2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

MARMnet: On the Application of AI to Video Quality Control for Broadcasting Companies

Riccardo Musmeci, *NTT DATA Italy*, riccardo.musmeci@nttdata.com
Adriano Manfré, *NTT DATA Italy*, adriano.manfre@nttdata.com
Tomohiro Ohtani, *NTT Data Corporation*, tomohiro.ohtani@nttdata.com
Seika Suzuki, *NTT Data Corporation*, seika.suzuki@nttdata.com

2021 13th International Conference on Quality of Multimedia Experience (QoMEX) | 978-1-6654-3589-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/QoMEX51781.2021.9465478

***Abstract*—Video Quality Control (VQC) is the process that media and broadcasting companies employ in order to detect potential anomalies in contents to release to the final users. This process is currently performed by human operators and presents many drawbacks in terms of costs and time spent on such activity. As an example, the operators leverage on expensive equipment that in few years needs to be changed. Additionally, the report generated by the operators present inconsistency in terms of anomalies identified. In this paper, we take the initial steps towards the integration of Artificial Intelligence into the VQC pipeline in order to introduce speed and uniformity in the report generation. Our solution, called Augmented Video Quality Control (AVQC), relies on the power of MARMnet, an ad-hoc neural network able to detect two anomalies of interest. Such model enables the AVQC to report the anomaly in half of the time needed by the current process. Moreover, quantitative and qualitative results show that MARMnet provides high performances in detecting the two anomalies in the videos to analyze.**

***Index Terms*—Video Quality Control, Broadcasting, Deep Learning, Human Assisted AI**

I. Introduction

In recent years, the global video content market has been growing exponentially. Particularly, television broadcasting and video streaming companies have seen a tremendous increase in terms of users and market value. For instance, it has been estimated that the value of the television broadcasting market will reach nearly $346.87 billion by 2022 [1], while video streaming services' market value reached $42.6 billion in 2019 and it is expected to have a *Compound Annual Growth Rate* (CAGR) of 20.4% from 2020 to 2027 [2].

In this scenario, the production process of video contents plays a fundamental role. Many steps are required in order to release a video content with satisfying quality for the final user: writing a script, shooting the video, editing, checking, and finally distributing. Clearly, every step of this pipeline is crucial for the success of video contents. Within this pipeline, the checking phase and the *Video Quality Control* (VQC) process have recently captured our interest and attention.

The VQC process is a complete inspection of a video, from head to tail, to ensure that the content is anomaly-free and meets all the technical specifications for distributors, networks, or clients. Usually, this process is performed by human operators, specifically trained on detecting anomalies in videos. Operators, by watching the content, annotate in a report the relevant technical anomalies encountered. Such anomalies may be related to a non-conforming color settings in the scene, a sudden black frame, deteriorated images that need restoration, a non-dubbed foreign video as well as inaccurate subtitles. Finally, when the report is completely filled, the video is fixed according to the anomalies reported.

978-1-6654-3589-5/21/$31.00 ©2021 IEEE

The current VQC process presents different limitations. Among these, the most important ones regard the effort spent by the operators in checking the content as well as the equipment needed to conduct the analysis. Additionally, the current VQC methodology is prone to a high number of inconsistencies among the reports produced during the process. In fact, operators often use different writing styles for the same anomaly leading to inconsistency.

Artificial Intelligence (AI) has been extensively used in different business fields for speeding up quality control processes that initially required human intervention solely [3]–[5]. The AI-based quality control processes have been engineered with the goal of supporting human operators. The idea is that the system indicates where and when some anomalies could appear. Then the operator validates such indications and reports their correctness. In such setting, the AI models' goal is to maximize the number of true positives and minimize the false positives. Having a high number of false positives means that the operators could waste their time in double checking false alarms, which may lead to a system that is not time convenient. Conversely, a high rate of true positive means that the system is able to correctly identify an anomaly when it appears.

To the best of our knowledge, VQC processes have not been integrated with AI capabilities yet. Therefore, we investigated the possibility of employing an AI setting into a VQC process. As a result, we propose an *Augmented Video Quality Control* (AVQC) system that leverages on Deep Learning techniques in order to support human operators in detecting anomalies and automatically filling reports, with the result of a standardized and more efficient process. To build such system we collaborated with one of the most important Media & Broadcasting company, interested in detecting two anomalies in the contents to broadcast. These anomalies regard the non-conforming level of colors as well as deteriorated images attributable to low-quality footage, e.g. videos from the 80s. The AVQC core that copes with these two anomalies consists of a neural network, called *MARMnet*, designed and developed in order to simulate how the operators work and to deliver high performance in

the quality control task.

II. Related Work

Media contents quality control through artificial intelligence has received significant attention in recent years [6]. Particularly, many authors have focused their research on finding deep learning models for the qualitative assessment of images and videos. Such models aim at replicating the human behavior in order to determine the visual quality perception of a media content.

In [7], the authors have found different approaches for the No Reference Video Quality Assessment (NRVQA). Such approaches leverage on the combination of different set of features, each with different nature (e.g. spatial, temporal, etc.), extracted from a single video, to produce an overall quality score of the content. Similarly, the authors in [8], defined an image quality assessment framework able to predict the visual sensitivity maps of an image agreeing with the human subjective opinions.

Process pattern recognition (PPR) is a field regarding the determination of specific patterns in a process. Detecting the anomalies based on the visual patterns in media contents makes the VQC belonging to such field. In the PPR area, different papers for the manufacturing processes have introduced artificial intelligence capabilities to detect specific patterns [9]–[11].

Despite the different applications found in literature, none of them regards the detection of potential anomalies during the video quality control performed by a broadcasting company. To this reason, in this paper, we present a novel approach based on deep learning capabilities to assess the presence of anomalies in media content. We focus on two specific anomalies, OLD QUALITY and COLOR, and we deliver MARMnet, a neural network able to simulate the human operator behaviors when analyzing a video and reporting one of the two anomalies.

III. The Video Quality Control Process

In a television broadcasting company, the VQC process aims at detecting all the potential anomalies in a video before it goes to the final release to the users. Currently, broadcasting companies leverage on teams of human operators trained on detecting VQC anomalies that by watching the video content, they first write down in a paper where and which errors they detected, and then they transfer all the notes on a digital report, called *Quality Control report* (QC report). Once the report is finalized, the video is checked and consequently edited at the time-codes reported.

This approach has many drawbacks. The first is related to the time spent on the visual inspection of the video. In fact, a human operator usually spends 2.5 times more of the duration of the content. Secondly, there is no standard for the digital report. Actually, each operator has his own manner to describe the errors, resulting in unstructured reports. Another drawback is related to the very expensive equipment needed to allow the operators to analyze the videos. As an example, each operator works in a room equipped with the current top notch TV monitors, audio systems, and supporting tools for general color and audio parameters. Such tools have to be changed every few years because they become obsolete and wear out easily.

Based on the knowledge and experience gained during the collaboration with Media & Broadcasting company, the anomalies to detect by the VQC are of different nature. Among these, some are related to non-dubbed audio and to the absence of subtitles. Other anomalies regard the missing translation of signs in salient areas of the scene (e.g. road signs, captions, etc.) or the sudden popping up of black frames. It is important to notice that each anomaly presents specific patterns and contexts that the operators learn to identify during their job training period. Hence, the anomaly detection does not rely on a subjective criteria. In this paper, two anomalies that regard the visual perception of the content quality are addressed: *COLOR* and *OLD QUALITY*.

The *COLOR* anomaly is referred to levels of Luma, Chroma, and primary colors (RGB) which result in off standard values. An example of COLOR anomaly is reported in Fig. 1a. However, reporting this anomaly strongly depends on the context of the video. For instance, in cases such as music TV shows, concerts, and talent TV shows, the lights around the stage play a significant role in deteriorating the video's colors, like shown in Fig. 1c. In this case, it is considered acceptable by the operators, and the anomaly is not reported.

The *OLD QUALITY* anomaly refers to a sequence of frames in which a grey-scale or deteriorated pattern is predominant in the sequence. This type of anomaly is reported only to mention that in that part of the video there might be needed an enhancement of the frames. In Fig. 2 two frames affected by the *OLD QUALITY* anomaly are shown. Moreover, the OLD QUALITY anomaly is not relevant when appears in only few frames. For instance, in a "normal" scene containing a dog shaking his head, few frames as the one on Fig. 1c may mislead the operator to report the anomaly, since the visual patterns resemble some OLD QUALITY cases. Conversely, by watching the entire scene, it is able to determine that this frame belongs to a normal scene and not to an anomalous one.

IV. Augmented Video Quality Control

Augmented Video Quality Control (AVQC) is designed to overcome the limitations of the current VQC process. To this aim, it introduces AI for detecting the anomalies followed by a standardization of the QC reports. Additionally, it supports the operators' work by generating a preliminary analysis of the content that they need to validate afterwards. This allows them to focus only on where the anomaly may appear and also drastically reduces the time needed to check a single content. Ultimately, it reduces the need of buying expensive equipment, such as monitors, to change every few years.

In our vision of AVQC, every anomaly to detect needs a specific custom model. In the following section, we present MARMnet, the neural network designed and developed to detect the OLD QUALITY and COLOR anomalies.

(a) *COLOR* anomaly (b) *COLOR* anomaly (c) Not a *COLOR* anomaly

Fig. 1: Frame samples for the *COLOR* anomaly. In (a) the numbers presents anomalous levels of red color. In (b) the blue area represents an anomaly since it has out-of-standard levels of blue. Figure (c) represents a case in which the context plays an important role to not report the COLOR anomaly. In fact, TV show stage scenes in which lights make the colors out-of-standards are considered as not anomalous.

(a) *OLD QUALITY* anomaly (b) *OLD QUALITY* anomaly (c) Not an *OLD QUALITY* anomaly

Fig. 2: Frame samples for the *OLD QUALITY* anomaly. In (a) a frame belonging to an old footage, while in (b) a frame belonging to gray-scale sequence. In (c) a dog shaking his head that produces a visual pattern similar to OLD QUALITY even though it belongs to a video with no OLD QUALITY anomaly.

A. *MARMnet*

MARMnet was designed with the goal of detecting an anomaly by simulating the behavior of the human operators during the quality control. Specifically, the operator assesses the presence of an anomaly by watching a small portion of a video. Such small sequence of consecutive frames is necessary for the operator to extrapolate the context of the scene and then possibly report the anomaly. For instance, when detecting the *COLOR* anomaly, the operator first understands the relevance of the scene, and then identifies the potential area with out-of-standard color levels. This results of great importance when the operators avoid possible false positive, such as the TV shows or concert scenes for the *COLOR* anomaly. As a consequence, MARMnet was designed in order to predict the presence of an anomaly based on a sequence of consecutive frames. With this setting, MARMnet can understand the context in which operates and decide whether the anomaly must be reported or not.

To replicate the operators behavior, MARMnet relies on the extraction of spatio-temporal features from the sequence of consecutive frames. Particularly, given the time t and a sequence $S_t = \{f_{t_1}, f_{t_2}, ..., f_{t_N}\}$ of N consecutive frames, f_{t_i}, taken at time $t+i$ in the video, MARMnet employs a Convolutional Neural Network (CNN) for each f_{t_i}. The CNN outputs a 1d vector containing B features, $v_i = \{b_0, b_1, \ldots, b_B\}$, for each f_{t_i}. By respecting the temporal order, each v_i is then used to create the matrix, $M_t = \{v_0, v_1, \ldots, v_N\}$, whose size is $N \times B$, and represents the spatial features extracted over time in the sequence S_t. The set of features M_t is then processed by a stack of Long Short Term Memory (LSTM) networks [12]. At the end, the final set of Q features $P_t = \{x_1, x_2, ..., x_Q\}$ is extracted and classified as either anomalous or normal. The overall MARMnet structure is shown in Fig. 3.

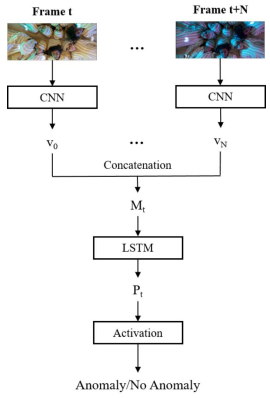


Fig. 3: MARMnet overall structure

The feature vector P_t can be interpreted as the mathematical representation of a potential anomaly considering the context of the scene. In fact, the spatial features allow to extrapolate information about the content of the frame, while the temporal feature are useful to map how this content evolves in the scene.

B. *Dataset Collection*

In our work, we collaborated with VQC operators in order to build two consistent datasets, one for COLOR and one for OLD QUALITY. Specifically, we extracted from around 5000 manually filled VQC reports, the video contents and the time codes at which the COLOR and OLD QUALITY were reported. Clearly, not every report presented the two anomalies we focused this paper on. Then, by leveraging on the expertise of VQC operators, we noticed that OLD QUALITY patterns

usually require 15 consecutive frames to be detected, while COLOR patterns require 10.

To extract the sequences, we had access to almost 700 videos for OLD QUALITY, while around 300 for COLOR. These videos are in 1920×1080 resolution at 30fps and every video is on average 30 seconds long. We collected 3355 sequences for COLOR, in which 1670 were affected by the anomaly. For the OLD QUALITY case, we collected 8965 sequences, in which 4090 were OLD QUALITY affected. Moreover, for the OLD QUALITY anomaly, on average 22 sequences were extracted from each video, while for COLOR around 17 sequences for video.

C. Transfer Learning

We considered two different settings for MARMnet, one for each anomaly. Both versions leverage the power of transfer learning for the feature extraction phase. This is due for two reasons: (1) pre-trained neural networks are able to extract general features (e.g. edges, shapes, etc.) in different contexts, and (2) the available data was not enough to properly train a custom CNN for feature extraction.

These models present some differences in their settings. Among these, the most important concerns the LSTM phase. In fact, compared to the OLD QUALITY version, the MARMnet setting for the COLOR anomaly presents less units in the LSTMs. The reason behind this choice is twofold. First fold, the sequence length is smaller for COLOR, so the feature extraction produces a smaller feature vector. Second fold, the dataset for COLOR is smaller. Therefore, to avoid an overfitting problem, we decided to decrease the number of units. The experiments in Section V describes the final settings of MARMnet for both anomalies.

V. MARMnet Setting Experiments

The experiments were conducted in order to determine the best setting of MARMnet for both anomalies. Specifically, the main goal was to find the setting that guarantees the best performance in terms of recall and precision, since they represent the target metrics to maximize in the context of the AVQC. Furthermore, it is important to take into account the inference time. In fact, the AVQC must be able to analyze a content in less than $2.5\times$ the duration of the video, which is the current duration spent by the operators.

To find the best setting of MARMnet, we took into account two of the most powerful pre-trained CNNs and applied the transfer learning technique for the feature extraction phase. We considered InceptionV3 [13] and Xception [14] since they provide high feature extraction capabilities as well as among the lightest computational requirements, which gives to MARMnet the ability to quickly take a decision on a sequence. Moreover, to consider also the inference time, we trained and tested MARMnet by considering, for each pre-trained neural network, two different portions of the convolutional base: (1) the very first half of the convolutional blocks, (2) the fully convolutional part of the pre-trained model. All in all, we had eight different settings (4 for each anomaly) for the feature extraction phase, and each of them produced a different set of features, both in size and content. Table I reports for each portion of the aforementioned CNNs, the size of the feature vectors generated.

Conv Base Portion	InceptionV3	Xception
Half	768	728
Full	2048	2048

TABLE I: Details on the feature vector size for each portion of the pre-trained models.

Clearly, since the features size strongly varies based on the feature extractor selected, we needed to analyze different architecture settings for the LSTM phase of MARMnet. Table II reports the different LSTM settings for the COLOR model and the OLD QUALITY model respectively. Specifically, the table reports the number of consecutive LSTM networks followed by the number of units within each network considering the different portions of CNN used for the feature extraction phase. Finally, every setting of MARMnet analyzed had a sigmoid activation layer that determines the presence of the anomaly.

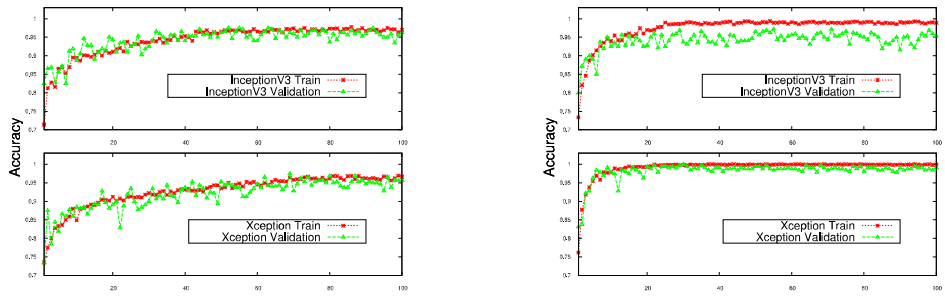
Anomaly	Conv Base Portion	# LSTM	# Units
COLOR	Half	2	$256, 64$
	Full	2	$512, 128$
OLD QUALITY	Half	2	$512, 128$
	Full	2	$1024, 256$

TABLE II: Details on the feature vector size for each portion of the pre-trained models.

Each model to evaluate was trained on a Dell Alienware Area 51 leveraging on two GPUs NVIDIA GeForce GTX 1080 Ti. Furthermore, we set the batch size to 64 and we trained the models for 100 epochs each. We split the dataset into 70% for training, 10% for validation, and the remaining 20% for testing. On average the training took 1.5hr for the OLD QUALITY version, while 1hr for the COLOR version.

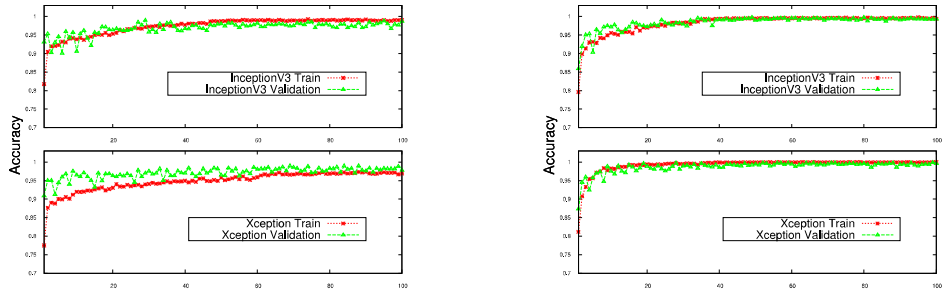
In Fig. 4 the training and validation accuracy over every epoch are reported for the MARMnet settings for the COLOR anomaly. Specifically, each sub-figure shows training and validation accuracies for InceptionV3 in the above graphs, while the one below shows how Xception performed. It can be seen that both convolutional bases offered positive results, since every MARMnet setting reached more than 95% of accuracy and did not show overfitting issues. Comparable behaviors were obtained also for the OLD QUALITY anomaly, whose reseults are shown in Fig. 5.

Tables III and IV report the metric scores obtained by every trained model on the testing dataset for both anomalies by setting a a prediction threshold to 0.5. Regarding the COLOR anomaly, the best setting for MARMnet is the combination of the full convolutional base of Xception followed by two LSTMs of 512 and 128 units respectively. In fact, such combination outperformed in both recall (97.4%) and precision all the other settings (99.1%). In the OLD QUALITY case, Table IV shows that the settings with the full CNN base of InceptionV3 and Xception was able to reach the maximum recall (98.4%). However, by employing only half of the InceptionV3 CNN base, MARMnet was able to reach



(a) Half portion (b) Full portion

Fig. 4: Training and validation accuracy over epochs for the MARMnet version for COLOR with the different portions of the convolution bases of InceptionV3 and Xception: (a) half CNN base, (b) full CNN base.



(a) Half portion (b) Full portion

Fig. 5: Training and validation accuracy over epochs for the MARMnet version for OLD QUALITY with the different portions of the convolution bases of InceptionV3 and Xception: (a) half CNN base, (b) full CNN base.

99.2% of precision and to outperform all the other versions. For this reason, we selected for the OLD QUALITY anomaly the setting of MARMnet that employs half of the InceptionV3 CNN base followed by two LSTMs of 512 and 128 units respectively.

Metric	Portion	InceptionV3	Xception
Recall [%]	Half	93.4	93.7
	Full	94.8	**97.4**
Precision [%]	Half	96.1	96.7
	Full	97.3	**99.1**

TABLE III: Score results on test set for the COLOR version of MARMnet.

Metric	Portion	InceptionV3	Xception
Recall [%]	Half	98	96.9
	Full	**98.4**	**98.4**
Precision [%]	Half	**99.2**	97.8
	Full	98.4	98.4

TABLE IV: Score results on test set for the OLD QUALITY version of MARMnet.

VI. MARMnet Qualitative Evaluation

To assess the ability and efficiency of our solution in a real-world scenario, we tested the AVQC capabilities in terms of inference time and prediction ability.

We had the AVQC processing around 1000 videos during our experiments, resulting in a report generation time of nearly $1\times$ the duration of the video. Additionally, the operators needed on average $0.3\times$ the content duration in order to check the predictions reported by the AVQC. This results in a quality control effort of $1.3\times$ the overall duration of the video, which is almost half of the current $2.5\times$.

In terms of prediction ability, MARMnet was tested on a dataset provided by the operators. For the OLD QUALITY version, the dataset consisted of 400 sequences containing normal and anomalous patterns, while the COLOR version was tested on 600 sequences. Overall, MARMnet was able to report every COLOR and OLD QUALITY sequence in the datasets. However, while for the COLOR cases the model presented strong capabilities on avoiding false positives, we noticed that the OLD QUALITY version of MARMnet slightly generated more cases of false alarms. Particularly, as shown in Fig. 6, we noticed that when some sequences present greenish patterns combined with strong white lights, MARMnet predicted them as anomalous. Also, if the sequence was bright and presented grayish patterns in some objects, like the castle in 6b, MARMnet was tricked and detected this sequence as an OLD QUALITY one. Nonetheless, MARMnet showed high ability in correctly learning patterns in this context. To this reason, we believe that by scouting other false positives, and training again the model, such alarms will be drastically reduced.

VII. Conclusion

In this paper, we take the initial steps towards the introduction of the Artificial Intelligence into the VQC process

(a) False alarm for the OLD QUALITY anomaly

(b) False alarm for the OLD QUALITY anomaly

Fig. 6: Two examples of false alarms generated by MARMnet for the OLD QUALITY anomaly. In a) the combination of green-ish colors and lights resemble some OLD QUALITY patterns for MARMnet. In b) the combination of a bright scene and the patterns in the castle makes MARMnet think this is an OLD QUALITY anomaly.

of the media and broadcasting companies. Specifically, we presented a novel neural network, called MARMnet, able to simulate the human operators behavior when performing the quality control of a content to broadcast. We designed and developed MARMnet in order to detect the COLOR and OLD QUALITY anomalies. MARMnet was trained and tested on a dataset built in collaboration with one of the main player in the broadcasting industry. The solution was tested in a real-world scenario giving promising results in terms of inference speed and prediction ability. In fact, MARMnet did not miss any anomaly in the video, while for the OLD QUALITY anomaly it generated few false alarms. However, we found that such cases can be easily investigated and inserted into the training dataset so that MARMnet is able to be more resilient to false alarms.

VIII. Future Work

As future work we plan to extend MARMnet for other types of video anomalies. Particularly, we are investigating its application in scenes that present frozen frame with sudden random audio noise. Currently, this anomaly requires a human intervention for its detection and the combination of MARMnet with an audio based model for detecting random noise could lead to a successful solution.

IX. Acknowledgement

This research was supported by NTT DATA Corporation. We thank our colleagues Davide Rezzonico, Silvia Cappelletto, Federico Ungolo, Stefano Turchetta, and Claudia Lunini who provided insight and expertise that greatly assisted the research.

References

[1] Television Broadcasting Global Market Report 2020. https://www.thebusinessresearchcompany.com/report/television-broadcasting-global-market-report, 2020.

[2] Video Streaming Market Size, Share Trends Analysis Report By Streaming Type, By Solution, By Platform, By Service, By Revenue Model, By Deployment Type, By User, And Segment Forecasts, 2020 - 2027. https://www.grandviewresearch.com/industry-analysis/video-streaming-market, 2020.

[3] How to Use Artificial Intelligence to Improve Quality Control. https://www.devteam.space/blog/how-to-use-artificial-intelligence-to-improve-quality-control/one, 2020.

[4] Using Artificial Intelligence to Improve Quality Control. https://www.qualitymag.com/blogs/14-quality-blog/post/94190-using-artificial-intelligence-to-improve-quality-control.

[5] How AI is Improving Quality in Manufacturing. https://industrytoday.com/how-ai-is-improving-quality-in-manufacturing/.

[6] Muhammad Shahid, Andreas Rossholm, Benny Lövström, and Hans-Jürgen Zepernick. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on Image and Video Processing*, 2014(1):40, Aug 2014.

[7] Hui Men, Hanhe Lin, and Dietmar Saupe. Spatiotemporal feature combination model for no-reference video quality assessment. In *2018 Tenth international conference on quality of multimedia experience (QoMEX)*, pages 1–3. IEEE, 2018.

[8] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. July 2017.

[9] Jianbo Yu, Xiaoyun Zheng, and Shijin Wang. A deep autoencoder feature learning method for process pattern recognition. *Journal of Process Control*, 79:1 – 15, 2019.

[10] Carlos A. Escobar and Ruben Morales-Menendez. Machine learning and pattern recognition techniques for information extraction to improve production control and design decisions. pages 286–300, 2017.

[11] Tao Zan, Zhihao Liu, Hui Wang, Min Wang, and Xiangsheng Gao. Control chart pattern recognition using the convolutional neural network. *Journal of Intelligent Manufacturing*, 31(3):703–716, Mar 2020.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.