A High-fidelity Partial Face Manipulation Dataset for Enhanced Deepfake Detection

Kaitai Tong, Junbin Zhang, Yixiao Wang, Hamidreza Tohidypour, and Panos Nasiopoulos
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
{h7i0b, zjbthomas, yixiaow, htohidyp, panosn}@ece.ubc.ca

***Abstract*—Deepfake technology has already impacted the integrity of news and may grow to hugely destructive political and social force. The realistic and convincing nature of deepfakes poses a threat to the authenticity of information, alarming individuals and organizations. While many studies have explored the issue of deepfakes, the majority of them have focused on swapping entire faces rather than partially manipulating them, which can be more difficult to detect. In this paper, we introduce a high-fidelity partially manipulated face dataset, aiming to fill the gap in the existing deepfake research by providing a comprehensive benchmark for partially manipulated face detection. Our dataset includes a diverse set of partially manipulated faces which is generated from high-quality facial images. Our proposed alignment pipeline ensures that the partially manipulated faces may be realistically integrated into the original images, providing a more challenging evaluation environment for deepfake detection models. Both objective and subjective evaluations of our proposed dataset have shown promising results, indicating its potential to become a significant benchmark for partially manipulated face detection.**

***Index Terms*—Deep learning, Deepfake, Local facial feature editing, Generative adversarial network**

I. Introduction

The advancement of internet and digital media technologies has made information exchange and sharing much more expedient and advantageous. Despite the benefits of this information explosion, great challenges have risen in maintaining the authenticity of data, particularly images and videos. Digital images and videos manipulation could potentially mislead social media for nefarious purposes, promoting a propaganda by creating fake news and highly convincing disinformation [1]. In particular, facial manipulation has become a rapidly emerging issue in the society due to its impersonation of influencers, such as politicians and celebrities [2]. Nowadays, advanced deep learning approaches, known as deepfakes [3], have gained popularity among fraudsters, due to the use of generative adversarial networks (GANs) [4] that are capable of producing realistic synthetic facial images that mimic real-life data. It is, therefore, crucial to develop intelligent deepfake detection systems to avoid the misuse of deepfake technology.

Existing deep learning based deepfake detectors [5]–[7] were trained and evaluated only on deepfake datasets on face-swapping scenarios [8]–[10], in which the entire face of the source image is replaced. However, partially modifying only some facial features leads to more convincing deepfakes and detecting such images is a much more challenging proposition. This is because the tampered regions are more narrowly focused and concentrated in the case of partially modified faces. As a result, partially modified faces are increasingly more sophisticated fake and have a significantly more destructive potential. For this reason, in this paper, we aim at addressing the lack of a benchmark dataset consisting of partially manipulated faces that will be invaluable in designing deep learning networks for deepfake detection.

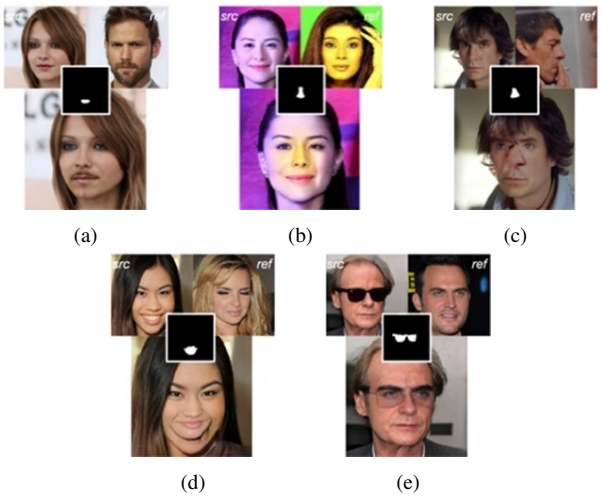

Fig. 1: Examples of typical low-quality partially manipulated face images due to (a) gender mismatch, (b) differences in skin color, (c) disparities in head orientations, (d) position offset between the facial parts in two faces (mouth in this example) and (e) artifacts appeared after removing eyeglasses from a face. In each subfigure, top left and right images are two faces that we attempted to mix. Binary masks that represent the to-be-modified parts of faces are placed in the middle. The bottom images are the manipulated faces.

Several techniques have been developed in recent years for changing only partial regions of the faces, e.g., only the eyes [11]–[14]. Among all the recent models that support editing parts of the faces, we adopted the state-of-the-art StyleMapGAN [14] to generate our dataset. Given two faces and the mask of the to-be-modified region, StyleMapGAN is able to mix the selected part of the faces from one to the other to generate partially manipulated faces. However, we notice that if the two selected faces are not aligned based on one or more visual aspects, the resultant face may look

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

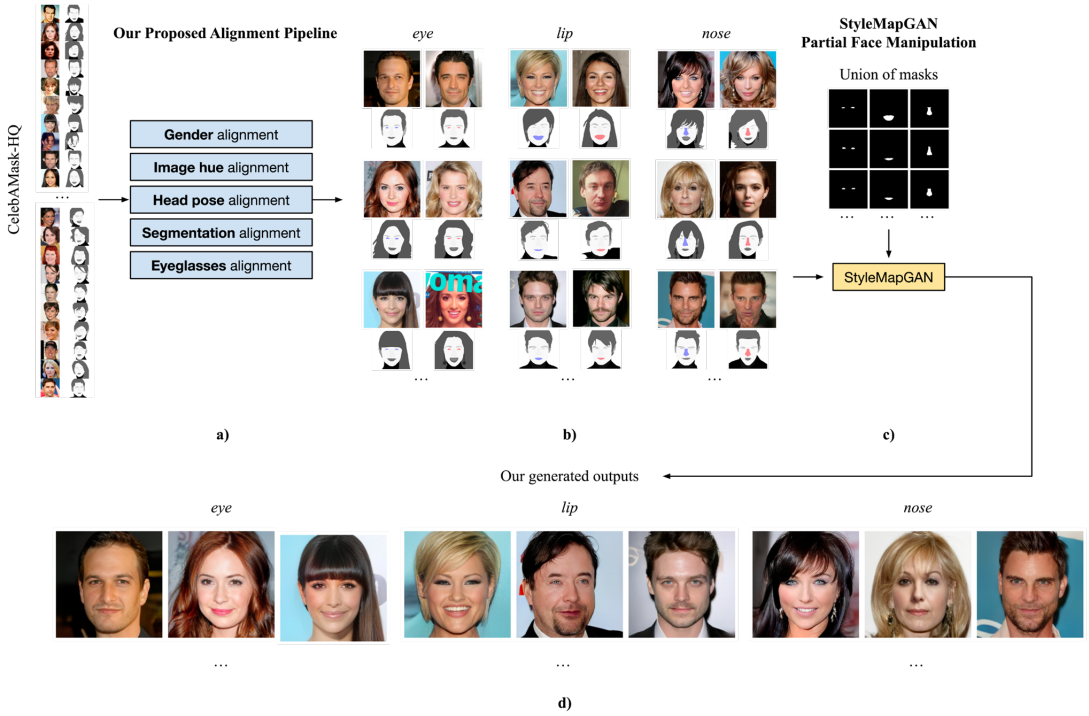

Fig. 2: An overview of the whole partial face manipulation workflow. a) It shows our proposed alignment pipeline, with multiple pre-processing steps. b) Some examples of selected pairs of faces for different parts of editing in our dataset, the left images are the source faces, and the right images are the reference faces. At the bottom of each pair, we also present the extracted masks from the CelebAMask-HQ, highlighted in blue and red. c) It shows the union of the extracted masks for each pair. The union mask is fed into the StyleMapGAN network along with the source and reference faces. d) The resulting high-quality faces that are generated after applying our proposed alignment pipeline.

unnatural. Fig. 1 depicts five typical examples that led to low-quality mixed faces. These mismatches will drastically decrease the overall fidelity of the deepfake datasets, resulting in significant performance degradation of deepfake detection models trained on the dataset. To build a natural-looking partially modified face dataset, we first identified the selection criteria for choosing two faces whose fusion by deepfake networks will most likely lead to a natural-looking face. Then, we proposed a series of pre-processing steps that are based on these criteria to generate a realistic dataset. Our objective and subjective evaluations showed the effectiveness of our proposed pre-processing steps for generating faces with partially modified eyes, lips, and noses.

The rest of the paper is organized as follows: in Section II we describe the steps of dataset generation and the proposed alignment pipeline. In Section III we present our experiments and discuss the results. We drew the conclusion in Section IV.

II. Proposed Method

The overall workflow for partially modifying faces with our proposed pre-processing steps is shown in Fig. 2. The approach consists of two parts, our proposed selection pipeline with multiple pre-processing steps, and the StyleMapGAN pipeline used for partially modifying faces. The following subsections first provide an overview of StyleMapGAN and then introduce our pre-processing steps.

A. Partial Face Manipulation

We adopted StyleMapGAN [14] to perform partial face manipulation tasks. StyleMapGAN is trained to modify faces using a large-scale celebrity facial dataset called the CelebAMask-HQ [15]. This dataset consists of 30,000 faces and the corresponding segmentation masks that separate different parts of the faces. Examples are shown in Fig. 3.

To mix two faces using StyleMapGAN, we provide a source face, which is the one we want to modify, and a reference image whose selected facial parts will be mixed with those of the source face. In addition, StyleMapGAN requires one mask that marks the to-be-modified regions of the source face. For this purpose, we create the segmentation masks of the two faces from the CelebAMask-HQ with the face parts (e.g., eye, lip, nose) that are supposed to be modified. We denote these masks as $\boldsymbol{M}^{e}_{src}$ and $\boldsymbol{M}^{e}_{ref}$, where e represents the to-be-modified parts of the faces, hereafter. Then, we take the union

Fig. 3: Sample faces from CelebAMask-HQ [15], along with their corresponding segmentation masks.

of these two masks ($\boldsymbol{M} = \boldsymbol{M}^e_{src} \cup \boldsymbol{M}^e_{ref}$).

Given these inputs, StyleMapGAN first uses an encoder structure to encode the two faces into low dimensional vectors that represent different parts of the faces in a disentangled manner. The resulting low dimensional vectors for the source and reference faces are denoted as $\boldsymbol{Z}_{src}$ and $\boldsymbol{Z}_{ref}$, respectively. Then, StyleMapGAN masks these vectors using the union of the masks to generate another vector to represent the mixed face as follows:

$$\ddot{\boldsymbol{Z}} = \boldsymbol{M}^e \otimes \boldsymbol{Z}_{ref} \oplus (1 - \boldsymbol{M}^e) \otimes \boldsymbol{Z}_{src}, \tag{1}$$

where $\oplus$ is the element-wise addition operation, and $\otimes$ is the element-wise multiplication. Finally, the manipulated face is produced by passing the new vector $\ddot{\boldsymbol{Z}}$ into a decoder. We refer readers to [14] for more about StyleMapGAN.

B. Our Proposed Alignment Pipeline

As illustrated in Fig. 1, it is crucial to identify reference faces that seem to be not aligned with the source face, to avoid generating unnatural faces. Therefore, we proposed a selection pipeline with five pre-processing steps for this purpose, each one designed to address one of the issues described in Fig. 1.

In the following subsections, we use index i to denote a source face and index j to denote a reference face. The total number of faces in the CelebAMask-HQ is $N = 30,000$. This means i and j cannot exceed N.

1) Gender alignment:

As shown in Fig. 1(a), the manipulated faces may seem to be unnatural when the source and reference faces belong to people of different genders. Therefore, we defined the criterion for gender comparison $\boldsymbol{G} \in \{0, 1\}$, as follows:

$$\boldsymbol{G}_{i,j} = \begin{cases} 0, & \text{if } g_i \neq g_j \\ 1, & \text{otherwise} \end{cases}, \tag{2}$$

where g_i and g_j represent the genders of the source and reference faces, respectively. A pair of source and reference faces would be excluded if $\boldsymbol{G}_{i,j} = 0$ (i.e., the genders of two faces are different).

2) Image hue alignment:

As can be seen in Fig. 1(b), partially modified faces that are generated by mixing two different image hues may have high contrast near the editing region. We handled this inconsistency by applying a threshold to the difference between average color of the skin areas of the source and reference faces. The inconsistency is computed using Delta E metric, which measures the differences of how humans perceive color in CIELAB color space [16]. We denoted it as $\Delta E^*_{ab}(\cdot)$, hereafter.

To get the facial areas, we relied on segmentation masks in the CelebAMask-HQ and extract the areas that were marked as "skin", which form a mask $\boldsymbol{M}^S$. Then, we convert the face images into CIELAB color space $\boldsymbol{I}^{L^*a^*b^*}$ and calculate the skin color difference $\boldsymbol{S}$ as follows:

$$\boldsymbol{S}_{i,j} = \Delta E^*_{ab}(\overline{\boldsymbol{M}^S_i \otimes \boldsymbol{I}^{L^*a^*b^*}_i}, \overline{\boldsymbol{M}^S_j \otimes \boldsymbol{I}^{L^*a^*b^*}_j}), \tag{3}$$

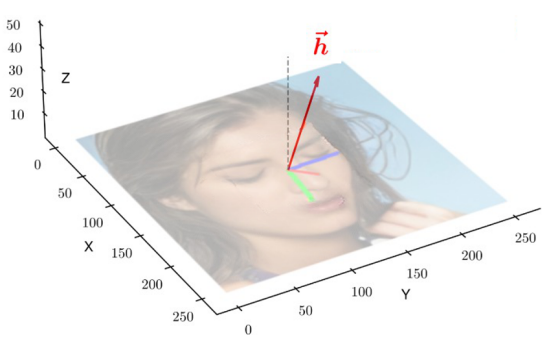

Fig. 4: An illustration of the head pose vector.

where $\otimes$ means element-wise multiplication, and the bar represents the average operation. Based on our experiments, we excluded pairs of source and reference faces if $\boldsymbol{S}_{i,j} > 2$.

3) Head pose alignment:

As it is shown in Fig. 1(c), the head pose is another crucial factor in creating more natural-looking faces. For this case, we take advantage of the state-of-the-art head pose detection methods to estimate head poses of two faces and calculate the similarity between the two head pose vectors using the cosine similarity function. We first used Hopenet [17] to estimate the head poses of faces as vectors $\overrightarrow{h}$ in the three-dimensional (3D) space ($\mathbb{R}^3$). The reason to use 3D vectors is that they contain depth information for more accurate head pose estimation. Such 3D vectors are visualized in Fig. 4. Then, the cosine similarity function was applied to the head pose vectors of the source and reference faces, which yielded the similarity measurement for head poses $\boldsymbol{H}$:

$$\boldsymbol{H}_{i,j} = \frac{\overrightarrow{h_i} \cdot \overrightarrow{h_j}}{\left\| \overrightarrow{h_i} \cdot \overrightarrow{h_j} \right\|}, \tag{4}$$

Based on our experiments, we excluded pairs of source and reference faces if $\boldsymbol{H}_{i,j} < 0.99$.

4) Segmentation mask alignment:

We noticed that if the positions of the parts of to-be-modified face have a large offset compared to the associated parts of the reference face, the resulting manipulated face may look unnatural (see Fig. 1(d)). In this case, the two extracted segmentation masks $\boldsymbol{M}^e_{src}$ and $\boldsymbol{M}^e_{ref}$ (as discussed in Section II-A) would have little intersection. As such, we estimated the intersection of two segmentation masks by computing the Dice Similarity Coefficient (DSC) [18] as follows:

$$\boldsymbol{D}_{i,j} = \frac{2 \sum \boldsymbol{M}^e_i \otimes \boldsymbol{M}^e_j}{\sum (\boldsymbol{M}^e_i)^2 \otimes \sum (\boldsymbol{M}^e_j)^2}, \tag{5}$$

Similar to previous pre-processing steps, we used the thresholding approach to exclude unwanted pairs of source and reference faces. However, in this case, the threshold varies based on the part of faces we want to modify. Specifically, we excluded pairs of faces: (a) for the to-be-modified eyes, if $\boldsymbol{D}_{i,j} > 1.0$ (i.e., we observed that for this case, other preprocessing steps are sufficient for producing high quality

TABLE I: The scores of NR-IQA metrics for different face parts. The symbol ↑ means the higher the better, while ↓ means the lower the better.

		eye			lip			nose		
		Ours	Random	[19]	Ours	Random	[19]	Ours	Random	[19]
NIQE [21]	↓	**3.411**	3.477	4.481	**3.462**	3.555	4.526	**3.356**	3.461	4.486
BRISTQUE [22]	↓	**22.546**	22.974	31.780	**22.707**	23.094	32.002	**22.714**	23.081	31.763
MetaIQA [23]	↑	**0.545**	0.535	0.518	**0.546**	0.535	0.517	**0.547**	0.536	0.520
MetaIQA+ [24]	↑	**0.836**	0.820	0.810	**0.834**	0.815	0.805	**0.838**	0.822	0.809

manipulated eyes, so no pair of faces is excluded in this step), (b) for the to-be-modified noses, if $\boldsymbol{D}_{i,j} \leq 0.9$ and, (c) for the to-be-modified lips, if $\boldsymbol{D}_{i,j} \leq 0.8$.

5) Eyeglasses alignment:

As shown in Fig. 1(e), we observed that when the source face has eyeglasses while the reference face does not, the output mixed face will contain artifacts. To address this, since the CelebAMask-HQ provides information on whether or not a face contains eyeglasses (denoted as $eg \in \{0, 1\}$), we defined the eyeglasses criterion $\boldsymbol{E} \in \{0, 1\}$ as follows:

$$\boldsymbol{E}_{i,j} = \begin{cases} 0, & \text{if } eg_i \neq eg_j \\ 1, & \text{otherwise} \end{cases}, \tag{6}$$

where eg_i and eg_j denote if the source and reference faces have eyeglasses, respectively. A pair of source and reference faces would be excluded if $\boldsymbol{E}_{i,j} = 0$.

III. Experimental Results

A. No-reference Image Quality Assessment

To evaluate the effectiveness of our proposed selection pipeline and the pre-processing steps, first we applied a vast variety of No-reference Image Quality Assessment (NR-IQA) metrics to compare the quality of the modified faces. We compared the generated partially modified faces in three cases: (1) modified faces generated using two randomly selected faces, (2) modified faces generated using two selected faces from our selection pipeline, and (3) modified faces created by the method in [19]. It is worth mentioning that the method proposed in [19] is the state-of-the-art approach that also modifies parts of a face by mixing two faces. It adopted a NR-IQA called Generated Image Quality Assessment (GIQA) [20] as a post-processing step to filter out low-quality faces.

For each of these three cases, we generated 1,000 modified faces for each of the three categories supported by our method, namely eye, lip, and nose. We applied four state-of-the-art NR-IQA metrics on each of the three cases: (1) Natural Image Quality Evaluator (NIQE) [21], (2) Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [22], (3) Meta-learning-based Image Quality Assessment (MetaIQA) [23] and (4) Meta-learning-based Image Quality Assessment Plus (MetaIQA+) [24].

We present the mean values of all metric scores over 1,000 faces for each modified part in Table I. It is observed that modified faces generated after applying our selection pipeline always achieve the best quality among the three cases. Note that, regardless of what face parts are modified, the objective quality scores generated using our proposed selection pipeline were always high, indicating the robustness of our proposed selection pipeline.

B. Subjective Test

As humans are our target audience that will decide the naturalness of modified faces, we conducted a subjective test. We generated 100 images for each of the modified parts (i.e., eye, lip, and nose) for two cases: (1) modified faces were generated using two randomly selected faces, (2) modified faces were generated using two faces selected by our selection pipeline. These images were shown to twenty (20) participants. Their task was to rate the naturalness of each single face by assigning a score between one (worst) and five (best) within 8 seconds. Table II presents the results of the Mean Opinion Score (MOS) of the two approaches. As can be seen, the images that were generated using our approach received the highest MOS scores. In addition, MOS results of our approach were stable as its standard deviation values were lower compared to those of the random selection approach. These results clearly show the superiority of our approach over random selection in creating realistic images.

IV. Conclusion

In this paper, we generated a novel high-fidelity partially manipulated face dataset that aims at filling the gap in existing deepfake research by providing a comprehensive benchmark for partially manipulated face detection. In this regard, we first identified the selection criteria that should be considered to choose two faces, whose fusion by deep learning deepfake models will most likely lead to a natural-looking face. Then, we proposed a series of novel pre-processing steps based on these criteria to generate a realistic and high-fidelity dataset. Our objective and subjective evaluations show the superiority of our approach in generating realistic faces compared to the state-of-the-art. Moreover, our evaluations indicate the great potential of our proposed dataset to become a significant benchmark for partially manipulated face detection.

TABLE II: Results of our subjective tests. The symbol ↑ represents the higher the better, while ↓ means the lower the better.

	MOS ↑		Standard Variation ↓	
	Ours	Random	Ours	Random
eye	**4.037**	3.419	**1.252**	1.490
lip	**4.132**	2.697	**1.213**	1.528
nose	**3.997**	2.612	**1.250**	1.552

REFERENCES

[1] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.

[2] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, p. 80–87, 2019.

[3] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, p. 910–932, 2020.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, 2020.

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *(WIFS)*, 2018, p. 1–7.

[6] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, p. 6105–6114.

[7] J. Zhang, H. Tohidypour, Y. Wang, and P. Nasiopoulos, "Shallow- and deep-fake image manipulation localization using deep learning," in *ICNC*, 2023.

[8] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, p. 2382–2390.

[9] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *CVPR*, 2020, p. 3207–3216.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, p. 1–11.

[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, p. 8789–8797.

[12] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Arbitrary facial attribute editing: Only change what you want," 2017.

[13] G. Yang, N. Fei, M. Ding, G. Liu, Z. Lu, and T. Xiang, "L2m-gan: Learning to manipulate latent space semantics for facial attribute editing," in *CVPR*, 2021, p. 2951–2960.

[14] H. Kim, Y. Choi, J. Kim, S. Yoo, and Y. Uh, "Exploiting spatial dimensions of latent in gan for real-time image editing," in *CVPR*, 2021, p. 852–861.

[15] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *CVPR*, 2020, p. 5549–5558.

[16] W. Mokrzycki and M. Tatol, "Colour difference delta e - a survey," *Mach. Graph. Vis*, vol. 20, no. 4, p. 383–411, 2011.

[17] N. Ruiz, E. Chong, and J. Rehg, "Fine-grained head pose estimation without keypoints," in *CVPRW*, 2018, p. 2074–2083.

[18] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, p. 297–302, 1945.

[19] R. Shao, T. Wu, and Z. Liu, "Detecting and recovering sequential deepfake manipulation," in *ECCV*, 2022, p. 712–728.

[20] S. Gu, J. Bao, D. Chen, and F. Wen, "Giqa: Generated image quality assessment," in *ECCV*, 2020, p. 369–385.

[21] A. Mittal, R. Soundararajan, and A. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, p. 209–212, 2012.

[22] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, p. 4695–4708, 2012.

[23] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *CVPR*, 2020, p. 14143–14152.

[24] ——, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, p. 1048–1060, 2021.