MixStyleFlow: Domain Generalization in Medical Image Segmentation using Normalizing Flows

Reza Safdari[1], Mohammad-Ali Nikouei Mahani[1], Mohamad Koohi-Moghadam[1*] and Kyongtae Tyler Bae[1*]

[1] The University of Hong Kong, Pokfulam, Hong Kong SAR, China
safdari@hku.hk, man.mahani@gmail.com,
koohi@hku.hk, baekt@hku.hk

Abstract. Despite the success of deep learning in medical image segmentation, domain shifts caused by variations in scanners and imaging protocols often degrade performance, limiting real-world clinical deployment. Domain generalization (DG) aims to address this issue by learning robust models that generalize well across different domains. While existing DG methods based on feature-space domain randomization have shown promise, they suffer from a limited and unordered search space of feature styles. In this work, we propose MixStyleFlow, a novel DG approach that utilizes normalizing flows to explicitly model the distribution of domain feature styles. By sampling domain feature styles from the learned normalizing flows and mixing them with original feature statistics along the feature channel dimension, our method effectively expands and diversifies domain features in a controllable manner. We evaluate MixStyleFlow on two medical segmentation tasks—prostate MRI and fundus imaging—demonstrating superior generalization performance on unseen target-domain data. Our results highlight the potential of normalizing flows for improving domain generalization in medical image segmentation, paving the way for more robust deep learning models capable of handling diverse clinical scenarios. The code is available at https://github.com/Reza-Safdari/MixStyleFlow.

Keywords: Domain Generalization, Medical Image Segmentation, Normalizing Flows, Feature Perturbation, Feature Style Augmentation.

1 Introduction

Deep learning has revolutionized medical image segmentation, enabling accurate delineation of anatomical structures and pathological regions in different imaging modalities like MRI, CT, and fundus photography [1]. However, a major challenge is the sensitivity of these models to domain shifts variations in image distributions due to different scanners, imaging protocols, patient populations, and acquisition settings [2, 3]. Models trained on specific domain data often perform poorly on unseen domains, limiting their generalizability and robustness in real-world clinical scenarios.

* Corresponding Author: Mohamad Koohi-Moghadam, Kyongtae Tyler Bae

Domain generalization (DG) addresses this challenge by learning models that can generalize well to unseen domains without requiring access to target-domain data during training [4, 5]. In medical imaging, achieving robust DG is crucial due to the heterogeneous nature of data arising from multi-institutional resources and diverse clinical practices [6]. Traditional approaches to tackle domain shifts include domain adaptation techniques, which require access to target-domain data during training [2]. However, collecting and annotating new data for each target domain is impractical and costly. Therefore, there is a pressing need for methodologies that enhance the generalization capability of segmentation models across diverse domains using only source-domain data.

Recent DG methods have explored feature-space domain randomization techniques aimed at learning domain-invariant representations by perturbing feature statistics during training [6]. Approaches such as MixStyle [7], MaxStyle [8] and their variants manipulate the style information in feature maps to simulate domain shifts, thereby encouraging the model to learn robust features. These methods operate on the assumption that content-preserving style transformations can be achieved by altering the statistics (e.g., mean and standard deviation) of the features. While these methods have shown promise results, using the statistics of source-domain data for feature perturbation may limit the search space and operate within a limited and unordered domain. Consequently, they may not fully capture the complex variations present in real-world domain shifts, potentially hindering the model's ability to generalize to unseen domains. To address these limitations, we propose MixStyleFlow, a novel domain generalization framework that utilizes the power of normalizing flows to explicitly model the distribution of domain feature styles.

2 Related Work

Feature-space domain randomization has emerged as a powerful approach for DG. Random Convolutions (RandConv) [9] introduce a domain randomization strategy by applying convolutional layers with randomly sampled weights to input images. This effectively alters local textures while preserving global shape information, encouraging models to rely more on structural cues rather than superficial textures. Dynamic Style Augmentation (DSU) [10] enhances DG by injecting Gaussian noise into feature statistics, generating diverse style variations beyond those present in the original data. This technique has shown particular effectiveness in medical image segmentation. Similarly, Treasure in Distribution (TriD) [11] expands feature perturbations by sampling statistics from a uniform distribution rather than relying solely on source-domain data. TriD further improves adaptability to domain shifts by employing a style-mixing strategy, blending original and augmented statistics along feature channels to promote domain-invariant representations.

Similarly, MixStyle [7] perturbs feature statistics by mixing instance-level mean and standard deviation from different training samples, implicitly simulating style variations across domains. MaxStyle [8] extends this idea by introducing adversarial style compositions, maximizing style diversity and generating more challenging training

examples. Beyond these methods, Exact Feature Distribution Matching (EFDM) [12] introduces a more precise feature perturbation approach by explicitly aligning the empirical cumulative distribution functions (eCDFs) of image features. Unlike approaches that rely only on first- and second-order statistics, EFDM performs exact histogram matching in feature space, ensuring a more faithful simulation of domain shifts. By offering a richer set of perturbations beyond traditional Gaussian-based transformations, EFDM significantly improves the generalization capability of segmentation models.

3 Method

3.1 Background on Normalizing Flows

Real Non-Volume Preserving (RealNVP) is a type of normalizing flow model used for density estimation and generative modeling [13]. It transforms a simple base distribution, such as Gaussian, into a complex target distribution through a sequence of invertible transformations [14]. The core idea behind RealNVP is the use of coupling layers, which enable efficient computation of both the forward transformation (sampling) and the inverse transformation (density evaluation). Each coupling layer splits the input x into two parts, x_1 and x_2, and applies an affine transformation to one part while keeping the other unchanged. The transformation is defined as:

$$\begin{aligned} x_1' &= x_1 \\ x_2' &= x_2 \odot \exp\big(s(x_1)\big) + t(x_1) \end{aligned} \tag{1}$$

where $s(x_1)$ and $t(x_1)$ are scale and translation functions, typically parameterized by neural networks. The symbol $\odot$ represents the elementwise (Hadamard) product. The inverse transformation, crucial for likelihood estimation, is easily computable as:

$$x_2 = \big(x_2' - t(x_1)\big) \odot \exp\big(-s(x_1)\big) \tag{2}$$

Since the transformation is invertible with a tractable Jacobian determinant, the log-likelihood of a sample x under the model is computed using the change of variables formula:

$$\log p_X(x) = \log p_Z\big(f(x)\big) + \sum_i \log \left|\det \frac{\partial f_i}{\partial x}\right| \tag{3}$$

where $f(x)$ maps the input to the latent space, $p_Z(z)$ is the probability density of the transformed sample under a simple prior distribution (e.g., Gaussian), and $\det \frac{\partial f_i}{\partial x}$ is the determinant of the Jacobian of the transformation. The loss function, derived from the negative log-likelihood (NLL) over N input samples $x_1, x_2, \ldots, x_N$, is given by:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p_X(x_i) \tag{4}$$

Minimizing this loss ensures the model effectively learns an expressive transformation that maps the data distribution to the Gaussian latent space while maintaining invertibility and efficient density estimation.

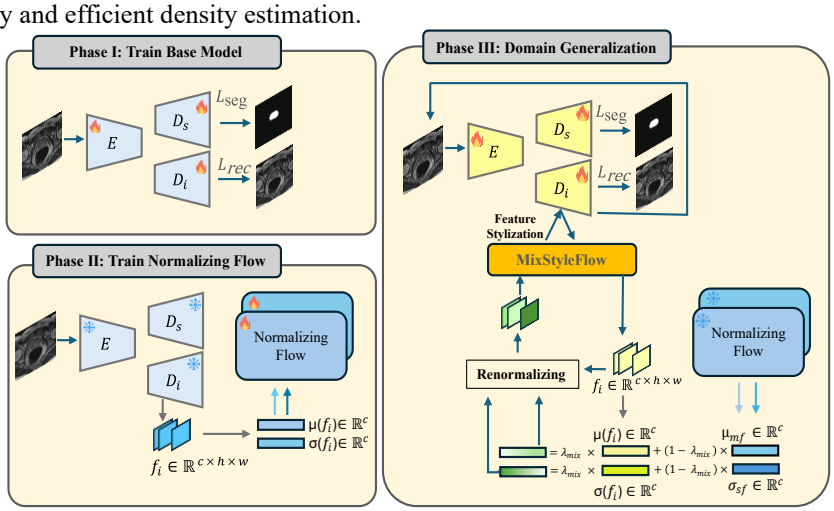

Fig. 1. The proposed MixStyleFlow framework. **Phase I** involves training the segmentation model on source datasets without domain generalization techniques. Subsequently, in **Phase II**, feature statistics (feature means and standard deviations) are extracted, and two normalizing flow models are trained to learn their distributions. In **Phase III**, the segmentation model is trained on source data and stylized images created through style feature augmentation at specific decoder layers. This augmentation is achieved by interpolating original feature statistics with in-domain and out-of-domain statistics sampled from the trained normalizing flow models.

3.2 Encoder with Dual-Decoder Architecture for Segmentation

We employed a dual-branch network presented in [8], consisting of one encoder (E) and two decoders. One decoder, called the segmentation decoder (D_s), is trained using segmentation loss to generate the semantic mask, while the other decoder, called the image decoder (D_i), is trained using reconstruction loss, allowing it to exploit both complementary image content features and task-specific shape features for the segmentation task **Fig. 1**. We also utilize the image decoder to apply style augmentation by randomizing the statistics (i.e., mean and standard deviation) of the feature maps in specific layers (the exact randomization method will be discussed in the next section). The reconstructed image from this decoder enables us to interpret and visualize the impact of the augmented features, ensuring that the randomization preserves their semantic and anatomical structures. For the segmentation loss L_{seg} , we use a combination of Binary Cross-Entropy (BCE) Loss L_{BCE} and Dice Loss L_{Dice}:

$$
\begin{aligned}
L_{\text{seg}} &= L_{\text{BCE}} + L_{\text{Dice}} \\
L_{\text{BCE}} &= -\sum_{i} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \\
L_{\text{Dice}} &= 1 - \frac{2 \sum_i y_i \hat{y}_i + \epsilon}{\sum_i y_i + \sum_i \hat{y}_i + \epsilon}
\end{aligned}
\tag{5}
$$

here yi and $\hat{y}_i$ represent the ground truth and predicted values, respectively, and ϵ is a small constant added for numerical stability. For the reconstruction loss, we use $l2$ norm as follows:

$$L_{\text{rec}} = \sum_i |I_i - \hat{I}_i|^2 \tag{6}$$

here I_i is the original input image and $\widehat{I_i}$ is the reconstructed image produced by the image decoder.

3.3 Proposed Method: MixStyleFlow for Domain Generalization

Our proposed MixStyleFlow framework consists of three phases that are presented in **Fig. 1**. In **Phase I**, we train the model on source datasets for segmentation using the architecture in Section 3.2, without domain generalization techniques. After training, in **Phase II**, we freeze its parameters and extract channel-wise feature statistics from the decoder. We then train two normalizing flow models (MuFlow for means, STDFlow for standard deviations) as described in Section 4.2, to learn their distributions. This enables (1) in-domain sampling to enrich the style space and (2) out-of-distribution sampling to expand it beyond the original dataset.

In **Phase III**, we train the segmentation architecture on the source data and stylized images generated by applying style feature augmentation at specific image decoder layers. Formally, let $f_i \in \mathbb{R}^{c \times h \times w}$ be c-dimensional feature maps extracted at a certain layer in the image decoder (D_i) for image x_i, where h and w represent the height and width of the feature maps, respectively. MixStyleFlow performs style augmentation for f_i by first normalizing it with its channel-wise means and standard deviations $\mu(f_i), \sigma(f_i) \in \mathbb{R}^c$: $f_i = \big(f_i - \mu(f_i)\big)/\sigma(f_i)$ and then transforming the normalized feature with a linear combination of its channel-wise statistics $\{\mu(f_i), \sigma(f_i)\}$ and sampled statistics from learned distributions $\{\mu_{mf},\ \sigma_{sf}\}$ as follows:

$$\begin{gathered} \text{MixStyleFlow}(f_i) = \gamma_{\text{mix}} \odot \overline{f_i} + \beta_{\text{mix}} \\ \gamma_{\text{mix}} = \lambda_{\text{mix}}\sigma(f_i) + (1 - \lambda_{\text{mix}})\,\sigma_{sf},\ \beta_{\text{mix}} = \lambda_{\text{mix}}\mu(f_i) + (1 - \lambda_{\text{mix}})\mu_{mf} \\ \sigma_{sf} \in \mathbb{R}^c \sim\ STDFlow, \qquad \mu_{mf} \in \mathbb{R}^c \sim\ MuFlow \end{gathered} \tag{7}$$

here λ_{mix} is a coefficient used for applying interpolation between original statistics and sampled statistics, randomly sampled from [0, 1].

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

To evaluate the effectiveness of MixStyleFlow in domain generalization for medical image segmentation, we conducted experiments on two segmentation tasks: prostate MRI segmentation and optic disc/optic cup (OD/OC) segmentation in fundus images. These datasets include multiple domains representing different imaging centers, scanners, and acquisition protocols, ensuring a diverse and challenging evaluation setting. For prostate MRI segmentation, We use the Medical Decathlon [15] for training and

intra-domain testing, while the remaining six sites from [16] are used for testing. We followed [17] to preprocess MRI volumes, retaining only prostate slices, resizing them to 384×384, and using them as 2D training data. For OD/OC segmentation, we used color fundus images from five different domains [18-21], covering a total of 1,441 images, training on one source domain and evaluating the model on the remaining domains (**Table 1**). We utilized the Dice Similarity Coefficient (DSC) for evaluating segmentation performance.

Table 1. Overview of the datasets used for training and evaluating MixStyleFlow

Task	Modality	Domain Names	Cases in each Domain
Prostate Segmentation	MRI	Medical Decathlon, Site A–F	32 30; 30; 19; 13; 12; 12
OD/OC Segmentation	Fundus Image	Domain1–5	195; 95; 400; 650; 101

4.2 Implementation Details

We trained three pairs of normalizing flows (NFs) to model the distributions of feature means and standard deviations from image decoder layers 2, 3, and 4, using temperatures of 3.5, 2.5, and 1.5, respectively, for 100 epochs. These decoder layers were selected experimentally based on preliminary results. Each NF model comprises four masked affine coupling flows with ActNorm layers [22]. The coupling network included two MLPs for scale and translation, with layer dimensions [32, 64, 32], [16, 32, 16], and [16, 32, 16] for image decoder layers 2, 3, and 4, respectively, capturing complex feature relationships. The segmentation model was trained with a mini-batch size of 20 using the AdamW optimizer. The learning rate was 0.0001 for prostate segmentation and 0.001 for OD/OC segmentation. The training steps 1,600 epochs for the prostate and 100 epochs for the OD/OC segmentation. We implemented our approach in PyTorch on a single NVIDIA RTX 6000 GPU.

4.3 Results

Comparing with Other DG methods. We evaluated MixStyleFlow on both intra-domain (IID) and cross-site prostate segmentation tasks (**Table 2**). The results demonstrate the effectiveness of our approach in improving generalization across unseen domains. For the IID, MaxStyle (87.27) and MixStyleFlow (86.47) showed competitive performance, indicating their ability to maintain segmentation accuracy within the source domain. For cross-site generalization, our proposed MixStyleFlow consistently outperformed other methods across multiple unseen domains (Sites A–F), demonstrating its robustness to domain shifts. Compared to MaxStyle, which also performed well on some sites, MixStyleFlow exhibited superior generalization, particularly in more challenging domains such as Site E (68.5 vs. 58.43) and Site F (74.38 vs. 57.72).

To further assess the generalization capability of MixStyleFlow, we evaluated it on the joint segmentation of the optic disc (OD) and optic cup (OC) in fundus images across multiple domains (**Table 3**). MixStyleFlow consistently outperformed other

methods in most unseen domains. Especially, in comparison with the second-best performing method, TriD, our model gets the best score in all most domains with comparable results on Domain 5.

Table 2. Evaluation results (Dice scores) on the prostate intra-domain and unseen cross-site test sets. The 'Avg' column represents the average Dice score across the unseen domains (Site A–F). The best results are highlighted with **bold**.

Methods	IID	A	B	C	D	E	F	Avg
Baseline	84.94	87.94	71.38	79.74	79.61	50.18	54.17	70.5
MixStyle	84.03	88.98	54.38	81.54	84.05	63.69	62.95	72.6
DSU	83.11	88.21	59.24	82.24	83.04	58.56	64.03	72.55
MaxStyle	**87.27**	90.20	**83.94**	83.28	85.8	58.43	57.72	76.56
MixStyleFlow	86.47	**90.25**	83.75	**84.06**	**85.85**	**68.5**	**74.38**	**81.13**

Table 3. Evaluation results for joint OD and OC segmentation in fundus images. Each cell shows the Dice scores for OD (first) and OC (second). The model is trained on one domain and evaluated on the others. The best results are in **bold**.

Methods	Domain1	Domain2	Domain3	Domain4	Domain5	Avg
SAN-SAW	76.42, 59.01	83.79, 73.23	84.17, 65.51	81.83, 62.36	87.00, 64.42	82.64, 64.91
RandConv	79.63, 64.14	85.00, 72.40	87.77, 69.57	83.08, 64.38	86.31, 60.37	84.36, 66.17
MixStyle	75.67, 60.84	86.35, 73.77	85.86, 66.60	84.86, 66.44	86.54, 65.99	83.36, 66.73
EFDM	78.79, 57.73	84.83, 72.30	85.25, 65.94	82.13, 61.62	85.43, 63.02	83.29, 64.12
DSU	76.88, 61.26	84.17, 74.10	89.12, 70.16	83.53, 63.19	87.09, 59.65	84.16, 65.67
MaxStyle	77.40, 65.44	86.95, 74.52	87.95, 67.62	84.69, 66.05	87.95, 64.84	84.99, 67.69
TriD	81.86, 66.67	**88.19**, 75.43	89.62, 70.85	84.81, 67.53	**87.88, 66.96**	86.47, 69.49
MixStyleFlow	**86.16, 66.99**	88.16, **75.64**	**90.58, 75.74**	**87.17, 74.91**	86.09, 62.97	**87.63, 71.25**

Low data regime. To evaluate performance in low-data settings, we trained our model with 10% and 30% of the prostate dataset [15] and assessed results across different domains [16]. As shown in **Fig. 2**, MixStyleFlow consistently outperformed MaxStyle across most domains. With only 10% training data (**Fig. 2**.**A**), MixStyleFlow achieved notable gains in domain B (38.80% vs. 24.93%) and E (20.31% vs. 6.96%), demonstrating robustness in extremely low-data scenarios. Increasing the training set to 30% improved both methods (**Fig. 2**.**B)**, but MixStyleFlow remained superior, particularly in domain B (72.40% vs. 64.18%) and E (34.32% vs. 15.46%). These results highlight MixStyleFlow's strong generalization even with limited supervision.

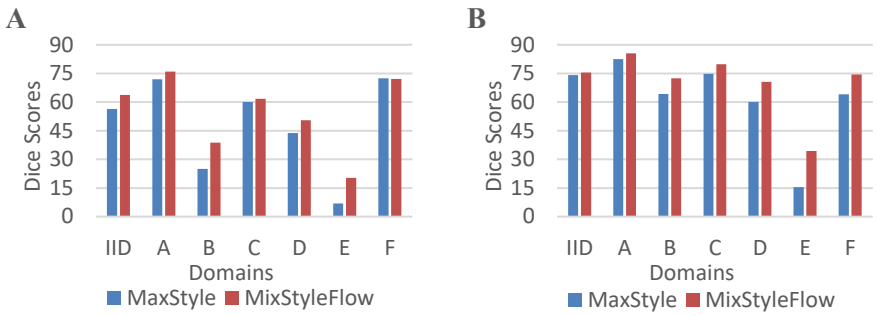

Fig. 2. Dice score evaluation on prostate intra-domain and cross-site test sets: (A) results with 10% training data; (B) results with 30% training data.

5 Conclusion

In this work, we introduced MixStyleFlow, a novel domain generalization framework that utilizes normalizing flows to model and perturb feature styles in a structured and expressive way. Unlike existing feature-space randomization techniques with limited and unordered perturbations, MixStyleFlow explicitly captures the distribution of domain feature styles, enabling controllable and diverse style augmentations. Our experiments on prostate MRI and optic disc/cup segmentation in fundus images demonstrated its superior generalizability to unseen domains. While we expect our proposed MixStyleFlow method will perform well with a high degree of generalizability in segmenting anatomical structures in various medical imaging applications, the computational overhead of normalizing flows may affect training efficiency compared with simpler perturbation methods. Future work could improve efficiency with optimized flow architectures and extend the approach to more imaging modalities and tasks.

Acknowledgments. This work was conducted in the JC STEM Lab of Innovative Medical Imaging Research funded by The Hong Kong Jockey Club Charities Trust.

Disclosure of Interests. The authors declare no competing financial or non-financial interests.

References

1. Azad, R., et al., *Medical image segmentation review: The success of u-net.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. **46**: p. 10076-10095.
2. Guan, H. and M. Liu, *Domain adaptation for medical image analysis: a survey.* IEEE Transactions on Biomedical Engineering, 2021. **69**(3): p. 1173-1185.
3. Xie, X., et al., *A survey on incorporating domain knowledge into deep learning for medical image analysis.* Medical Image Analysis, 2021. **69**: p. 101985.

4. Wang, J., et al., *Generalizing to unseen domains: A survey on domain generalization.* IEEE transactions on knowledge and data engineering, 2022. **35**(8): p. 8052-8072.
5. Zhou, K., et al., *Domain generalization: A survey.* IEEE transactions on pattern analysis and machine intelligence, 2022. **45**(4): p. 4396-4415.
6. Niu, Z., et al., *A survey on domain generalization for medical image analysis.* arXiv preprint arXiv:2402.05035, 2024.
7. Zhou, K., et al., *Domain Generalization with MixStyle*, in *In Proceedings of Ninth International Conference on Learning Representations*. 2021. p. 1-15.
8. Chen, C., et al. *Maxstyle: Adversarial style composition for robust medical image segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
9. Xu, Z., et al., *Robust and generalizable visual representation learning via random convolutions.* arXiv preprint arXiv:2007.13003, 2020.
10. Li, X., et al., *Uncertainty modeling for out-of-distribution generalization*, in *International Conference on Learning Representations*. 2022.
11. Chen, Z., et al. *Treasure in distribution: A domain randomization based multi-source domain generalization for 2d medical image segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023. Springer.
12. Zhang, Y., et al. *Exact feature distribution matching for arbitrary style transfer and domain generalization*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
13. Dinh, L., J. Sohl-Dickstein, and S. Bengio, *Density estimation using real nvp.* arXiv preprint arXiv:1605.08803, 2016.
14. Rezende, D. and S. Mohamed. *Variational inference with normalizing flows*. in *International conference on machine learning*. 2015. PMLR.
15. Antonelli, M., et al., *The medical segmentation decathlon.* Nature communications, 2022. **13**(1): p. 4128.
16. Liu, Q., et al. *Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
17. Hu, S., et al., *Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation.* IEEE Transactions on Medical Imaging, 2022. **42**(1): p. 233-244.
18. Almazroa, A., et al. *Retinal fundus images for glaucoma analysis: the RIGA dataset*. in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. 2018. SPIE.
19. Orlando, J.I., et al., *Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs.* Medical image analysis, 2020. **59**: p. 101570.
20. Sivaswamy, J., et al. *Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation*. in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*. 2014. IEEE.

21. Zhang, Z., et al. *Origa-light: An online retinal fundus image database for glaucoma analysis and research*. in *2010 Annual international conference of the IEEE engineering in medicine and biology*. 2010. IEEE.
22. Kingma, D.P. and P. Dhariwal, *Glow: Generative flow with invertible 1x1 convolutions.* Advances in neural information processing systems, 2018. **31**.