Generalize Ultrasound Image Segmentation via Instant and Plug & Play Style Transfer

GENERALIZE ULTRASOUND IMAGE SEGMENTATION VIA INSTANT AND PLUG & PLAY STYLE TRANSFER

Zhendong Liu[1,2*], *Xiaoqiong Huang*[1,2*], *Xin Yang*[1,2], *Rui Gao*[1,2], *Rui Li*[1,2], *Yuanji Zhang*[3], *Yankai Huang*[3], *Guangquan Zhou*[4], *Yi Xiong*[3], *Alejandro F Frangi*[1,5,6], *Dong Ni*[1,2 †]

[1]School of Biomedical Engineering, Health Science Center, Shenzhen University, China
[2] Medical UltraSound Image Computing (MUSIC) Lab, Shenzhen University, China
[3]Department of Ultrasound, Luohu People's Hospital, Shenzhen, China
[4] School of Biological Sciences and Medical Engineering, Southeast University, China
[5] Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK
[6] Medical Imaging Research Center (MIRC), University Hospital Gasthuisberg, Electrical Engineering Department, KU Leuven, Leuven, Belgium

ABSTRACT

Deep segmentation models that generalize to images with unknown appearance are important for realworld medical image analysis. Retraining models leads to high latency and complex pipelines, which are impractical in clinical settings. The situation becomes more severe for ultrasound image analysis because of their large appearance shifts. In this paper, we propose a novel method for robust segmentation under unknown appearance shifts. Our contribution is threefold. First, we advance a onestage plugandplay solution by embedding hierarchical style transfer units into a segmentation architecture. Our solution can remove appearance shifts and perform segmentation simultaneously. Second, we adopt Dynamic Instance Normalization to conduct precise and dynamic style transfer in a learnable manner, rather than previously fixed style normalization. Third, our solution is fast and lightweight for routine clinical adoption. Given 400 $\times$ 400 image input, our solution only needs an additional 0.2 ms and 1.92M FLOPs to handle appearance shifts compared to the baseline pipeline. Extensive experiments are conducted on a large dataset from three vendors demonstrate our proposed method enhances the robustness of deep segmentation models.

Index Terms— Ultrasound, Style transfer, Segmentation

1. INTRODUCTION

The tremendous success of deep neural networks (DNNs) has benefitted medical image analysis [1]. However, deployment of DNN models in real clinical scenarios is threatened by appearance shifts that degrade their performance. Different sources of appearance variation affect routine medical image acquisition, including operators, protocols, vendors, parameters and tissue properties, all of which can lead to unpredictable image appearance changes [2, 3]. The adverse effect of appearance shift on ultrasound image segmentation can be observed in Fig.1. Making DNNs robust against appearance shift is along the last mile before they can be clinically adopted.

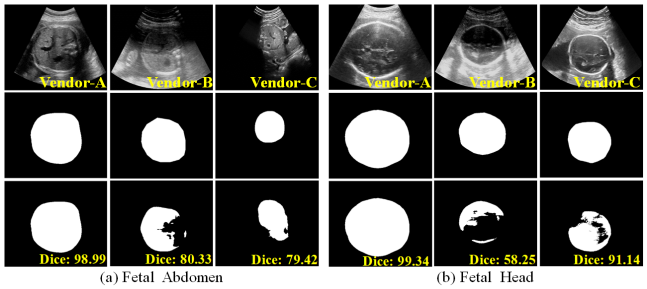

Fig. 1. Illustration of segmentation degradation on (a) fetal abdomen and (b) fetal head ultrasound images acquired from different vendors. The second row shows the segmentation ground truth, and the third row shows the predictions by a deep model trained on Vendor-A.

Automated solutions to remove image appearance shift are highly desirable but challenging. Real clinical applications often require the solution to be fast, practical and straightforward. Whereas, limited computation resources, the impossibility of retraining, and unpredictable appearance shifts are stringent constraints faced by these solutions, especially for ultrasound image analysis. Recently, Domain Adaptation (DA) has been proposed for dealing with image appearance shifts. Aligning appearance spaces [4, 5] or fea-

*Authors contributed equally.
†Corresponding author: nidong@szu.edu.cn.

ture spaces [6, 7, 8] of different domains were explored using generative adversarial learning. Although DA is attractive, it depends heavily on sufficient data from the target domain for complex retraining. The need to find a source-target domain mapping confines these solutions to cases where both domains are defined a priori. By revisiting the basic definition of appearance shift, style transfer [9] (ST) inspires a new and intuitive way to tackle this problem. ST removes appearance shift by rendering the appearance of the target content image as a reference style image. Compared to DA, ST is independent on the target domain, retraining-free, and suitable for images with unknown appearance shift. Ma *et al.* [10] made the early attempt to exploit an online ST to reduce the appearance variation for better cardiac MR segmentation. Liu *et al.* [11] proposed an Adaptive Instance Normalization (AdaIN) based style transfer module for vendor adaptation. However, these methods only regard ST as a pre-processing module isolated from segmentation model. This two-stage setting not only consumes extra computation resources but also blocks ST-segmentation interaction.

In this paper, we propose a novel ST based one-stage framework for robust ultrasound image segmentation against unknown appearance shift. Our contribution is three-fold. *First*, with the plug-and-play design of the ST module, we unify the ST and segmentation model into a one-stage framework to remove unknown appearance shift and perform segmentation simultaneously. *Second*, we adopt the Dynamic Instance Normalization (DIN) [12] to conduct style transfer at multiple layers of segmentation model through a learnable manner, which provides precise affine parameters for more accurate style transfer. *Third*, our solution is lightweight and reduces the computation burden for better deployment in clinical scenarios. Given a 400×400 image input, our solution only needs 0.2*ms* and 1.92M FLOPs to handle appearance shift. Segmentation models following our solution need no samples for time-consuming retraining before they can take domain-unknown images. Extensive experiments are conducted on 6,532 fetal head (FH) and abdomen (FA) ultrasound images from three vendors. The results show our proposed method outperforms state-of-the-art methods.

2. METHODOLOGY

Figure 2 presents a schematic view of our method. We first train the segmentation model on Vendor A (style) images by using a U-net architecture [13], which achieved remarkable success especially in medical image segmentation. We then freeze this model to train hierarchical DIN-nets by only using its encoder part and construct the final plug-and-play model (DINSeg) for segmenting content images from other vendors.

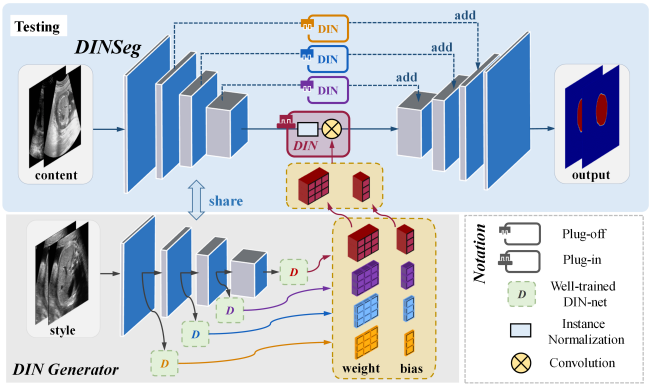

Fig. 2. Schematic view of our proposed framework.

2.1. One-stage Plug & Play Framework

Real clinical scenarios hold tough challenges for DNNs to combat appearance shift. Recently, ST has been proposed to remove the gap by rendering the appearance of unknown content images as style images. Previous ST-based methods [10, 11] trained two models and developed two-stage systems for segmenting images with appearance shift. In the first stage, the content image is transferred into a stylized image by the ST model. In the second stage, the segmentation model trained on style images is performed on the stylized image and get the result. This two-stage setting is straightforward, but it does not only consume extra computation resources but also blocks the interplay between ST and segmentation.

As shown in Fig. 2, we propose to simplify the pipeline into a novel one-stage design. Instead of a separate model, the DIN unit serves as the plug-and-play module for segmentation. Specifically, layer-specific DIN units are extracted from the trained DIN generator and then can be plugged into the segmentation model. These units parameterized by the appearance of a style image can align statistics of the content feature with those of the style features. In this plug-and-play setting, the choice of ST units is critical.

2.2. Choice of Style Transfer Units

Huang *et al.* [14] observed that matching feature statistics can achieve arbitrary style conversion. They proposed AdaIN to align style statistics after performing *Instance Normalization* (IN) on feature maps of the content image. AdaIN is powerful, but the affine parameters for style transfer are empirically defined as the basic statistics of mean and standard deviation, which may lead to the suboptimal ST performance. We propose to introduce *Dynamic Instance Normalization* (DIN) [12] to generate required plug-and-play DIN units for the first time in the literature. DIN units are developed in a learnable way and thus can encode a sophisticated style patterns.

As shown in Fig. 3, given a style image I_s, the affine parameters W^L and b^L are learned from the feature maps $\mathcal{F}_s^L$ of

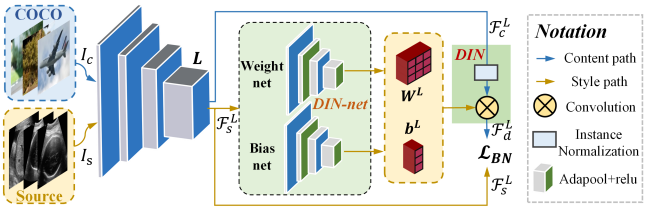

Fig. 3. System design to train DIN-net. The weight W^L and bias b^L are the affine parameters of DIN unit at layer L.

I_s by *DIN-net*. Then, normalized $\mathcal{F}_c^L$ is stylized by dynamic convolution with the learned weight W^L and bias b^L. DIN operation can be formulated as:

$$\textbf{DIN}(\mathcal{F}_c^L) = \textbf{IN}(\mathcal{F}_c^L) \otimes W^L + b^L \quad (1)$$

2.3. Training of Dynamic Instance Normalization Unit

Figure 3 depicts the details of training DIN-net. Unlike [12], which uses an encoder-decoder architecture to train DIN-net for ST, we only keep the encoder as our system core to simplify the training, which will generate the plug-and-play units. This modification also accelerates and stabilizes the training process. Specifically, our system shares the same encoder as the segmentation model. Only the DIN-net, including weight and bias branches, need to be trained. These two branches consist of two convolutional and adaptive pooling operations for handling arbitrary input sizes. To provide enough instances to improve and verify the ST performance, we use a large natural images corpus (COCO dataset [15]) as content image inputs for DIN-net training. With the training, DIN-net can learn complex and rich style patterns.

As shown in Fig. 3, ultrasound images serve as style images I_s. The frozen encoder extracts the style feature maps $\mathcal{F}_s^L$ at layer L. A layer-specific DIN-net is attached at layer L. The weight net and bias net are trained and then generate the affine parameters of DIN unit at layer L. Suggested by Li *et al.* [16], measuring the alignment loss via Batch Normalization (BN) statistics can achieve style transfer. Hence, given μ and σ as the channel-wise mean and standard deviation, we minimize the BN-Statistics Matching loss to convey the style information to DIN output feature $\mathcal{F}_d^L$ gradually as follow:

$$\mathcal{L}_{BN} = \|\mu(\mathcal{F}_d^L) - \mu(\mathcal{F}_s^L)\|_2 + \|\sigma(\mathcal{F}_d^L) - \sigma(\mathcal{F}_s^L)\|_2 \quad (2)$$

3. EXPERIMENTAL RESULTS

3.1. Dataset and Training Details.

Our datasets consist of the fetal head (FH), and abdomen (FA) images acquired from three ultrasound vendors, denoted as Ven-A, Ven-B and Ven-C. Images from Ven-B and Ven-C were further adjusted with different time gain compensations (TGC) to mimic more complex appearance shift variations. Our experiments involved 4200, 1516 and 816 ultrasound images from Ven-A, Ven-B and Ven-C, respectively. Experienced experts provided ground truth for all images. Ven-A images were set as style images to train the segmentation network. In contrast, images from the other two vendors were set as the unseen content images with unknown appearance shift. All the data acquisition was approved by local IRB and anonymized for segmentation.

We trained DIN-nets using 10,000 images from Microsoft COCO dataset as content images and 200 images from Ven-A as style images. We noted that DIN-net learns the affine parameters from ultrasound image in Vendor-A (style), rather than COCO content images. The COCO dataset we adopted can simulate complex style distribution and enable DIN-net to learn richer style patterns. Moreover, it is challenging to collect massive amount of ultrasound images to train DIN-nets fully but versatile COCO datasets are easy to obtain. Adam optimizer with learning rate 10^{-3} is set to minimize $\mathcal{L}_{BN}$. We trained a typical U-net segmentation model on Ven-A images and froze it during testing (baseline). We conducted all experiments with PyTorch on an NVIDIA GTX 2080 Ti GPU.

3.2. Quantitative and Qualitative Evaluation.

Based on our framework and the usage of DIN plug-ins, we devised two variants of our methods, namely Single-DINSeg (*S-DINSeg*) and Multi-DINSeg (*M-DINSeg*). As depicted in Fig. 2, *S-DINSeg* only plugs a single DIN at the end of the segmentation encoder, while *M-DINSeg* has an extra DIN in each of the last three skip connections. We compared them with the solution replacing DIN with AdaIN [14], and got two variants, *S-AdaINSeg* and *M-AdaINSeg*. We also implemented two typical two-stage frameworks for comparison, including the *StyleSegor* [10] and *WaveCT-AIN* (WCT-AIN) [11]. GAN based DA methods are not considered for comparison in this work, since they need samples from Ven-B and Ven-C for retraining. We used in total 6 indicators for evaluation, including Dice coefficient (Dice), Jaccard index (JAC), Hausdorff Distance of Boundaries (HDB), Average Surface Distance (ASD), Precision (PRE) and Recall (REC).

Table 1. Quantitative evaluation across different vendors.

	Metric	Baseline	S-DINSeg	M-DINSeg
		FA/FH	FA/FH	FA/FH
Ven-B	**Dice(%)**	88.63/94.93	92.27/96.20	93.58/96.97
	HDB(pixel)	17.03/11.87	13.39/7.52	12.50/6.07
Ven-C	**Dice(%)**	94.69/95.69	95.13/96.27	95.00/96.97
	HDB(pixel)	17.77/11.21	10.05/8.45	10.34/7.31

Table 1 shows the segmentation results of fetal images from two vendors. Both S-DINSeg and M-DINSeg get consistent improvements on two datasets, even when the baseline on Ven-C is already competitive. This indicates the efficacy of our proposed framework against unknown appearance shift.

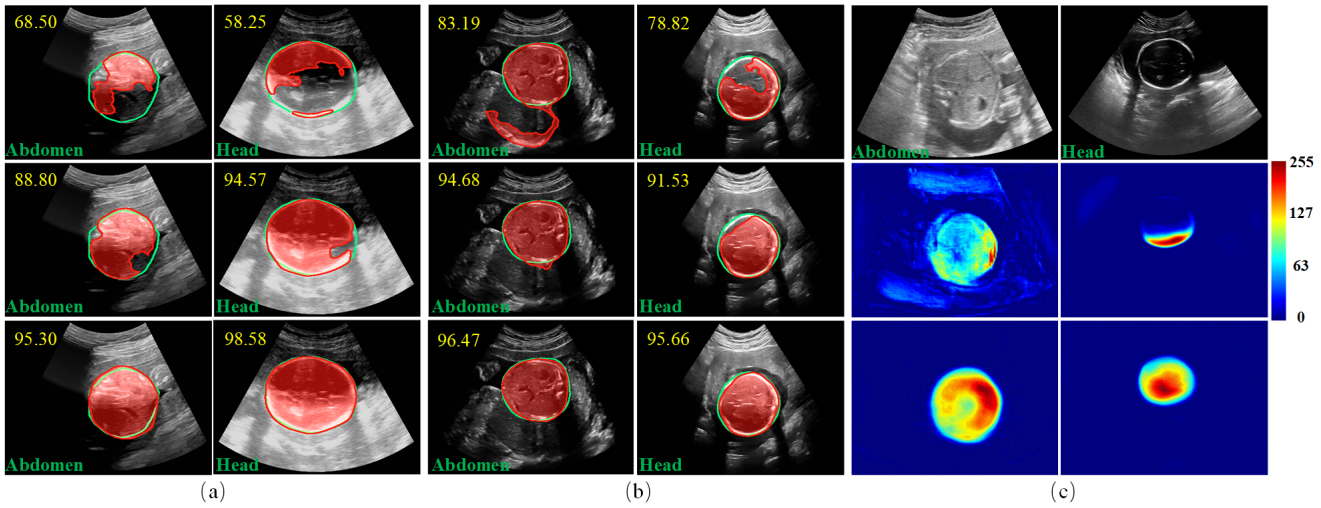

Fig. 4. (a)-(b) Segmentation results on Ven-B and Ven-C FA (left) and FH (right) images. Baseline, S-DINSeg and M-DINSeg results are listed from top to bottom. Green curve, red area and yellow digits denote ground truth, segmentation and Dice, respectively. (c) Feature maps before (middle row) and after (bottom row) DIN.

With the DIN used in multiple sites of the segmentation network, the results of M-DINSeg demonstrates that hierarchical DIN units can remove appearance shift better by encoding rich style information of different levels.

Table 2. Quantitative comparisons among different methods.

Methods	Dice(%)	JAC(%)	HDB	ASD	PRE(%)	REC(%)
	FA/FH	FA/FH	FA/FH	FA/FH	FA/FH	FA/FH
Baseline	88.63/94.93	83.42/91.67	17.03/11.87	4.11/2.86	95.35/97.29	86.11/94.03
HistEqual	84.16/87.69	78.68/83.54	20.47/14.94	4.63/3.15	93.93/93.59	81.21/85.86
StyleSegor [10]	89.89/95.11	84.54/91.74	16.88/11.78	3.97/2.81	95.33/96.56	88.12/94.42
WCT-AIN [11]	91.87/96.21	86.82/93.10	14.76/7.34	3.71/2.11	96.36/97.13	89.63/95.33
S-AdaINSeg	91.65/96.24	86.63/93.41	14.17/7.30	3.74/2.07	96.20/97.30	89.70/95.84
M-AdaINSeg	92.52/96.42	87.71/93.64	13.33/7.22	3.52/2.02	95.77/97.00	91.26/96.37
S-DINSeg	92.27/96.20	86.85/93.03	13.39/7.52	3.85/2.17	**96.58/97.54**	89.77/95.19
M-DINSeg	**93.58/96.97**	**88.76/94.38**	**12.50/6.07**	**3.52/1.76**	95.80/96.70	**92.42/97.43**

Table 2 lists the quantitative comparisons among our methods and other ST-based methods on Ven-B images, due to the space limitation. This indicates the efficacy of ST for segmenting images with appearance shift. Best results are achieved by our proposed method. M-DINSeg improves the Dice index by about 5 percent for FA and 2 percent for FH over their baselines. Among these methods, one-stage solutions, including M-AdaINSeg and DINSeg, show consistent advantages over the two-stage frameworks, like the StyleSegor and WCT-AIN. Hence, simplifying the pipeline and directly conducting ST in segmentation network are tractable and superior in removing appearance shift.

Figure 4 (a-b) visualize the results of our methods on segmenting FA and FH from Ven-B and Ven-C, respectively. With DIN plug-in, S-DINSeg can almost recover the segmentation from the broken masks obtained by the baseline. When DIN units further plug into the skip connections, M-DINSeg presents the most visually plausible segmentation and highest Dice when compared to the ground truth. Fig. 4 (c) gives insights into the change in feature maps before and after DIN-based style transfer. DIN removes appearance shift and makes the segmentation get strong activations only around regions of interest.

Table 3. Model complexity evaluation.

Methods	FLOPs(G)		Params(M)		Time(ms)	
	Transfer	Whole	Transfer	Whole	Transfer	Whole
StyleSegor [10]	49.17	209.12	138.36	172.89	3000	3029.88
WCT-AIN [11]	70.77	230.72	10.66	45.19	75	104.98
S-AdaINSeg	6.4e-4	159.95	/	34.53	0.17	30.15
M-AdaINSeg	9.81e-3	159.95	/	34.53	0.88	30.86
S-DINSeg	1.92e-3	159.95	0.002	34.532	0.20	30.26
M-DINSeg	1.94e-2	159.95	0.004	34.534	1.99	31.87

Table 3 investigates the computation complexity of different methods (input size is 400×400). The number of floating-point operations (FLOPs), total parameters (Params) and inference time (ms) are reported criteria. Using the same segmentation network, we evaluated the time of ST (Transfer) and the whole pipeline (Whole) separately. Results show that our system based on one-stage plug-ins solution enhance segmentation robustness in terms of efficiency and complexity.

4. CONCLUSION

In this work, we propose *DINSeg*, which unifies DIN and segmentation network in a single model to increase both generalization capacity and inference efficiency. We show that *DINSeg* achieves consistent improvement over existing effective systems on generalizing deep model across a new domain. The proposed system is high-speed and lightweight being thus amenable for deployment in clinical settings.

5. ACKNOWLEDGMENTS

This work was supported by the grant from National Key R&D Program of China (No.2019YFC0118300), Shenzhen Peacock Plan (No. KQTD2016053112051497, KQJSCX201 80328095606003).

6. COMPLIANCE WITH ETHICAL STANDARDS

We state that our work is an academic study for which no ethical approval was required.

7. REFERENCES

[1] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: a review," *Engineering*, 2019.

[2] E. Gibson, Y. Hu, N. Ghavami, H. U Ahmed, C. Moore, M. Emberton, H. J Huisman, and D. C Barratt, "Inter-site variability in prostate segmentation accuracy using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 506–514.

[3] W. Yan, Y. Wang, M. Xia, and Q Tao, "Edge-guided output adaptor: Highly efficient adaptation module for cross-vendor medical image segmentation," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1593–1597, 2019.

[4] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE TBME*, 2018.

[5] X. Yang, H. Dou, R. Li, X. Wang, C. Bian, S. Li, D Ni, and P.A. Heng, "Generalizing deep models for ultrasound image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 497–505.

[6] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 597–609.

[7] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G Abramson, and B. A Landman, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *ISBI*. IEEE, 2018, pp. 1217–1220.

[8] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network," in *CVPR*, 2018, pp. 9242–9251.

[9] L. A Gatys, A. S Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016, pp. 2414–2423.

[10] C. Ma, Z. Ji, and M. Gao, "Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 128–136.

[11] Z. Liu, X. Yang, R. Gao, S. Liu, H. Dou, S. He, Y. Huang, Y. Huang, H. Luo, Y. Zhang, et al., "Remove appearance shift for ultrasound image segmentation via fast and universal style transfer," *arXiv preprint arXiv:2002.05844*, 2020.

[12] Y. Jing, X. Liu, Y. Ding, X. Wang, E. Ding, M. Song, and S. Wen, "Dynamic instance normalization for arbitrary style transfer," *arXiv preprint arXiv:1911.06953*, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[15] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[16] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," *arXiv preprint arXiv:1701.01036*, 2017.