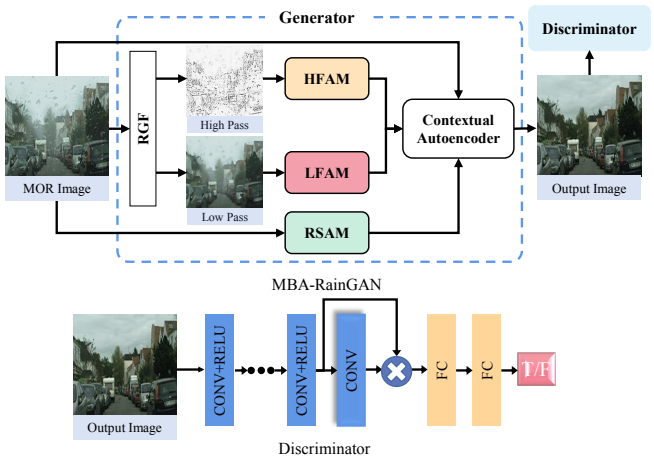

Fig. 2: Overall architecture of the multi-branch attention generative adversarial network (MBA-RainGAN).

3.1. Generative Network

To remove the mixture of rain (MOR) (see Figure 2), we develop a three-stage decomposition strategy in the generative network. First, a scale-aware Rolling Guidance Filter (RGF) [15] is used to separate rain streaks and rainy haze from the rainy image, in order to ease the burden of multiple attention learning. Then, the separated rain streak layer and rainy haze layer are fed into the high frequency attentive module (HFAM) and low frequency attentive module (LFAM) respectively, while the raindrop attention map is directly learned from the input image using the raindrop spatial attentive module (RSAM). Lastly, a contextual autoencoder is guided by the three attention maps to progressively remove MOR.

High/Low Frequency Attentive Module. To predict the rain streak attention map and the rainy haze attention map, the high frequency attentive module (HFAM) and the low frequency attentive module (LFAM) process the high pass component and the low pass component of the RGF output, respectively. As shown in Figure 3, HFAM and LFAM share the same network structure, which is composed of two 3×3

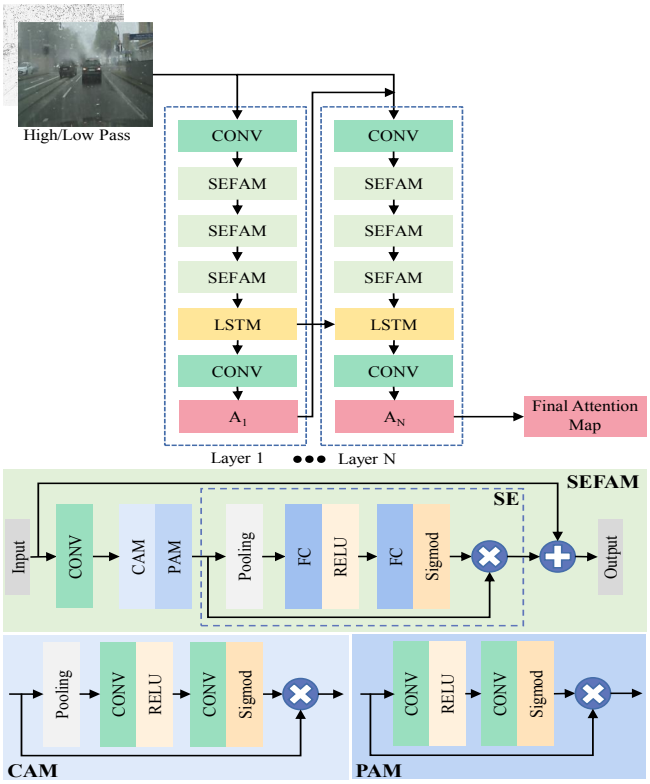

Fig. 3: Architecture of our HFAM and LFAM, A_N represents the final attention map. Note that we set N as 2.

convolutional layers, three SE Feature Attention Module (SEFAM) and a convolutional LSTM unit [16].

Through the observation of various real MOR images, we incorporate uneven rain streak and rainy haze distribution into consideration. Motivated by [17], we treat different features and pixel regions unequally and propose an SE Feature Attention Module (SEFAM) to produce additional flexibility for dealing with uneven rain streak and rainy haze distribution. As shown in Figure 3, we first adopt a Channel Attention Module (CAM) [18] and a Pixel Attention Module (PAM) [17] to generate different weights for channel-wise and pixel-wise features, respectively. Then, because SE can model a correlation between different feature channels, we develop the SEFAM to intensify the feature channel that has more context information by giving a larger weight. After that, the feature maps from SEFAM are fed into an LSTM unit and a convolutional layer to generate the 2D attention map.

Raindrop Spatial Attentive Module. As for raindrops, due to the randomness of their distributions and the complex reflection effect in the contaminated areas, we build a raindrop spatial attentive module (RSAM) to learn the raindrop attention map. Motivated by [19, 20, 4], we apply the two-round four-directional IRNN to accumulate the global contextual information, which substantially enlarges the receptive field for extracting raindrop features. As shown in Figure

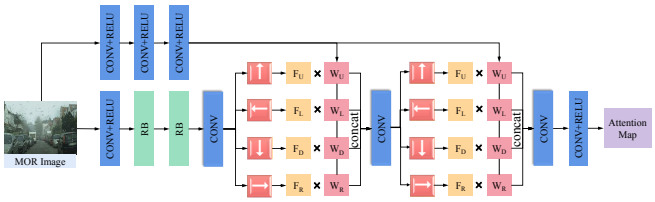

Fig. 4: Architecture of our RSAM.

4, for each position at the input feature map, four-directional (up, left, down, right) recurrent convolutional operations are performed to collect the horizontal and vertical neighborhood information between two-round IRNN structures:

$$f_{i,j} \leftarrow max(\alpha_{dir} f_{i,j-1} + f_{i,j}, 0) \tag{6}$$

where $f_{i,j}$ denotes the feature at the location (i, j) and α represents the weight parameter in the recurrent convolution layer for each direction. Besides, we add another branch to capture the spatial contextual information to selectively highlight the projected raindrop features.

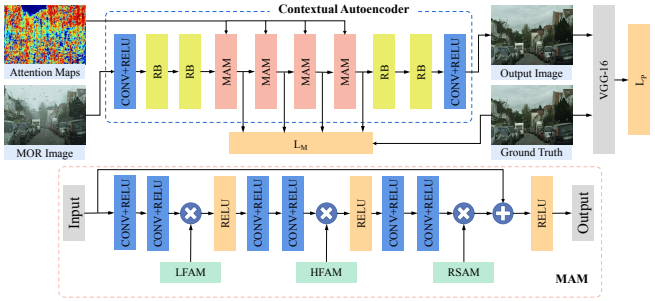

Fig. 5: Architecture of our contextual autoencoder and MAM.

Contextual Autoencoder. Our contextual autoencoder generates clean images that are free of MOR. As shown in Figure 5, the first convolutional layer extracts the image features, followed by two residual blocks to obtain deeper embeddings. The key part of contextual autoencoder consists of four MAMs that reflect the complex entanglement of MOR.

To take advantage of these attention maps, we propose a multi-attentive module (MAM), where the prediction of network embeddings is progressively guided by different attention maps. The attention maps of rainy haze, rain streaks, and raindrops are sequentially added into MAM (see Figure 5), leading to two benefits: 1) it avoids the confusion caused by simultaneously introducing the information from multiple disparate attention maps; 2) according to the established formulation of MOR, it considers the superposition of MOR that rainy haze, rain streaks and raindrops happen in the bottom, medium and top layers, respectively. The i-th scale feature M_i extracted from each MAM and the same scale ground-truth T_i

are utilized in the multi-scale losses L_M as:

$$\mathbf{L_M} = \sum_{i=1}^{N} \lambda_i \mathbf{L_{MSE}}(M_i, T_i) \tag{7}$$

where the values of λ are set to be 1.0, 0.8, 0.6, 0.4 for multi-scale features of 1, $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$ of the original size, respectively. Besides, we adopt the perceptual loss [21] to measure the global discrepancy between the prediction O and the GT T:

$$\mathbf{L_P} = \mathbf{L_{MSE}}(VGG(O), VGG(T)) \tag{8}$$

The VGG used in the perceptual loss is the $conv2,3$ layer of pre-trained VGG-16 network [22]. The overall loss for the generative network is formulated as:

$$\mathbf{L_{GN}} = \mathbf{L_P} + \mathbf{L_M} \tag{9}$$

3.2. Discriminative Network

The discriminative network accepts the output of generative network and checks if it looks like the ground truth. We constrain extracted features from interior layers of the discriminator, by minimizing the distance between the attention maps and the output after feeding them into another CNN:

$$\begin{aligned}\mathbf{L_{map}} =& \mathbf{L_{MSE}}(D_{map}(O), A_{HN}) + \mathbf{L_{MSE}}(D_{map}(O), A_{LN}) \\ &+ \mathbf{L_{MSE}}(D_{map}(O), A_{RN}) + \mathbf{L_{MSE}}(D_{map}(T), 0)\end{aligned} \tag{10}$$

where D_{map}, A_{HN}, A_{LN} and A_{RN} represent the 2D map produced by the discriminative network, HFAM, LFAM and RSAM respectively. The whole loss function of the discriminative network can be expressed as:

$$\mathbf{L_{DN}} = -\log(D(T)) - \log(1 - D(O)) + \gamma \mathbf{L_{map}} \tag{11}$$

where the balancing weight γ is set to be 0.1.

Method		RainCityscapes++	Runtime
Derain	Eigen's	17.18/0.71	1.43s/0.46s
	AttentGAN	24.03/0.84	1.62s/0.54s
	MSPFN	22.19/0.70	1.78s/0.67s
	DGCN	23.64/0.79	1.72s/0.58s
	MPRNet	22.87/0.76	1.32s/0.41s
	SPA-Net	23.12/0.82	1.53s/0.49s
Dehaze	MSBDN-DFF	22.37/0.79	1.46s/0.47s
	AOD-Net	19.22/0.76	1.21s/0.33s
	MSCNN	20.70/0.77	1.02s/0.26s
Derain+Dehaze	DAF-Net	25.21/0.85	1.46s/0.48s
	DGNL-Net	26.89/0.86	1.49s/0.52s
	Ours	**29.16/0.91**	**1.47s/0.47s**

Table 1: Averaged PNSR, SSIM and time (CPU/GPU) on RainCityscapes++ of SOTAs for removing MOR.

4. EXPERIMENTS

4.1. Experimental Dataset

We enrich the popular RainCityscapes by considering raindrops, named RainCityscapes++, which contains a total of 8580 MOR images. Then, we divide the total 8580 images in RainCityscapes++ into the training set (containing 7580 images) and the testing set (containing 1000 images). We also download 400 MOR photos from the Internet by searching with keywords should be "rain and fog photo".

Scheme	PSNR	SSIM
A	17.79	0.72
H+A	22.56	0.79
L+A	24.73	0.83
H+L+A	27.16	0.87
H+L+R+A	28.67	0.88
H+L+R+A+D	**29.16**	**0.91**

Table 2: The decomposition for ablation study. Note that A, H, L, R, and D denote autoencoder, HFAM, LFAM, RSAM, and discriminator respectively.

4.2. Experimental results

As shown in Figure 1 and Table 1, compared with other deraining methods, our method achieves the best performance in MOR removal on both synthetic images and real photos, and better preserves the color and structure of the image. We also conduct the ablation study on key components of MBA-RainGAN. From Table 2, we can observe that the complete multi-branch attention scheme achieves the best results both quantitatively and qualitatively. Furthermore, the corresponding attentive discriminator enhances the image details and renders our results more realistic for human perception.

5. CONCLUSION

In this work, we explore the visual effects of MOR and formulate the rain imaging model with rain streaks, rainy haze, and raindrops simultaneously. To cope with the MOR problem, we create a new dataset named RainCityscapes++, and propose a multi-branch attention generative adversarial network (termed an MBA-RainGAN), which develops a three-stage decomposition strategy to disentangle the MOR effects, i.e., the streak-aware decomposition with RGF, attention-level decomposition by the multi-branch attentive network, and the final image decomposition by the autoencoder. Our deraining result is finally validated by an attentive discriminator. Extensive experiments show that our method outperforms state-of-the-art deraining methods in complex rainy scenes, both quantitatively and qualitatively. In future, we will consider the MOR problem by unsupervised learning.

6. REFERENCES

[1] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang, "Multi-scale progressive fusion network for single image deraining," in *Proceedings of the IEEE/CVF confer-*

ence on computer vision and pattern recognition, 2020, pp. 8346–8355.

[2] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding, "Rain streak removal via dual graph convolutional network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–9.

[3] David Eigen, Dilip Krishnan, and Rob Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 633–640.

[4] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12270–12279.

[5] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2482–2491.

[6] Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng, "Single-image real-time rain removal based on depth-guided non-local features," *IEEE Transactions on Image Processing*, vol. 30, pp. 1759–1770, 2021.

[7] Shuangli Du, Yiguang Liu, Mao Ye, Zhenyu Xu, Jie Li, and Jianguo Liu, "Single image deraining via decorrelating the rain streaks and background scene in gradient domain," *Pattern Recognition*, vol. 79, pp. 303–317, 2018.

[8] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan, "Deep joint rain detection and removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1357–1366.

[9] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7007–7016.

[10] Xu Qin and Zhilin Wang, "Nasnet: A neuron attention stage-by-stage net for single image deraining," *arXiv preprint arXiv:1912.03151*, 2019.

[11] Sen Deng, Mingqiang Wei, Jun Wang, Yidan Feng, Luming Liang, Haoran Xie, Fu Lee Wang, and Meng Wang, "Detail-recovery image deraining via context aggregation networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14548–14557.

[12] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng, "Depth-attentional features for single-image rain removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8022–8031.

[13] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[14] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2020, pp. 11908–11915.

[15] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia, "Rolling guidance filter," in *European conference on computer vision*, 2014, pp. 815–830.

[16] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[17] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11908–11915.

[18] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[19] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2874–2883.

[20] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng, "Direction-aware spatial context features for shadow detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7454–7462.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016, pp. 694–711.

[22] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.