Physical parameter prediction by embedding human perceptual parameter for 3D garment modeling

Seongmin Lee*, Woojae Kim, Sewoong Ahn, Jaekyung Kim and Sanghoon Lee*
* Yonsei University, Seoul, Republic of Korea
E-mail: {lseong721, wooyoa, anse3832, jkkproject, slee}@yonsei.ac.kr Tel/Fax: +82-2-2123-7734

***Abstract*—To model garments into a virtual environment, it is crucial to predict the physical parameters of the simulated model. However, it is troublesome for a user or technical director to intuitively reflect their aesthetic intention using physical parameters. In this paper, we propose a framework that predicts various physical parameters (*e.g.,* stretch resistance, bend resistance, ...) by embedding human perceptual parameters (*e.g.,* wrinkly, stretchy, ...) in multi-task learning (MTL) perspective. By predicting both physical and perceptual parameters, we can effectively solve this problem, and can give an important cue to model a 3D garment maximizing users visual presence. Furthermore, by taking a class activation mapping method, our model seeks the intermediate visual understanding of physical and perceptual parameters. Through the rigorous experiments, we demonstrate that the predicted physical and perceptual parameters agree with subjective values.**

I. Introduction

In recent years, due to the development of head-mounted display and rendering methods, the demand for VR/AR content is increasing, resulting in rapid growth of 3D application markets. As a result, the understanding of 3D models became a critical factor to design vivid 3D scenes. Among various 3D modeling methods, a parametric simulation method that reflects physical effects into garments plays an important role in improving the realism of 3D garment model.

Generally, 3D garment models are simulated by graphics tools such as Maya, 3ds Max, and Cinema4D by tuning physical parameters of the material. In the case of Maya's nCloth, which is one of the well-utilized garment simulators, there are various physical parameters (*e.g.,* stretch resistance, bend resistance, ...).

Technical directors manipulate these physical parameters into the garment model to mimic the desired texture of the target material. In this manner, there are several studies which estimate physical parameters over garment videos [1], [2], [3]. However, since the physical parameters are determined by mathematics-based physics formula, there is a limitation of intuitive understanding between target garment and 3D model [4], [5]. For this reason, the technical director needs to acquire physical prior-knowledge to implement various garment in 3D scenes. Therefore, it is remains a tough problem to the technical directors to simulate realistic and vivid garment.

To address this problem, Sigal *et al.* [6] proposed a perceptual control space. In [6], they defined the term *perceptual parameters*, which are understandable at a glance such as wrinkly, stretchy, and so on. Although they conduct various simulations using the perceptual parameters using the method of subjective assessment, there is a lack of objective prediction of the physical parameters. Therefore, objective analysis of a new garment sample is still difficult, and it is hard to be applied in practical 3D modeling fields.

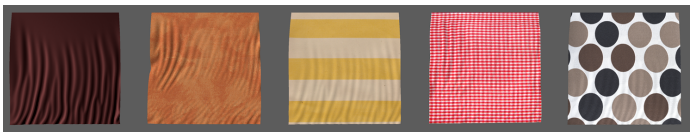


Fig. 1. **Garment patterns.** From left to right, it represents solid color, light pattern, simple pattern, complex pattern, and circular pattern.

Therefore, we propose a novel multi-stage framework, which is termed multi-task learning-based Garment Perceptual Physical Parameter Assessor (G3PA), fully utilizes the advantages of the human perceptual opinion and physical mechanism of the garment model. By using G3PA, it is applicable for the user to model 3D garment easily in the graphics applications, and can help intuitive understanding of garment materials. Furthermore, to visualize how the human visually perceives garment movements, we employ class activation map (CAM) algorithm [7]. In addition, to verify the performance of G3PA, we build the 3D garment video database.

The contributions of G3PA are 1) constructing a 3D garment video database which considers various garment patterns and physical parameters, 2) predicting both physical and perceptual parameters effectively by embedding perceptual parameters in a multi-task learning, 3) visualizing the activation map as an intermediate step to analyze the perceptual motion of the 3D garment.

II. 3D Garment Video Database

There were few studies analyzing the physical parameters by using the database introduced in [8]. This database is composed of real-world garments, therefore, it is hard to apply in graphics application directly. To tackle this problem, we construct a 3D garment database modeled by various physical types and 2D pattern types. Each garment video has 1024×1024 spatial resolution of 192 frames. 3D model based video sequences include various external physical motions generated by wind effect (wind gives textural motion to 3D garments as shown in Fig. 1), and they are simulated by nCloth in Maya.

TABLE I
DATA DETAILS.

Physical parameter	Value	Physical meaning
Stretch resistance	60, 300, 540, 780, 1020	Specifies the amount the current object resists stretching when under tension.
Bend resistance	0.2, 1, 1.8, 2.6, 3.4	Specifies the amount the object resists bending across edges when under strain.
Deform resistance	0.1, 0.5, 0.9, 1.3, 1.7	Specifies how much the current object wants to maintain its current shape.
Lift	0.1, 0.5, 0.9, 1.3, 1.7	The component of aerodynamic force perpendicular to the relative wind.
Drag	0.1, 0.5, 0.9, 1.3, 1.7	The component of aerodynamic force parallel to the relative wind.

In general, professional simulation artists have utilized 11 physical parameters to construct a general 3D model [6]. However, these parameters are correlated to each other, these can be combined to create similar garment movements. Therefore, it makes difficult to train the model stably. To address this, we choose 5 physical parameters (stretch/bend/deform resistance, lift and drag) out of 11 physical parameters that play an important role in the 3D garment model. The details of physical parameters are depicted in Table I.

The value of each physical parameter shown in Table I was chosen so that the users could recognize the physical changes. Then, each video in our database is made up of a combination of five physical parameters sampled randomly among the values of the selected physical parameters. Note that, due to the diversity of the database, we use 5 garment pattern types (solid, light, simple, complex and circular), which are depicted in Fig. 1. Therefore the size of our database is 5×25=125 videos, where 5 garment patterns and 25 physical parameter combinations.

In accordance with ITU-R BT.500-13 [9], we conducted a subjective experiment by using constructed garment videos. Firstly, we selected eleven candidates of perceptual parameters which have been analyzed by previous work [6]. Among them, we carefully chose four perceptual parameters (wrinkly, stretchy, heavy and smooth) to exclude the duplicate physical tendency on the 3D garment model. Each of 22 non-experts from the age group of 20-30 years participated in the experiment. All subjects were screened for normal visual acuity on the Landolt chart. Each subject watched garment videos and scored each perceptual parameter scale from 1 to 5 (Likert-like scale: 1 is *not <perceptual parameter>*, 5 equal to *very < perceptual parameter >*). The mean opinion score (MOS) of each individually evaluated parameter were used as ground-truth of the perceptual parameter.

III. FRAMEWORK OF G3PA

A. Overview of the proposed G3PA

Fig. 2 shows the overall framework of the G3PA. G3PA is composed of two stages. The first stage is pre-processing which considers motion characteristics as inputs, and it is described in Section III-B. The second stage is the multi-task learning, making the G3PA to predict the physical and perceptual parameters simultaneously. In Section III-C, we describe the details of the training strategies of the second stage.

B. Stage1: Motion Factor Extraction

Bouman *et al.* [1] experimentally showed that the motion is crucial when human perceives the physical properties of a garment. In particular, they also showed that the second order spatio-temporal derivatives (*i.e.,* a change in frame difference) are the most important motion factors for perceiving garment properties [1], [10].

The most intuitive way of motion estimation is to compute the frame difference. However, there is a problem in conducting the frame difference to predict the garment parameters. When a user perceives a texture of garment material, the local region having low-spatial frequency is more sensitive than those in the high-spatial-frequency. For instance, in Fig. 1, when the human seeks to perceive the texture of the garment, they tend to look the blown region by the wind than each patterns of the garment. Therefore, a simple Gaussian low-pass filter is applied to avoid input maps being focused in the unnecessary pattern area while meeting human visual characteristics. The frame difference map $D^{t,a}$ is computed as

$$D^{t,a} = \left| GLP(I^{t+a}) - GLP(I^{t}) \right|, \tag{1}$$

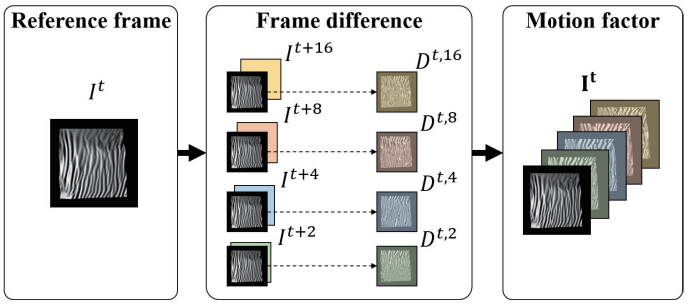
where I^t is *t-th* frame of a video resized to 256×256 resolution, and $GLP(\cdot)$ is Gaussian low-pass filter, and *a* is sampling range to consider various temporal variances from the reference frame (we use 2, 4, 8 and 16 in this paper). Here, we used the Gaussian low-pass filter of 3×3 kernel size and variation 1. After computing the frame difference map, we concatenate the frame difference maps in the reference frame I^t as channels. Therefore, the *t-th* sample set is defined as follows

$$\mathbf{I^t} = \left\{ I^t; D^{t,2}; D^{t,4}; D^{t,8}; D^{t,16} \mid t \in [1, T-16] \right\}, \tag{2}$$

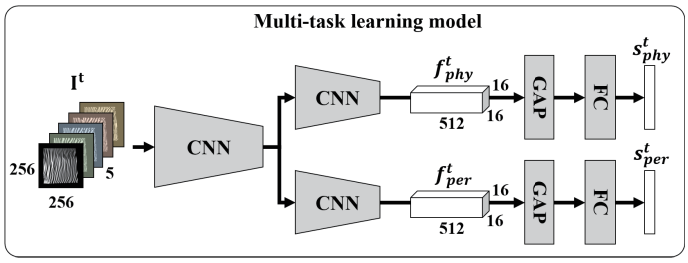
where T is the number of frames in one video sample. Since G3PA has I^t as and input, which has motion factor for the reference frame, we assume that entire frames in one video of our database has the same MOS and computes loss for the perceptual parameters.

C. Stage 2: Multi-task learning

Multi-task learning is a way to solve multiple tasks while training a generalized feature representation along the tasks [11]. To employ multi-task learning, there must be strong correlation across tasks [12], [13]. In other words, each task in multi-task learning should be related in terms of their purpose. In our approach, we believe the perceptual parameters and physical parameters have a strong relationship since Sigal *et al.* [6] has experimentally proved their relational importance. Therefore, we employ multi-task learning in physical parameter prediction by embedding perceptual parameter.



(a) Stage 1: motion factor extraction.



(b) Stage 2: multi-task learning of physical parameters and perceptual parameters.

Fig. 2. **The proposed G3PA framework.** For each reference frame, we design the input $\mathbf{I^t}$ that reflects the motion factor by concatenating the frame difference map. The multi-task learning model takes $\mathbf{I^t}$ as an input and predicts the physical parameters and perceptual parameters.

1) Multi-task loss: The proposed G3PA takes $\mathbf{I^t}$ as an input and it is fed to the first CNN consists of six convolution layers (3×3 kernel, 1 stride and 1 padding) with four max pooling layers. The feature map, which is the output of the first CNN, has learned the common traits of the garment movement caused by physical and perceptual parameters. Then, it is divided into two CNN branches to infer each task individually. The each second CNN has one convolution layer (3×3 kernel, 1 stride and 1 padding) without down-sampling. The output of the second CNN is $16 \times 16 \times 512$ size feature map. The feature map of the physical parameters and the perceptual parameters are termed f^t_{phy}, f^t_{per} at frame *t*, respectively. Then, each feature map f^t_{phy} and f^t_{per} are regressed onto each ground-truth parameters after global average pooling (GAP) layer and fully connected (FC) layer. To optimize the proposed model, we construct a multi-task learning loss with several constraint terms. Basically, the mean squared error is applied to minimize each error between each tasks' parameters and ground-truth parameters. Each tasks' loss l_p defined as

$$l_p = \sum_{i=1}^{N} \left\| s_p^{t,i} - GT_p^i \right\|^2 \quad (p \in \{phy, per\}), \tag{3}$$

where N is total number of physical or perceptual parameters, $s_p^{t,i}$ is *i-th* predicted parameter, and GT_p^i is *i-th* ground-truth parameters ($p \in \{phy, per\}$). However, $s_{phy}^{t,i}$ and $s_{per}^{t,i}$ are predicted for 16 frame lengths of the reference frame I^t. Therefore, the physical and perceptual parameters for one video are computed as the temporal average pooling of $s_{phy}^{t,i}$ and $s_{per}^{t,i}$. Consequently, the predicted *i-th* physical parameter S_{phy}^i and the *i-th* perceptual parameter S_{per}^i for a video are defined as

$$S_p^i = \frac{1}{T-16} \sum_{t=1}^{T-16} s_p^{t,i} \quad (p \in \{phy, per\}). \tag{4}$$

2) Total variation loss: If the model is optimized to minimize the MSE without any constraints, then the feature representation is affected by undesirable noise [14]. Therefore, we apply a smoothing restriction that penalizes high frequencies in the feature map using the total variation (*TV*) L_2 standard [15], [16], [17].

$$TV(f_p^t) = \frac{1}{H \cdot W} \sum_{x,y} (\mathbf{s}_{horz,p}(x,y)^2 + \mathbf{s}_{vert,p}(x,y)^2), \tag{5}$$

where *H, W* indicate the height and width of $\boldsymbol{s}$, and $\mathbf{s}^t_{horz,p}$, $\mathbf{s}^t_{vert,p}$ are Sobel-filtered feature maps in the horizontal and vertical directions, respectively ($p \in \{phy, per\}$), and (*x,y*) means spatial location. Total variation loss is defined as $l_{TV} = \sum_k TV(f^t_{p,k})$, where k indicate the channel of the feature map (k=1, 2, ..., 512). Final loss function of G3PA, l_{G3PA} is

$$l_{G3PA} = l_{phy} + \alpha \cdot l_{per} + \beta \cdot l_{TV} + \gamma \cdot l_2, \tag{6}$$

where l_2 is L_2 regularization term and α, β, γ are hyper parameters that determine the learning ratio between each loss. These hyper parameters are tuned by validation set ($\alpha = 0.8$, $\beta = 10^{-7}$ and $\gamma = 10^{-5}$).

D. Visualize Motion Factor

In order to visualize the perceptual motion factor, we compute class activation map from the extracted feature. Class activation map algorithm highlights the attention region that is relevant to the predicted values [7], [18]. As a result, the class activation maps M^i_{phy}, M^i_{per} for the *i-th* physical parameter and the *i-th* perceptual parameter is

$$M_p^{t,i}(x,y) = \sum_k w_{p,k}^i f_{p,k}^t(x,y) \quad (p \in \{phy, per\}), \tag{7}$$

where $w^i_{p,k}$ is the *k-th* weight of the fully connected layer regress into *i-th* physical or perceptual parameter $s_p^{t,i}$ ($k = 1, 2, ..., 512$).

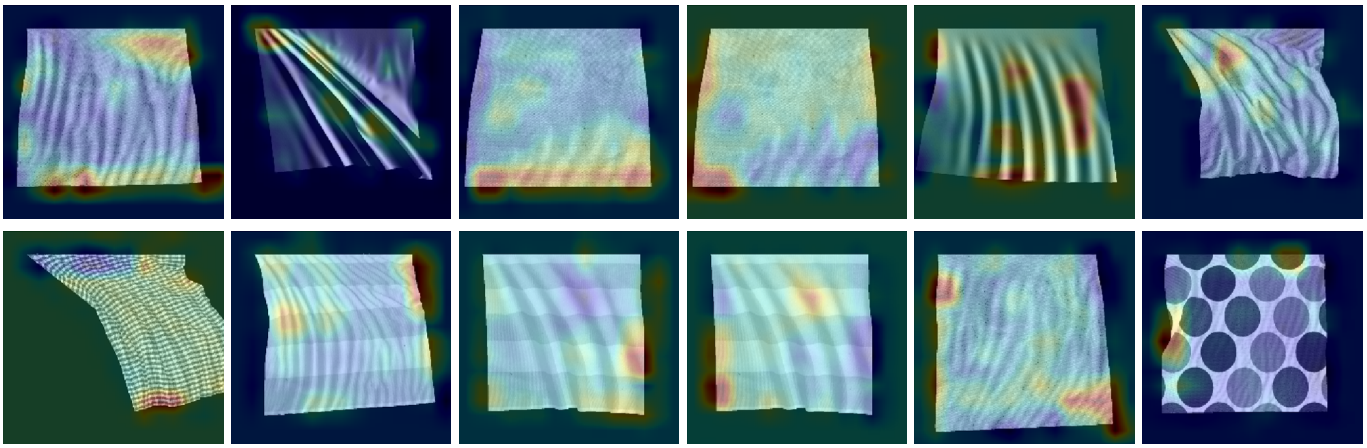
IV. Experiments

During the experiment, we randomly divided our database into three subset, 80% for training, 10% for validation and 10% for testing, and all parameters are normalized between 0 to 1, divided by the maximum value (*e.g.,* all stretch resistance is divide by 1020).

In order to validate performance of G3PA, an ablation test and visual analysis of the intermediate results are performed. To compare the performance of G3PA, the well-known correlation measurements were used: Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SROCC). Table II shows the ablation test results of three cases of the G3PA and subscripts means used loss.

TABLE II
ABLATION TEST RESULTS OF G3PA.

		Stretch resistance	Bend resistance	Deform resistance	Lift	Drag	Wrinkly	Stretchy	Heavy	Smooth
G3PA$_{phy}$	PLCC	0.741	0.700	0.861	0.837	0.554	-	-	-	-
	SROCC	0.776	0.715	0.883	0.806	0.545	-	-	-	-
G3PA$_{phy,per}$	PLCC	0.745	0.733	0.912	0.874	0.610	0.861	0.613	0.893	0.821
	SROCC	0.784	0.731	0.917	0.856	0.590	0.846	**0.675**	0.889	0.797
G3PA$_{phy,per,TV}$	PLCC	**0.776**	**0.747**	**0.913**	**0.888**	**0.639**	**0.879**	**0.629**	**0.895**	**0.846**
	SROCC	**0.804**	**0.767**	**0.919**	**0.864**	**0.616**	**0.859**	0.619	**0.898**	**0.820**



(a) Stretch resistance (b) Deform resistance (c) Lift (d) Drag (e) Wrinkly (f) Heavy

Fig. 3. **CAM visualization.** Results of class activation map visualization for each parameter.

The main point of the ablation test is that multi-task learning was applied to G3PA$_{phy,per}$ and G3PA$_{phy,per,TV}$. By comparing G3PA$_{phy}$ and G3PA$_{phy,per}$, we could see that the use of perceptual parameters improves physical parameter prediction performance. G3PA$_{phy,per,TV}$ shows the result of adding total variation loss term, along with physical and perceptual parameters. As a result, we show that multi-task learning about physical and perceptual parameters has noticeable performance improvements compared to the model that has only learned about physical parameters. Note that total variation loss reduces undesirable noise, resulting performance in superior performance. Therefore proposed G3PA is reasonable for understanding the 3D garment models.

Fig. 3 shows the class activation map $M_{phy}^{t,i}$, $M_{per}^{t,i}$ of G3PA. It shows that each parameter has a different activation region. Stretch resistance determines the degree of stretching on 3D garment. Therefore, edge and wrinkled regions are activated. In addition, when the garment was entirely stretched, we can see most of the area was activated except for the fixed upper part. Deform resistance decide how much the garment is trying to maintain its current shape. Since the amount of wrinkle intensity can be interpreted as the degree of keeping the current shape, deform resistance mainly activates on the wrinkled region. Lift is the perpendicular component of acting force, and conversely drag is the parallel component. According to this, activation map of lift and drag is highlighted on wrinkled and flat region, respectively. Consequently, we can see that activated regions of lift and drag are in an inverted relationship, and it is consistent to the physical meaning of parameters. In the case of perceptual parameters, since human perceives garment properties by motion factor, class activation map of every perceptual parameter is activated in the wrinkled region, but its details are different. We could see that Wrinkly activates the entire wrinkled region but Heavy largely activates in the deep wrinkled region.

V. CONCLUSION

In this paper, we proposed a novel garment parameters prediction framework named G3PA. The G3PA is a multi-stage framework, and we employ multi-task learning to predict both physical and perceptual parameter simultaneously. To verify the performance of G3PA, we construct a 3D garment database and show that G3PA has a remarkable performance. Furthermore, we visualize the intermediate results of G3PA and analyze them. Based on this, we believe that the G3PA provides an important motion factor that reflects the human visual characteristics, and it will help technical directors to design vivid 3D scenes.

VI. ACKNOWLEDGMENTS

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1702-08

REFERENCES

[1] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman, "Estimating the material properties of fabric from video," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1984–1991.

[2] Shan Yang, Junbang Liang, and Ming C Lin, "Learning-based cloth material recovery from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4383–4393.

[3] Kiran S Bhat, Christopher D Twigg, Jessica K Hodgins, Pradeep K Khosla, Zoran Popović, and Steven M Seitz, "Estimating cloth simulation parameters from video," in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2003, pp. 37–51.

[4] David Baraff and Andrew Witkin, "Large steps in cloth simulation," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 1998, pp. 43–54.

[5] Sofien Bouaziz, Sebastian Martin, Tiantian Liu, Ladislav Kavan, and Mark Pauly, "Projective dynamics: fusing constraint projections for fast simulation," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 154, 2014.

[6] Leonid Sigal, Moshe Mahler, Spencer Diaz, Kyna McIntosh, Elizabeth Carter, Timothy Richards, and Jessica Hodgins, "A perceptual control space for garment simulation," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 117, 2015.

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[8] K. L. Bouman, B. Xiao, P. Battaglia, and W. T. Freeman, "Estimating the material properties of fabric from video," *International Conference on Computer Vision (ICCV)*, 2013.

[9] RECOMMENDATION ITU-R BT, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[10] Bela Julesz, "Visual pattern discrimination," *IRE transactions on Information Theory*, vol. 8, no. 2, pp. 84–92, 1962.

[11] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[12] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Natalia Neverova, Alain Trémeau, and Christian Wolf, "Multi-task, multi-domain learning: application to semantic segmentation and pose regression," *Neurocomputing*, vol. 251, pp. 68–80, 2017.

[13] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[14] Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[15] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee, "Deep cnn-based blind image quality predictor," *IEEE transactions on neural networks and learning systems*, , no. 99, pp. 1–14, 2018.

[16] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.

[17] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1969–1977.

[18] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.