# Deep multi-task learning for image/video distortions identification

Zoubida Ameur, Sid Ahmed Fezza, Wassim Hamidouche

# Deep Multi-Task Learning for Image/Video Distortions Identification

**Zoubida Ameur · Sid Ahmed Fezza · Wassim Hamidouche**

Received: date / Accepted: date

**Abstract** Identifying distortions in images and videos is important and useful in various visual applications, such as image quality enhancement and assessment techniques. Instead of applying them blindly, these techniques can be applied or adjusted depending on the type of distortion identified. In this paper, we propose a deep multi-task learning (MTL) model for identifying the types of distortion in both images and videos, considering both single and multiple distortions. The proposed MTL model is composed of one convolutional neural network (CNN) shared between all tasks and $N$ parallel classifiers, where each classifier is dedicated to identify a type of distortion. The proposed architecture also allows to adjust the number of tasks according to the number of distortion types considered, making the solution scalable. The proposed method has been evaluated on natural scene images and laparoscopic videos databases, each presenting a rich set of distortions. The experimental results demonstrate that our model achieves the best performance among the state-of-art methods for both single and multiple distortions[1].

**Keywords** Distortion · Identification · Multi-task learning · Deep learning · Natural image · Laparoscopic video · Multiple distortions · Scalability.

Z. Ameur and W. Hamidouche
Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
E-mail: firstname.lastname@insa-rennes.fr

SA. Fezza
National Institute of Telecommunications and ICT, Oran, Algeria
E-mail: sfezza@inttic.dz

[1] Code is available at: https://github.com/zoubidaameur/Deep-Multi-Task-Learning-for-Image-Video-Distortions-Identification

## 1 Introduction

Nowadays, multimedia content, especially images and videos, is used in a variety of fields and applications including healthcare, industry, security and social media. Images and videos go through several stages before being received and perceived by an end-user. The processing stages may deteriorate the image quality by introducing different kinds of distortion. This can occur at different stages including acquisition, processing, compression, transmission and storing. For instance, during acquisition, due to defocus or motion, blur could be introduced. At the compression stage, lossy compression codecs like JPEG can introduce a blocking artifact. Moreover, during transmission, some data packets may be lost, which introduces degradation during image or video reconstruction. Therefore, identifying distortions can be very useful for image enhancement and correction techniques. Instead of blindly applying such techniques [1–9], quality enhancement methods or other image processing algorithms can be performed or adjusted depending on the types of distortion determined. In addition, a priori knowledge of the type of distortion improves the performance of image quality assessment (IQA) metrics. Since such knowledge is essential in determining the most appropriate distortion-specific IQA metric [10–12], which are in most cases more efficient in specific distortions than universal metrics. Moreover, some IQA metrics take advantage of distortion information by providing useful information/features to the IQA algorithm [30–36].

The most common distortions encountered in real-word applications are blur, noise and blocking artifacts. Blur can have two forms: motion blur and defocus blur. The blur distortion can come from different sources, such as atmospheric turbulence, diffraction, optical de-

focusing and camera shaking [9]. Noise is a random variation of brightness or color information and can be introduced during image/video acquisition, processing and other practices [13]. Finally, blocking distortion may be more noticeable by lossy compression, especially at low bit rates. Such annoying distortion is mainly due to the use of block-based coding approach, where each block is transformed and quantized independently.

In addition, there are other types of distortions specific to certain fields of application. In laparoscopic surgery, for instance, the surgeon uses a tiny camera to project the internal organs of the abdomen on a monitor. The displayed laparoscopic videos may be distorted due to side effects of the instruments used or due to technical problems [14]. Smoke distortion may be introduced when using diathermy, while uneven illumination distortion may occur if the field of view is not evenly illuminated [15].

The human visual system (HVS) has the ability to distinguish the type of distortion in an image/video. It is therefore very interesting to develop an automatic method that mimic this HVS's capability. In recent years, much efforts have been devoted to the problem of identifying distortion in images and videos [18–34]. Most studies have focused on identifying a single type of distortion [26], however, images and videos can suffer from multiple distortions in most real-world applications [38], which is more challenging to address given the complex interactions and masking effect among distortions.

Considering the importance of identifying image and video distortions, in this paper, we propose a novel approach that performs the multi-distortion identification/classification task by decomposing it into several sub-tasks, where each sub-task is responsible for identifying a single type of distortion. Our proposed method is a deep multi-task learning (MTL) model that efficiently and accurately classifies single and multiple distortions. It consists of a pre-trained CNN shared between all tasks and $N$ separate parallel fully connected (FC) neural networks, where each FC neural network is responsible for identifying a specific type of distortion. The proposed method has been evaluated on natural scene image databases including popular IQA databases. Furthermore, an evaluation on laparoscopic videos database was conducted to show the efficiency and adaptability of our proposed framework to different applications.

The rest of this paper is organized as follows. In Section 2, we give an overview on distortion identification techniques. Section 3 provides a full description of the proposed method. In Section 4, the experimental results are presented and analyzed. Finally, Section 5 concludes the paper.

## 2 Related Work

In this section, we briefly review some of the previous works on identifying image distortion. These works are grouped according to whether the identification of distortion is considered as a single main task or as a secondary sub-task of an IQA process.

### 2.1 Distortion identification as a main task

Praneeth *et al.* [20] proposed content and perception-based features for efficient distortion classification, called as COPDIC. Given an input image, features are derived from local block level characteristics to classify common distortion types using local mean removal, divisive normalization followed by natural scene statistics (NSS) features extraction. Then, a pre-trained multi-stage support vector machine (SVM) classifier is used to identify the distortion type of this input image. Image quality metrics can be useful for distortion identification process by providing the most relevant visual features to the classifier. Chetouani *et al.* [18] proposed to use the output of image quality metrics (IQMs) as input for an artificial neural networks classifier to determine the distortion types within a given image. In [19], Gabor filters are applied to both clean reference and distorted images, then mean squared error (MSE) is calculated between them, the results represent the features used by a quadratic bayes normal classifier (QNBC) to classify the type of distortion present in the distorted image. However, since these last two described works use full-reference IQA metrics, they require the presence of the clean reference image to identify the type of distortion, which may not be practical in most real-word applications.

Namhyuk Ahn *et al.* [21] addressed the problem of distortion classification in an image without a reference image using CNNs architectures. In order to reach convergence quickly, they used pre-trained models on ImageNet dataset [39] including VGG 16 [40] and ResNet-101 [41]. They also created a new Flickr-Distortion dataset to train their model. Mateusz *et al.* [22] proposed ensemble learning method composed of two CNNs for distortion classification. The two CNNs architectures are similar in terms of number of layers and layout, but differ in the size and number of parameters used during the training. Both architectures contain three convolutional and max pooling layers followed by two FC layers. Their proposed solution out-

performs other SVM-based solutions by more than 10% in terms of accuracy. Bianco *et al.* in [23] proposed to first analyze features extracted from different layers of different deep neural network (DNN) architectures, then evaluate their relevance for distortion classification using clustering. The best features were then used for the recognition of the type of distortion. The obtained results showed that deep visual representations can be exploited even in an unsupervised way to efficiently recognize various types of image distortion.

### 2.2 Distortion identification as a secondary task

Distortion identification is useful in the process of IQA, either by helping to select the most appropriate metric for quality evaluation or by providing useful information about distortions for the metric itself. Therefore, distortion classification can be included in the process of IQA for better performance.

Falk *et al.* [24] proposed to select from a pool of full-reference IQMs those representing the most relevant features for each type of distortion. Then, they designed composite measures for each distortion type based on a linear combination of the selected features. Therefore, the quality of distorted image is evaluated and its distortion type is classified based on the distortion-specific features. Moorthy *et al.* [25] demonstrated that each distortion affects the statistics of natural images in a characteristic way. Thus, they build a classifier that can classify a given image into a particular distortion category solely on the basis of distorted image statistics (DIS) which are an extension of NSS. Based on this study, they proposed two no-reference IQMs: blind image quality index (BIQI) [26] and distortion identification-based image verity and integrity evaluation (DIIVINE) [27], which are based on a two-stages framework involving a distortion identification followed by a distortion-specific quality assessment. Peng *et al.* [28] also proposed a two-stage scheme for quality assessment. At the first stage, the image distortion type is predicted using a SVM. At the second stage, based on knowledge of distortion type, a fusion of three existing IQMs is performed using the K-nearest neighbors (KNN) regression. Chetouani *et al.* [29] proposed a framework for estimating image quality based on the assumption that there is no universal IQM that can efficiently estimate image quality across all distortion types. They proposed a method based on linear discriminant analysis (LDA) to classify the distortions before estimating image quality, where the classification stage uses quality scores derived from different IQMs applied on the reference image and its degraded version. The classification of distortions helps to determine the types of IQM that should be considered for the quality evaluation stage. In the same vein, Zohaib *et al.* proposed in [15] a distortion identification step followed by the quality evaluation for laparoscopic video. They used four distortion-specific classification methods, which consist of no-reference distortion-specific IQMs.

Considering the great successes achieved by deep CNN on various computer vision tasks, it was natural to adopt it for IQA and distortion identification tasks. Kang *et al.* [32] proposed a CNN model that simultaneously estimates image quality and identifies the type of distortion. The baseline structure of the proposed model for quality assessment, called as IQA-CNN, has one convolutional layer, one pooling layer, two FC layers and one output layer. It was then extended for the distortion classification by adding a classification layer and was referred to as IQA-CNN+ and later enhanced to IQA-CNN++ [42]. In this case, the quality estimation is considered as the main task while distortion identification as the secondary task. Wang *et al.* [33] also proposed a CNN-based approach to identify the type of distortion and assess the quality without the clean reference image. The parameters of the proposed CNN model were learned to fit both tasks. Kede *et al.* [34] proposed a CNN-based IQM which consists of two sub-networks: 1) a distortion identification network and 2) a quality prediction network, sharing the early layers. First, the distortion identification sub-network is trained, then, the quality prediction sub-network is trained starting from the pre-trained early layers and the output of the first sub-network. The same idea has been followed in [35] and [36], for instance Huang *et al.* [35] proposed a full reference IQM named mask gated convolutional network (MGCN) that evaluates the image quality score and identifies distortions simultaneously. In the MGCN metric, an encoder block is designed to encode the transformation between reference and distorted images as low level features. Next, the abstractor extracts high level features from the low level features. Finally, the high level features are exploited by the predictor which predicts the image quality score and classifies distortions simultaneously.

The above described works have investigated distortion identification and achieved considerable results. However, the proposed solutions suffer from three major drawbacks. First, most of the proposed solutions are knowledge-driven approaches, which means that the feature descriptors must be designed manually such as: NSS, DIS or even IQMs. Second, in many methods, the classification of distortion is seen as a secondary task or as a sub-task of IQA, while it is also important to design a standalone distortion identification algorithm considering its utility in various applications, including

image enhancement and restoration techniques. Third, the proposed classifiers are designed to consider a limited and predefined number of distortions, which makes them not scalable and not general-purpose distortion classifiers.

Our proposed solution has been designed to overcome these drawbacks with the following main contributions:

- Our proposed method is a data-driven approach, which is based on a CNN model trained in an end-to-end manner responsible for extracting the most relevant features without any hand-engineering.
- Our proposed solution only performs distortion identification as the main and unique task. The proposed MTL approach classifies several distortion types by performing several tasks, where each task is responsible for identifying a single and specific distortion type.
- The number of types of distortion considered by our proposed MTL model is adjustable and scalable with respect to both data and application.

## 3 Proposed Method

The aim of our work is to develop a robust and reliable method for the identification of distortions without needing to access the clean reference image. To achieve that, we propose a deep multi-task learning (MTL) approach that identifies and classifies distortions in both images and videos. We first give a formal definition of MTL and the notations used, then we describe in detail our proposed method.

### 3.1 Multi-task learning: definition and notations

Let us consider $N$ related learning tasks $\{\mathcal{T}_i\}_{i=1}^{N}$, MTL aims to enhance the learning of a model $\mathcal{M}$ of task $\mathcal{T}_i$ by exploiting the knowledge contained in all or a subset of the $N$ tasks [16]. In the case of supervised learning, a task $\mathcal{T}_i$ is associated to a training dataset $\mathcal{D}_i$ composed of $M$ training samples $\mathcal{D}_i = \{\boldsymbol{z}_j^i, y_j^i\}_{j=1}^{M}$ with $\boldsymbol{z}_j^i \in \mathbb{R}^{d_i}$ is the jth training sample of dimension $d_i$ and $y_j^i$ its label. The training dataset of a task $\mathcal{T}_i$ can be represented by a matrix $\boldsymbol{Z}^i = \{\boldsymbol{z}_1^i, \ldots, \boldsymbol{z}_M^i\} \in \mathbb{R}^{M \times d_i}$. We can distinguish homogeneous MTL that, in opposite to heterogeneous MTL, consists in tasks of one type i.e., classification or regression [44]. On the other hand, according to the dimension $d_i$ of the feature space of input samples $i$, we can distinguish heterogeneous-feature MTL from homogeneous-feature MTL implying that the input samples have the same dimension $d_i = d_j,\ \forall i \neq j$. In this paper, our distortion identification problem is modeled as a homogeneous MTL for classification tasks and homogeneous-feature MTL with the same input image dimension or patches dimension. Researchers in MTL address three main issues of when to share, what to share, and how to share [16]. To answer the first issue of when to share, MTL is motivated in this paper by the strong relation among distortions identification in a single input image. The what to share and how to share, specifying the knowledge to share and the concrete way to share the knowledge, respectively, are both addressed in detail in the next sections.

### 3.2 Proposed multi-task learning model

Our proposed model $\mathcal{M}$ takes an image $I$ as input and predicts its distortion types. The proposed model performs $N$ tasks, where each task is responsible for identifying a specific distortion type. Therefore, given an input image $I$, our model outputs $N$ values, where each value represents the probability that $I$ contains a certain distortion type. The proposed model can be formulated as follows:

$$\{\hat{p}_1, \hat{p}_2, ..., \hat{p}_N\} = \mathcal{M}(I), \tag{1}$$

where $\hat{p}_i$ denotes the probability of the presence of a distortion $d_i$ in the image $I$ estimated by the model $\mathcal{M}$ with $i \in \{1, \ldots, N\}$, and $N$ represents the total number of distortion types considered and therefore the number of tasks performed by our MTL model.

Figure 1 shows the flowchart of the proposed approach. The model $\mathcal{M}$ is based on MTL framework which typically consists of a shared part that branches out into task-specific heads [43]. The shared part is a feature extractor $f_\theta(I)$ that takes as input an image $I$ and outputs a vector of features $\boldsymbol{x} \in \mathbb{R}^F$, with $F$ the number of features and $\theta$ represents the parameters of the shared CNN. The task-specific heads are classifiers $g_{\phi_i}(\boldsymbol{x})$ that takes as input the same vector of features $\boldsymbol{x}$ but each identifies a specific type of distortion, where $\phi_i$ stands for the parameters of the classifier. The feature extractor $f_\theta(I)$ consists of a pre-trained CNN on ImageNet dataset [39], while all classifiers consist of a network of FC layers randomly initialized.

Overall, MTL aims to improve the generalization performance of multiple prediction tasks by appropriately sharing relevant information across them. Given that the distortion identification is performed on the same input and CNNs are known to be efficient in extracting the most salient and relevant features, we chose to share the CNN among all tasks, so the features in the shared layers do not need to be recalculated for the

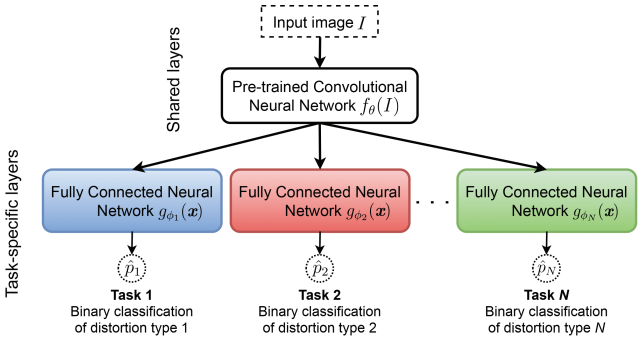



**Fig. 1** The framework of our proposed approach for identifying image distortions.

different tasks. However, since each distortion can be identified based on the presence or absence of certain features, the classifiers are kept task-specific. The objective of each classifier is to take the results from the shared CNN and use them to classify the image into a distortion-class. Thus, each network of FC layers will compile the data extracted by the shared CNN to form the final output representing a probability.

The pre-trained CNN used in the feature extraction block is DenseNet 169 [45], for which we obtained the best results, however, any other CNN architecture can be considered. In the DenseNet 169 network, in each layer, information is obtained from all preceding layers, i.e., feature maps from all previous layers are concatenated, processed and passed to all subsequent layers. This technique encourages the reuse of features, reduce the number of parameters and strengthen the features propagation. DenseNet 169 architecture consists of convolutional layers, pooling layers, dense blocks, transition blocks and FC layers, as shown in Figure 2. A dense block consists of $1 \times 1$ convolution layer to reduce feature maps followed by $3 \times 3$ convolution layer to improve the computational efficiency. Between each two contiguous dense blocks, there is a transition block which consists of $1 \times 1$ convolution layer and a pooling layer to reduce the number of feature maps.

In order to use the DenseNet 169 as feature extractor, the last dense layers are removed and the output of the last dense block is flattened to form a vector of features $\boldsymbol{x}$. Given an input image $I$, the shared layers $f_\theta(I)$ responsible for extracting relevant features outputs a vector of features $\boldsymbol{x}$

$$\boldsymbol{x} = f_\theta(I). \tag{2}$$

The number of tasks that our model performs is fixed according to the number of distortion types to be detected. If $N$ types of distortion are addressed in the considered application, our model performs $N$ binary classification tasks using $N$ different classifiers. Each classifier $g_{\phi_i}(\boldsymbol{x})$ is composed of two stacked FC layers of size 512 and an output layer of size 1. The last output layer of each classifier has a sigmoid activation layer that outputs values between $[0, 1]$ representing the probability $\hat{p}_i$ indicating the presence of a distortion $d_i$ in the input image $I$

$$\hat{p}_i = g_{\phi_i}(\boldsymbol{x}),\ \forall i \in \{1, \ldots, N\}. \tag{3}$$

The end-to-end model is illustrated in Figure 2 and can also be formulated as follows:

$$\{\hat{p}_1, \ldots, \hat{p}_N\} = \{g_{\phi_1} \circ f_\theta(I), \ldots, g_{\phi_N} \circ f_\theta(I)\}. \tag{4}$$

### 3.3 Input and output processing

The image datasets adopted in this study are not large enough to train the deep DenseNet 169 from scratch, hence the choice of using a pre-trained DenseNet 169 on ImageNet. To further compensate the lack of training samples while respecting input dimensions, we chose to perform patch-wise training. In patch-wise training, the CNN is trained on a small patch of the image instead of the whole image. Therefore, given an input image $I$, a fixed number of patches of size $224 \times 224$ is extracted over which our MTL model loops. Each patch is treated individually, this means that for each patch $k$, each classification block $g_{\phi_i}(\boldsymbol{x})$ outputs a probability $\hat{p}_{i,k}$ indicating the presence of a specific distortion $d_i$ in this patch. An average of all predictions is computed and rounded at the end of each classifier to label the image as follows:

$$\hat{p}_i = \frac{1}{K} \sum_{k=1}^{K} \hat{p}_{i,k}, \tag{5}$$

where $K$ denotes the total number of patches extracted from image $I$.

### 3.4 Loss function

In multi-task learning, a joint loss function must be defined for several tasks. While a single task has a well-defined loss function, multiple tasks result in multiple losses. Considering a MTL model that performs $N$ tasks with task-specific loss functions noted as $\mathcal{L}_i$ and task-specific weights noted as $\omega_i$, the loss function of the MTL model is expressed as follows:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{i=1}^{N} \omega_i \mathcal{L}_i. \tag{6}$$

All tasks in our case perform a classification, binary cross-entropy (BCE) is the default loss function to use

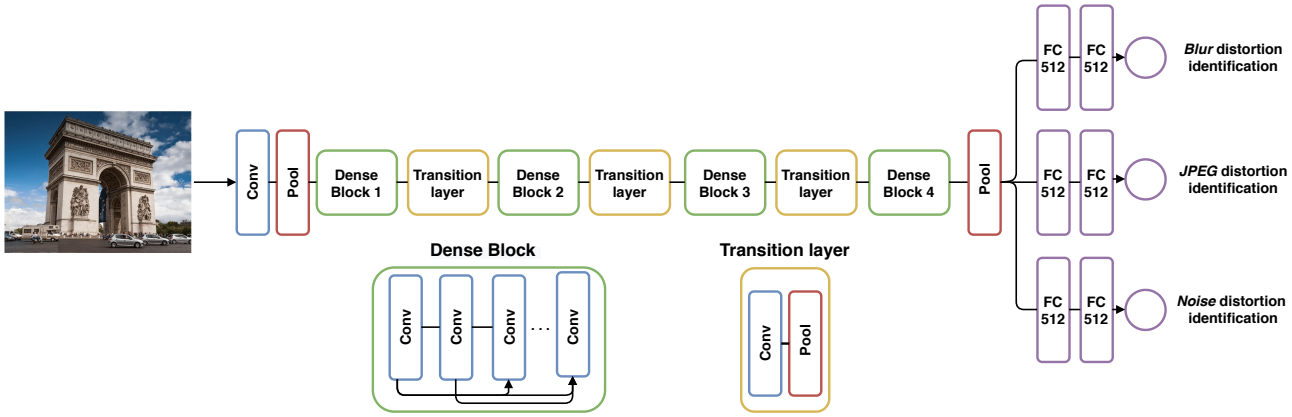



**Fig. 2** Deep multi-task learning model for the classification of distortions in natural scene images. Given an input image patch, our MTL model passes the patch through the CNN DenseNet 169 to extract its most relevant features in the form of a vector that will be fed into the different classifiers, then each classifier (FC layers network) is fed with this vector to output a value indicating the probability of the presence of a specific distortion.

for binary classification problems where the target values are in the range of $[0, 1]$. Thus, we adopted the task-specific loss function

$$\mathcal{L}_i = -\frac{1}{M}\sum_{m=1}^{M} p_i^{(m)} \log(\hat{p}_i^{(m)}) + (1-p_i^{(m)}) \log(1-\hat{p}_i^{(m)}), \quad (7)$$

where $p_i^{(m)}$ is the ground truth label and $\hat{p}_i^{(m)}$ is the predicted probability of the distortion $d_i$ at an image sample $m$, while $M$ represents the number of images per batch.

One of the challenges of jointly learning multiple tasks is to properly weighting the task-specific loss functions [46, 47]. In our case, since the loss scales are the same and task importance is the same, a uniform weighing is considered. The total loss is then an uniform weight sum of the $N$ loss functions, i.e., each task is assigned a loss weight of $\omega_i = 1/N$, $\forall i \in \{1, \ldots, N\}$.

Since each image consists of $K$ patches, the single loss function is expressed as follows

$$\begin{aligned}\mathcal{L}_i(p_i, \hat{p}_i) = &-\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} p_{i,k}^{(m)} \log\left(\hat{p}_{i,k}^{(m)}\right) \\ &+(1-p_{i,k}^{(m)}) \log\left(1-\hat{p}_{i,k}^{(m)}\right).\end{aligned} \quad (8)$$

### 3.5 Temporal extension

In order to show the efficiency and adaptability of our proposed method to various and multiple distortion types encountered in different applications, we also considered another challenging issue which is distortion classification in laparoscopic videos. It is important to note that we addressed this application as part of a challenge organized in a leading international conference, for which we obtained the 3rd place in grand challenge[2].

Laparoscopic videos are used in real-time, which requires a real-time distortion classifier. Therefore, we extended our model to cover the temporal aspect by applying some modifications to the preprocessing steps and the architecture of the model. Specifically, in order to perform classification in real-time, the CNN used as features extractor which is DenseNet 169 has been substituted by VGG16 [40], because it is less deep and consequently takes less time in processing the input. In addition, the task-specific part contains five parallel classifiers, because the provided laparoscopic video quality (LVQ) dataset covers five distortion types [15], including noise, defocus blur, motion blur, smoke and uneven illumination. The end-to-end MTL model is then composed of a shared pre-trained VGG16 followed by five parallel classifiers each responsible for identifying a single type of distortion as shown in Figure 3.

For hardware constraints and more precisely the limits in terms of memory, patch extraction cannot be done on video frames. Instead, frames extraction is performed which allows to use the model as if it was fed with input images and at the same time increase the number of training samples. Since the distortions are applied to all frames of a given reference video when building the dataset [15], therefore, frames of the same video are assigned the same labels. Given an input video $V$, a fixed number of frames is extracted then fed into the MTL model. Each frame is treated individually,

[2] Real-time Distortion Classification in Laparoscopic Videos – ICIP 2020 Challenge: https://2020.ieeeicip.org/challenge/


**Fig. 3** Deep multi-task learning model for the classification of distortions in laparoscopic videos. Given an input video frame, our MTL model passes the frame through the CNN VGG16 to extract its most relevant feeatures in the form of a vector that will be fed into the different classifiers, then each classifier (FC layers network) is fed with this vector to output a value indicating the probability of the presence of a specific distortion.

then an average of all predictions is computed and rounded to label the video.

## 4 Experiments

In this section, we first introduce the datasets used to train and evaluate the proposed MTL model. Next, we evaluate and compare the performance of our model with respect to state-of-the-art methods on the different datasets. Finally, we conduct a series of ablation experiments to analyse the behavior of the proposed method.

### 4.1 Datasets

#### *4.1.1 Natural scene image datasets*

In order to evaluate the proposed method on natural scene images, four publicly available image quality databases are exploited: KonstanzLin artificially distorted image quality database (KADID-10K) [49], computational and subjective image quality (CSIQ) [48], Tampere image dataset (TID2013) [13] and laboratory for image & video multi distortion (LIVEMD) [37]. KADID-10K contains 81 pristine images of size $512 \times 384$, each degraded by 25 distortions at five severity levels. CSIQ is composed of 866 images that were obtained from 30 clean original images of size $512 \times 512$ using six distortion types and four to five severity levels. TID2013 provides 25 reference images of size $512 \times 384$ and 3000 distorted images at five levels of distortion. Unlike KADID-10K, CSIQ and TID2013 datasets, which are single distortion datasets, LIVEMD is a multiply distorted image dataset, i.e., it contains distorted images counting multiple distortion types simultaneously. LIVEMD dataset [37] contains 15 reference images from which two subsets of distorted images are created using three distortion types, including blur, JPEG and noise. The first subset is obtained by applying blur followed by JPEG, each at three different severity levels. In the same way, the second subset is obtained by applying blur followed by noise. This results in a total of 450 distorted image.

Table 1: Features of the considered natural scene image datasets. All datasets constrain three types of distortion: {blur, noise and JPEG}. Multi-distorted images of blur-JPEG and blur-noise are also included in the LIVEMD dataset resulting in five calsses.

| Dataset | Number of images | Number of classes | multi-distortion |
|---|---|---|---|
| KADID-10K [49] | 1215 | 3 | ✗ |
| CSIQ [48] | 450 | 3 | ✗ |
| TID2013 [13] | 375 | 3 | ✗ |
| LIVEMD [37] | 405 | 5 | ✓ |

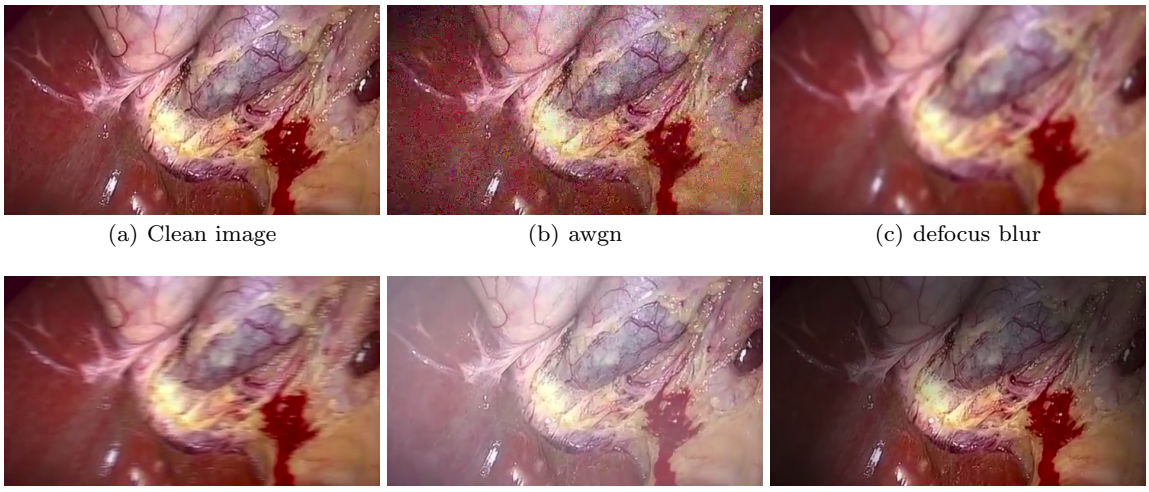
We considered three distortion types that are common in the three databases: (a) JPEG compression (JPEG), (b) white noise contamination (noise) and (c) Gaussian blur (blur). This leads us to use 1215, 450, 375 and 405 images respectively from the KADID-10K, CSIQ, TID2013 and LIVEMD datasets. Table 1 summarises the main features of the considered natural images datasets.

Figure 4 illustrates a clean reference image from LIVEMD dataset and its distorted versions, including blur, JPEG, noise, blur-JPEG and blur-noise distor-

(a) Clean image (b) blur (c) JPEG

(d) noise (e) blur-JPEG (f) blur-noise

**Fig. 4** A clean reference image and its distorted versions, from LIVEMD database, containing blur, JPEG, noise and combinations of blur-JPEG, blur-noise distortions.



(a) Clean image (b) awgn (c) defocus blur

(d) motion blur (e) smoke (f) uneven illumination

**Fig. 5** An example of a frame extracted from laparoscopic video and its distorted versions covering each distortion type in the LVQ dataset.

tions. As mentioned before, these distortions are the most common and widely studied types of distortion.

#### *4.1.2 Laparoscopic video dataset*

Laparoscopic video quality dataset [15] consists of a total of 20 reference videos, each of 10 seconds duration. These videos were extracted from Cholec80 dataset [50]. The selection of videos was made to include maximum possible variations of scene content and temporal information. For scene content, ten different categories were chosen: bleeding, grasping and burning, multiple instruments, irrigation, clipping, stretching away, cutting, stretching forward, organ extraction and burning. Each reference video is distorted by five different kinds of distortion (single or multiple distortions) with four different levels, resulting in a total of 800 videos. The resolution of videos is $512 \times 288$ with a 16:9 aspect ratio and a frame-rate of 25 frames per second (FPS). Figure 5 shows the different types of distortion included in LVQ dataset. These distortions are the most common affecting laparoscopic videos.

### 4.2 Implementation details

For natural scene image datasets, the training is performed with back-propagation using ADAM optimiza-

tion [51,52] with a learning rate of 0.0001, a $10^{-6}$ decay and momentum of 0.9. A dropout is applied after each FC layer of size 512 with a rate equals to 0.25. Dropout is a regularization technique used to reduce overfitting in artificial neural networks. It consists of randomly ignoring some nodes of a given layer during training. This method forces each node to assume different degrees of responsibility for the inputs as they are either activated or dropped out randomly during the training process. This has the effect of making the training process noisy, thus avoiding overfitting [53].

For an end-to-end training, the model is trained iteratively by back-propagation over 60 epochs. In each epoch, the training set is divided into batches for batch-wise optimization. Each mini-batch contains six images ($M = 6$), each represented by a fixed number of patches ($K = 4$ for CSIQ and TID2013, $K = 6$ for KADID-10K and $K = 8$ for LIVEMD) of size $244 \times 224$, thus leading to an effective batch size of $M \times K$ patches.

For laparoscopic video dataset, in each epoch, the training set is divided into batches of videos for batch-wise optimization. Each mini-batch contains 10 videos, each represented by ten frames ($M = 100$), thus leading to the effective batch size of $M \times K = 100$ frames ($K = 1$).

## 4.3 Evaluation procedure

In order to train and evaluate our proposed MTL model, we randomly split each dataset into two subsets of non overlapping content, 80% for training and 20% for testing. This procedure is repeated 10 times and the median of overall accuracy, precision, recall and F1-score values are reported and used to evaluate the performance of our model on the test set.

Precision is defined as the number of true positives (TPs) divided by the number of TPs plus the number of false positives (FPs) as follows :

$$precision = \frac{TP}{TP + FP}, \tag{9}$$

in our case, TP is when the model perfectly identifies the presence of a distortion type. While, FP is when the model confuses the presence of a distortion type with another.

Recall is defined as the number of TPs divided by the number of TPs plus the number of false negatives (FNs) as follows :

$$recall = \frac{TP}{TP + FN}, \tag{10}$$

in our case, FN is when the model mistakenly predicts the absence of a distortion type. Finally, the F1-score is calculated as follows:

$$F1\text{-score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \tag{11}$$

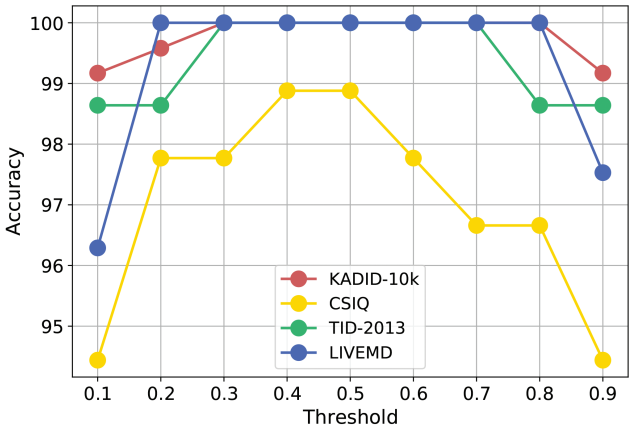



**Fig. 6** Classification accuracy with respect to different threshold values for all considered datasets.

## 4.4 Experimental Results

### *4.4.1 Classification threshold*

In the proposed model, the last layers are assigned sigmoid activation functions which make them output a value between 0 and 1 representing the probability that an image contains a specific distortion. The decision for converting the predicted probability into a class label is governed by a parameter referred to as the *decision threshold*, *discrimination threshold* or simply the *threshold*, which typically has a default value of 0.5. In the case of a binary classification with class labels 0 and 1, values below the threshold are assigned to class 0 and values greater than or equal to the threshold are assigned to class 1. In our case, our MTL model performs $N$ binary classifications such that each of them serves to identify a specific type of distortion, so that 0 indicates the absence of distortion, while 1 indicates its presence. The default threshold may not represent an optimal interpretation of the predicted probabilities, so they must always be adapted to the problem under consideration, and therefore adjust the threshold value accordingly. Hence, we studied the performance of our model, in all datasets, according to the values of the threshold.

Figure 6 shows the accuracy of our classification model as a function of the threshold. The threshold has been varied within a range of $[0.1, 0.9]$ with a step of 0.1. The accuracy is then calculated according to this hard threshold. From this figure, one can notice that

Table 2: Overall accuracy on single and multi-distortion image datasets. The top result is highlighted in boldface.

| Method | Dataset | | | |
|---|---|---|---|---|
| | KADID-10K | CSIQ | TID2013 | LIVEMD |
| BRISQUE [54] | 91.71 | 80.54 | 88 | 88.56 |
| DIIVINE [27] | 49.44 | 56.24 | 55.28 | 74.71 |
| IQM-G | 90.80 | 79.57 | 82.66 | 88.67 |
| COPDIC [20] | 90.57 | 90.96 | 88.25 | 90.64 |
| Mateusz [22] | 92.45 | 91.08 | 89.32 | 85.05 |
| Kang [32] | 93.64 | 90.71 | 92.26 | 86.14 |
| Kede [34] | 98.52 | 97.03 | 95 | 88.57 |
| Golestaneh [36] | 97.68 | 93.67 | 92.53 | 95.13 |
| MGCN [35] | 96.96 | 97.10 | 94.79 | 88.23 |
| Bianco [23] | - | 79.60 | 85.40 | 90.90 |
| Our model | **100** | **98.88** | **100** | **100** |

our model achieves relatively low performance for very low or very high threshold values, this is particularly clear for CSIQ and TID2013 datasets. Because a very low threshold value makes the model really sensitive, while a very high threshold value makes it indifferent for a wide range of probabilities. Our model achieves its best performance for a threshold value between 0.4 and 0.5. Therefore, a threshold value of 0.5 was adopted and the results provided below were obtained based on this threshold value.

### *4.4.2 Comparison and discussion*

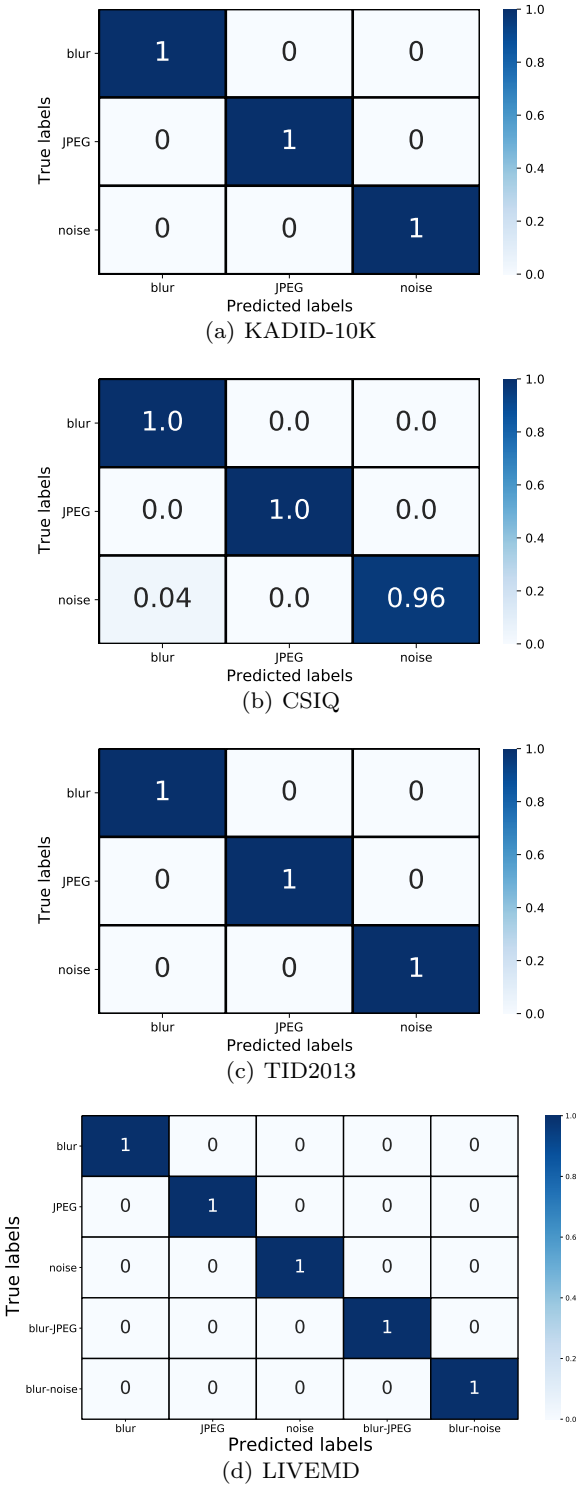
The performance of our model is evaluated and compared to ten state-of-the-art methods, including distortion identification-based image verity and integrity evaluation (DIIVINE) [27], blind/referenceless image spatial quality evaluator (BRISQUE) [54], distortion-specific IQMs, COntent & Perception based features for DIstortion Classification (COPDIC) [20], Mateusz's method [22], Kang's method [32], Kede's method [34], Golestaneh's method [36], mask gated convolutional network (MGCN) [35] and Bianco's method [23]. The first four methods are based on hand-crafted features that are fed into a SVM classifier. For instance, the third method (noted as IQM-G) groups three distortion-specific IQMs, so that each of them quantifies a specific distortion (blur, JPEG and noise) in a given input image. To quantify blur distortion, an algorithm designed to measure local perceived sharpness in images relying on both spectral and spatial properties is used [55]. To quantify JPEG distortion, a no-reference quality measurement algorithm for JPEG compressed images is exploited [12]. To quantify noise distortion, a technique to estimate the noise level proposed in [56] is adopted. The scores obtained from each of these three distortion-specific IQMs are used to form a vector of features that will be introduced in a SVM to perform the classification of distortion. The remaining six methods are deep learning-based approaches.

The comparison of our model with state-of-the-art methods is first performed in terms of accuracy, as reported in Table 2. From this table, we can notice that our model outperforms all considered methods in all datasets, this is particularly remarkable for LIVEMD dataset, which is the most challenging. In addition, BRISQUE provides fairly good results for a handcrafted method, while DIIVINE performance is poor on all datasets. IQM-G method also manages to provide quite good results on the different datasets, which means that the three distortion-specific IQMs chosen as feature extractors are complementary and their combination succeeds to discriminate the distortion type. Each of the handcrafted-based methods provides balanced results on the different datasets, which means that their performance is not strongly tied to the content of any dataset. However, their results generally remain inferior to those achieved by deep learning-based methods. We can also notice that the Kede [34] and Golestaneh [36] methods achieve the best performance among deep learning-based approaches, but without going beyond the proposed method.

Overall, it can be noted that handcrafted methods have proven that they can be competitive with deep learning-based methods if the extracted features are relevant for distortion identification and adapted to the datasets. The importance of the feature extraction step is also well illustrated through the performance of BRISQUE, DIIVINE and IQM-G methods, since all of them use a SVM classifier but provide different accuracy, because each one extracts different types and number of features. The difficulty with such approaches is choosing features that cover all types of distortion. However, since our method is a data driven approach, the selection of distortion-specific features is done automatically across the different towers, i.e., the classifiers specializing in a single distortion type.

In addition, we computed the precision and recall for each distortion type separately. Precision for all considered datasets is reported in Tables 3, 4, 5 and 6, while recall scores are reported in the diagonal of the normalized confusion matrices plotted in Figure 7. From Table 3, we can notice that our model obtained 100% precision for all the distortions considered. Likewise, most of the deep learning-based methods provide good results, this is due to the fact that KADID-10K is the dataset with the largest number of distorted images, thus allowing to better learn the features of each distortion type.

From Table 4, we can notice that our model delivered 100% precision for JPEG and noise distortions




(a) KADID-10K

(b) CSIQ

(c) TID2013

(d) LIVEMD

**Fig. 7** Confusion matrices on natural scenes image datasets.

on CSIQ dataset, while obtained 96.66% for blur distortion. Also, IQM-based metric obtained a precision of 100% for noise distortion on CSIQ dataset, while provided low performance for blur and JPEG distortions. Moreover, DIIVINE method provided the lowest

Table 3: Precision per distortion type on KADID-10K dataset. The top result is highlighted in boldface.

| Method | Distortion | | |
|---|---|---|---|
| | blur | JPEG | noise |
| BRISQUE [54] | 87.65 | 90.5 | 98.9 |
| DIIVINE [27] | 53.65 | 43.03 | 50.31 |
| IQM-G | 80.75 | 92.3 | **100** |
| COPDIC [20] | 92.18 | 87.88 | 93.04 |
| Mateusz [22] | 94.57 | 90.21 | 89.95 |
| Kang [32] | 96.88 | 93.47 | 90.49 |
| Kede [34] | 98.54 | **100** | 97.06 |
| Golestaneh [36] | 98.55 | 97.71 | 98.29 |
| Our model | **100** | **100** | **100** |

Table 4: Precision per distortion type on CSIQ dataset. The top result is highlighted in boldface.

| Method | Distortion | | |
|---|---|---|---|
| | blur | JPEG | noise |
| BRISQUE [54] | 83.3 | 63.33 | 96.66 |
| DIIVINE [27] | 69.83 | 30.30 | 68.23 |
| IQM-G | 78.12 | 60.58 | **100** |
| COPDIC [20] | 87.29 | 91.16 | 90.41 |
| Mateusz [22] | 89.94 | 94.55 | 90.75 |
| Kang [32] | 91.57 | 86.46 | 93.19 |
| Kede [34] | 98.50 | 95.31 | 94.48 |
| Golestaneh [36] | 93.40 | 95.52 | 90.96 |
| MGCN [35] | 95.61 | 97.72 | 97.62 |
| Our model | **96.66** | **100** | **100** |

Table 5: Precision per distortion type on TID2013 dataset. The top result is highlighted in boldface.

| Method | Distortion | | |
|---|---|---|---|
| | blur | JPEG | noise |
| BRISQUE [54] | 92.85 | 75 | **100** |
| DIIVINE [27] | 64.53 | 49.09 | 50.87 |
| IQM-G | **100** | 60.71 | 92.85 |
| COPDIC [20] | 90.84 | 85.27 | 86.69 |
| Mateusz [22] | 85.21 | 89.50 | 91.74 |
| Kang [32] | 91.63 | 93.04 | 90.71 |
| Kede [34] | 94.93 | 96.28 | 95.29 |
| Golestaneh [36] | 94.81 | 90.36 | 89.18 |
| MGCN [35] | 98.02 | 91.26 | 90.89 |
| Our model | **100** | **100** | **100** |

performance, whereas BRISQUE performed well except for JPEG distortion. In addition, Kede [34], Golestaneh [36] and MGCN [35] methods obtained high and stable results for all the distortions of CSIQ dataset.

From Table 5, we can observe that some methods achieved high precision up to 100% for a specific distortion type. This means that features of these methods are well-designed for a specific distortion type but do not generalize well to all distortion types. On the other hand, the proposed method provided very stable results for all distortion types considered.

Table 6: Precision per distortion type on LIVEMD dataset. The top result is highlighted in boldface.

| Method | Distortion | | |
|---|---|---|---|
| | blur | JPEG | noise |
| BRISQUE [54] | 78 | 91.82 | 92.18 |
| DIIVINE [27] | 88.48 | 62.17 | 65.46 |
| IQA-G | 80.47 | 79.04 | 92.76 |
| COPDIC [20] | 85.53 | 89.61 | 92.68 |
| Mateusz [22] | 82.42 | 90.34 | 84.55 |
| Kang [32] | 86.05 | 84.82 | 88.17 |
| Kede [34] | 89.67 | 85.63 | 90.28 |
| Golestaneh [36] | 93.50 | 96.57 | 93.84 |
| MGCN [35] | 90.08 | 84.26 | 88.22 |
| Our model | **100** | **100** | **100** |

Table 7: F1-score for each distortion type on natural scene image datasets.

| Dataset | Distortion | | |
|---|---|---|---|
| | blur | JPEG | noise |
| KADID-10K | 100 | 100 | 100 |
| CSIQ | 100 | 100 | 98.18 |
| TID2013 | 100 | 100 | 100 |
| LIVEMD | 100 | 100 | 100 |

Finally, from Table 6, we can notice a significant decrease in performance of all methods, except DIIVINE, compared to their performance on the other datasets. Because, LIVEMD dataset contains images with multi-distortion types making them difficult to recognize. Despite this, our model succeed in recognizing perfectly all types of distortion thanks to the adoption of the MTL architecture.

#### *4.4.3 Distortion identification confusion*

For a complete evaluation of our model, normalized confusion matrices are plotted to discuss mis-classified distortions. Each row in a confusion matrix represents an actual label which we call *true label*, while each column represents a predicted label. The purpose behind plotting confusion matrices is to show the number occurrences or probability for a class being classified as another one. Figure 7 depicts normalized confusion matrices of our model on the four considered datasets, from which we can notice that the proposed model always perfectly identified and classified distortion types of all datasets, except for CSIQ dataset, where noise distortion is confused with blur distortion.

Because neither precision nor recall, alone, gives all the information on the performance of a model, F1-score of our model for each distortion type is computed and provided in Table 7. The good result reported in this table shows that our model is accurate and robust in the classification of distortions.

Table 8: Classification accuracy using different pre-trained CNN architectures.

| | CSIQ | TID2013 | KADID-10K | LIVEMD |
|---|---|---|---|---|
| VGG16 | 87.77 | 100 | 99.58 | 97.53 |
| VGG19 | 83.33 | 98.64 | 99.58 | 100 |
| ResNet50 | 91.11 | 100 | 100 | 98.76 |
| DenseNet 169 | 98.88 | 100 | 100 | 100 |

#### *4.4.4 Ablation experiments*

To investigate the effectiveness of our model, we conduct a series of ablation experiments. First, different pre-trained CNN architectures have been considered as feature extraction block, including VGG16, VGG19, ResNet50 and Densenet. Table 8 shows the classification performance of each architecture on the four datasets. It is clear that the using DenseNet architecture as a features extractor provides the best performance. Furthermore, in addition to providing the highest performance, this architecture is the one with the lowest number of parameters, thus requiring less memory and computational resources.

Second, in Table 9 and Figure 8, we show the classification performance when the proportions of training data vary between 20% and 80% on the four datasets considered. We can conclude that the classification performance is not strongly dependent on the size of training data. From 75 images as a training dataset, we can achieve a high classification accuracy of over 80% for all datasets, which shows the efficiency of our model.

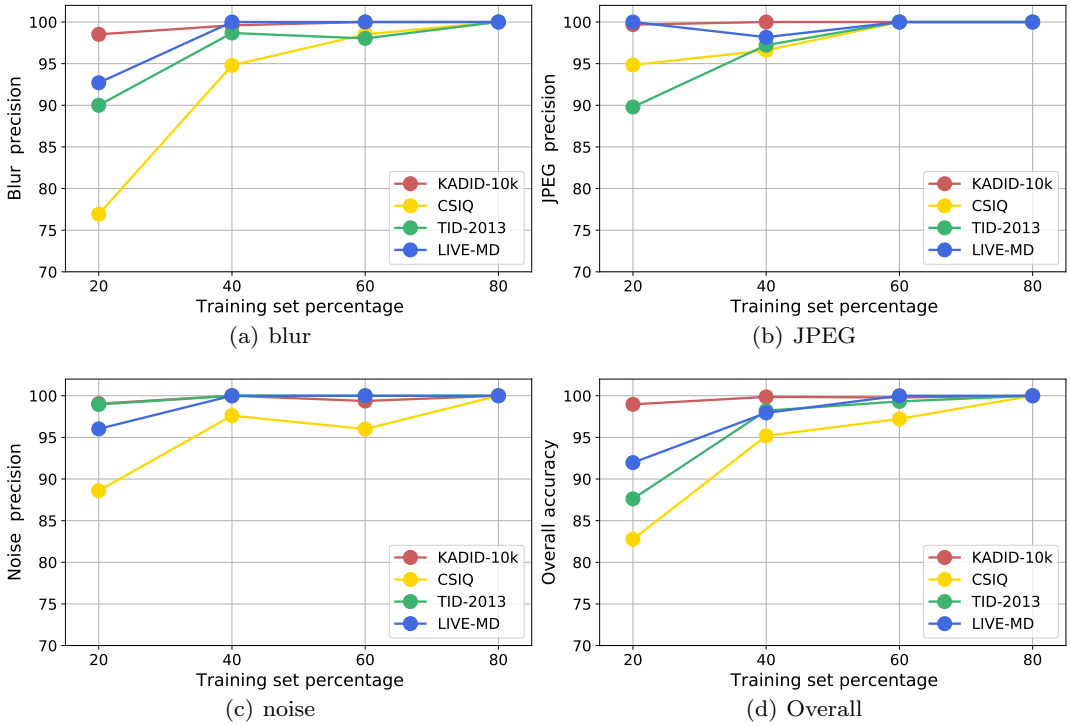
#### *4.4.5 Evaluation on LVQ dataset*

For distortion classification in laparoscopic videos, our model is evaluated on the LVQ dataset and compared to multiple state-of-the-art methods, including DIIVINE, BRISQUE and Zohaib's method [15]. It is important to specify that the LVQ dataset contains both single and multiple distortions, which makes it more challenging.

Table 10 shows the overall accuracy and the precision per distortion of our model compared to state-of-the-art methods. We can notice that DIIVINE and BRISQUE methods perform relatively poorly on smoke and uneven illumination distortions, because they were designed to process natural scene images and not laparoscopic videos content where such distortions are common. Zohaib's method provides good results but with different performance for the different distortions. However, our proposed method offers the best accuracy and outperforms all the considered methods using a single end-to-end model. This illustrates the efficiency and adaptability of our proposed model to different types of distortion encountered in different applications.

Table 9: Effect of different proportions of training data on the classification performance.

| Training(%) | CSIQ | | | | | TID2013 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $N$ samples | Precision | | | Accuracy | $N$ samples | Precision | | | Accuracy |
| | | blur | JPEG | noise | | | blur | JPEG | noise | |
| 20% | 90 | 76.92 | 94.84 | 88.59 | 82.77 | 75 | 90 | 89.79 | 98.95 | 87.62 |
| 40% | 180 | 94.79 | 96.59 | 97.61 | 95.18 | 150 | 98.68 | 97.22 | 100 | 98.21 |
| 60% | 270 | 98.52 | 100 | 96 | 97.22 | 225 | 98.03 | 100 | 100 | 99.32 |
| 80% | 360 | 96.66 | 100 | 100 | 98.88 | 300 | 100 | 100 | 100 | 100 |
| Training(%) | KADID-10K | | | | | LIVEMD | | | | |
| | $N$ samples | Precision | | | Accuracy | $N$ samples | Precision | | | Accuracy |
| | | blur | JPEG | noise | | | blur | JPEG | noise | |
| 20% | 243 | 98.52 | 99.68 | 99.05 | 98.97 | 81 | 92.70 | 100 | 96.02 | 91.97 |
| 40% | 486 | 99.60 | 100 | 100 | 99.86 | 162 | 98.46 | 100 | 100 | 97.94 |
| 60% | 729 | 100 | 100 | 99.37 | 99.79 | 243 | 100 | 100 | 100 | 100 |
| 80% | 972 | 100 | 100 | 100 | 100 | 324 | 100 | 100 | 100 | 100 |




(a) blur (b) JPEG

(c) noise (d) Overall

**Fig. 8** Precision per distortion and overall classification accuracy according to the training data size for the four considered natural scene image datasets.

Table 10: Precision per distortion and overall classification accuracy on the LVQ dataset. The top result is highlighted in boldface.

| Method | Distortion | | | | | |
|---|---|---|---|---|---|---|
| | awgn | defocus blur | motion blur | smoke | uneven illumination | Overall |
| DIIVINE [27] | 98.21 | 96.10 | 98.30 | 80.15 | 75.55 | 65.38 |
| BRISQUE [54] | **100** | 89.67 | 90 | 68.38 | 62.58 | 50.53 |
| Zohaib [15] | **100** | 91.5 | 89 | 87 | 88.5 | - |
| Our model | **100** | **100** | **100** | **100** | **99.37** | **99.37** |

Figure 9 shows the confusion matrix of our model for each of the distortions on the LVQ dataset. It provides additional information, in particular, allows to visualize which distortions are confused with others. We can see that the videos containing defocus blur are sometimes classified as containing both defocus blur and uneven illumination, this is due to the fact that the dataset does not contain enough samples of videos containing

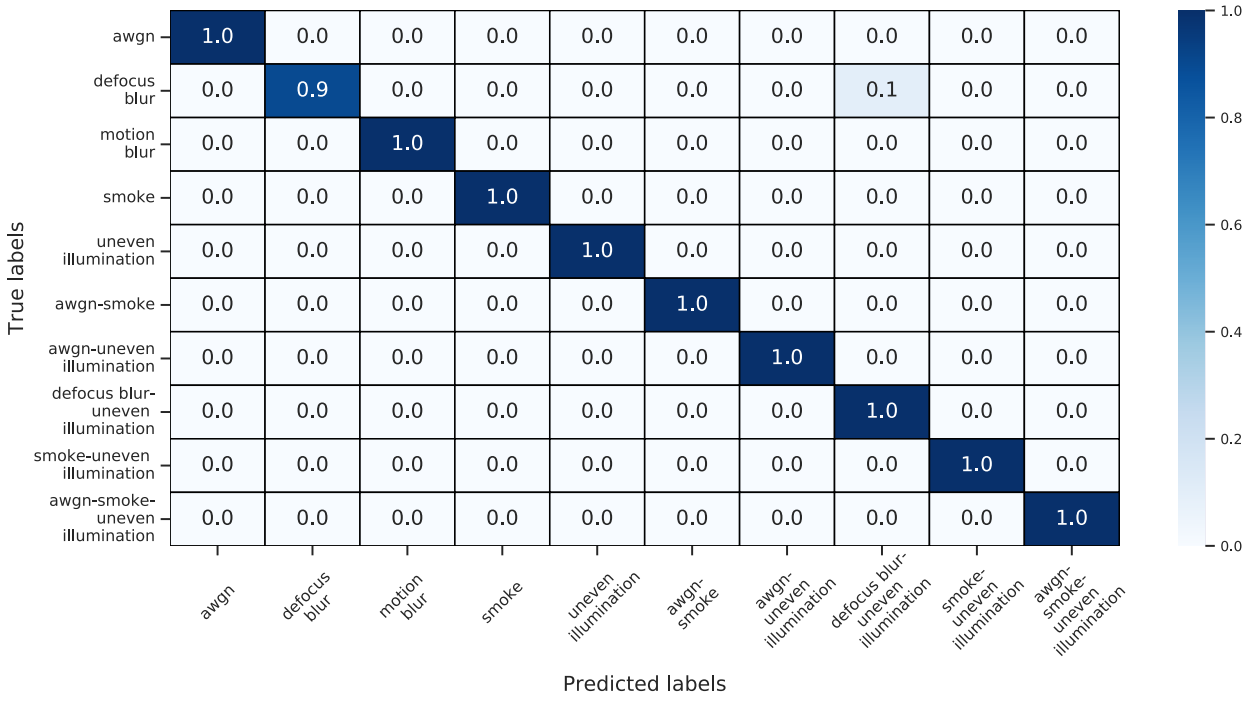



**Fig. 9** Confusion matrix on LVQ dataset.

both defocus blur and uneven illumination distortions, so that the model learns to efficiently distinguish when a video contains solely defocus and when it contains defocus with uneven illumination.

Table 11 shows the performance of our model in terms of precision and F1-score for single and multiple distortions on the LVQ dataset. From this table, we can notice that our model identifies and classifies perfectly videos containing a single distortion type, while 98.61% precision and 99.30% F1-score are reported for videos with multiple distortions due to the the complex interactions and masking effect between distortions.

Another important aspect to consider when classifying distortions in laparoscopic videos is the run time. For such application, real-time performance are required. Table 12 shows the average classification time per frame for DIIVINE, BRISQUE and our model, obtained on intel core i7 system with 32GB RAM and NVIDIA GeForce GTX 1080 running on Windows OS. We can observe that even though our model is based on deep learning techniques, which are known to be computationally heavy, it supports 0.05 FPS, thus making our approach a faster method than the other methods while still providing better accuracy. In addition, the obtained average classification time per frame allows real-time distortion classification.

As mentioned before, the task of distortion classification of laparoscopic videos has been addressed as part of a grand challenge. The evaluation of our model

Table 11: Precision and F1-score of our model for single and multiple distortions on the LVQ dataset.

| | Precision | F1-score |
|---|---|---|
| Single distortion | 100 | 99.43 |
| Multiple distortions | 98.61 | 99.30 |

Table 12: Average classification time per frame evaluation.

| Method | Average time per frame (seconds) |
|---|---|
| DIIVINE [27] | 3.30 |
| BRISQUE [54] | 0.08 |
| Our model | **0.05** |

by challenge organizers on a private dataset containing different laparoscopic videos than those provided in the training dataset, yielded an F1-score for single distortion equals to 90.7 and F1-score for a mixture of single and multiple distortions equals to 93.3. The ranking was done based on a weighted combination of classification accuracy and F1-score as defined in Eq. (12), on the basis of which our solution was ranked 3rd.

$$\begin{aligned} Finalscore = {} & 0.35\, rank\_f1\_single\_multi \\ & + 0.35\, rank\_accuracy \\ & + 0.15\, rank\_f1\_single \\ & + 0.15\, rank\_time. \end{aligned} \tag{12}$$

## 5 Conclusion

In this paper, we have proposed a deep MTL model for distortion identification. The proposed MTL model consists of a shared features extractor and a set of parallel classifiers. Each classifier is responsible for identifying a single type of distortion. This MTL architecture allows each classifier to focus on the features related to the distortion for which it is responsible, instead of covering all the distortions at the same time, making it more specialized.

The experimental results showed that our proposed method provides better performance than state-of-the-art approaches for single and multiple distortions. In addition, our MTL model offers the best trade-off between prediction accuracy and computation time. Finally, the proposed solution is scalable with respect to the number of distortion types. As future work, we plan to extend our MTL model to perform both distortion identification and severity level estimation.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References


1. Chen, F., Ma, J.: An empirical identification method of gaussian blur parameter for image deblurring. IEEE Transactions on signal processing **57**(7), 2467–2478 (2009)
2. Levin, A., Nadler, B.: Natural image denoising: Optimality and inherent bounds. In: CVPR 2011, pp. 2833–2840. IEEE (2011)
3. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on image processing **16**(8), 2080–2095 (2007)
4. List, P., Joch, A., Lainema, J., Bjontegaard, G., Karczewicz, M.: Adaptive deblocking filter. IEEE transactions on circuits and systems for video technology **13**(7), 614–619 (2003)
5. Foi, A., Katkovnik, V., Egiazarian, K.: Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. IEEE transactions on image processing **16**(5), 1395–1411 (2007)
6. Dong, C., Deng, Y., Change Loy, C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 576–584 (2015)
7. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8183–8192 (2018)
8. Mao, X.J., Shen, C., Yang, Y.B.: Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv preprint arXiv:1606.08921 (2016)
9. Kundur, D., Hatzinakos, D.: Blind image deconvolution. IEEE signal processing magazine **13**(3), 43–64 (1996)
10. Immerkaer, J.: Fast noise variance estimation. Computer vision and image understanding **64**(2), 300–302 (1996)
11. Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T.: A no-reference perceptual blur metric. In: Proceedings. International Conference on Image Processing, vol. 3, pp. III–III. IEEE (2002)
12. Wang, Z., Sheikh, H.R., Bovik, A.C.: No-reference perceptual quality assessment of jpeg compressed images. In: Proceedings. International Conference on Image Processing, vol. 1, pp. I–I. IEEE (2002)
13. Ponomarenko, N., Ieremeiev, O., Lukin, V., Jin, L., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: A new color image database tid2013: Innovations and results. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 402–413. Springer (2013)
14. Verdaasdonk, E.G., Stassen, L.P., van der Elst, M., Karsten, T.M., Dankelman, J.: Problems with technical equipment during laparoscopic surgery. Surgical endoscopy **21**(2), 275–279 (2007)
15. Khan, Z.A., Beghdadi, A., Cheikh, F.A., Kaaniche, M., Pelanis, E., Palomar, R., Fretland, Å.A., Edwin, B., Elle, O.J.: Towards a video quality assessment based framework for enhancement of laparoscopic videos. In: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, vol. 11316, p. 113160P. International Society for Optics and Photonics (2020)
16. Yu Zhang and Qiang Yang: A Survey on Multi-Task Learning . arXiv preprint arXiv:1707.08114v2 (2018)
17. Praneeth, D., Venkatanath, N., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind distortion classification using content and perception based features. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 71–75. IEEE (2014)
18. Chetouani, A., Beghdadi, A., Deriche, M.: Image distortion analysis and classification scheme using a neural approach. In: 2010 2nd European Workshop on Visual Information Processing (EUVIP), pp. 183–186. IEEE (2010)
19. Ortiz-Jaramillo, B., Garcia-Alvarez, J.C., Führ, H., Castellanos-Dominguez, G., Philips, W.: Quantifying image distortion based on gabor filter bank and multiple regression analysis. In: Image Quality and System Performance IX, vol. 8293, p. 82930E. International Society for Optics and Photonics (2012)
20. Praneeth, D., Venkatanath, N., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind distortion classification using content and perception based features. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 71–75. IEEE (2014)
21. Ahn, N., Kang, B., Sohn, K.A.: Image distortion detection using convolutional neural network. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), pp. 220–225. IEEE (2017)
22. Buczkowski, M., Stasiński, R.: Convolutional neural network-based image distortion classification. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 275–279. IEEE (2019)
23. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: Disentangling image distortions in deep feature space. arXiv preprint arXiv:2002.11409 (2020)
24. Falk, T.H., Guo, Y., Chan, W.Y.: Improving robustness of image quality measurement with degradation classification and machine learning. In: 2007 Conference Record

of the Forty-First Asilomar Conference on Signals, Systems and Computers, pp. 503–507. IEEE (2007)
25. Moorthy, A.K., Bovik, A.C.: Statistics of natural image distortions. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 962–965. IEEE (2010)
26. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal processing letters **17**(5), 513–516 (2010)
27. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE transactions on Image Processing **20**(12), 3350–3364 (2011)
28. Peng, P., Li, Z.: Image quality assessment based on distortion-aware decision fusion. In: International Conference on Intelligent Science and Intelligent Data Engineering, pp. 644–651. Springer (2011)
29. Chetouani, A., Beghdadi, A., Deriche, M.: A hybrid system for distortion classification and image quality evaluation. Signal Processing: Image Communication **27**(9), 948–960 (2012)
30. Fezza, S.A., Chetouani, A., Larabi, M.C.: Using distortion and asymmetry determination for blind stereoscopic image quality assessment strategy. Journal of Visual Communication and Image Representation **49**, 115–128 (2017)
31. Wang, H., Zuo, L., Fu, J.: Distortion recognition for image quality assessment with convolutional neural network. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2016)
32. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP), pp. 2791–2795. IEEE (2015)
33. Wang, H., Zuo, L., Fu, J.: Distortion recognition for image quality assessment with convolutional neural network. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
34. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-end blind image quality assessment using deep neural networks. IEEE Transactions on Image Processing **27**(3), 1202–1213 (2017)
35. Huang, C., Jiang, T., Jiang, M.: Encoding distortions for multi-task full-reference image quality assessment. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1864–1869. IEEE (2019)
36. Golestaneh, S.A., Kitani, K.: No-reference image quality assessment via feature fusion and multi-task learning. arXiv preprint arXiv:2006.03783 (2020)
37. Jayaraman, D., Mittal, A., Moorthy, A.K., Bovik, A.C.: Objective quality assessment of multiply distorted images. In: 2012 Conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR), pp. 1693–1697. IEEE (2012)
38. Zhang, Y., Chandler, D.M.: Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation. IEEE Transactions on Image Processing **27**(11), 5433–5448 (2018)
39. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee (2009)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
41. He, K., Zhang, X., Ren, S., Sun, J.: 2016 IEEE conference on computer vision and pattern recognition (cvpr). IEEE Las Vegas, USA (2016)
42. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2791–2795 (2015)
43. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
44. X. Yang, S. Kim, and E. P. Xing: Heterogeneous multitask learning with joint sparsity constraints. in NIPS (2009)
45. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. corr abs/1608.06993 (2016). arXiv preprint arXiv:1608.06993 (2016)
46. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7482–7491 (2018)
47. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International Conference on Machine Learning, pp. 794–803. PMLR (2018)
48. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging **19**(1), 011006 (2010)
49. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: 2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–3. IEEE (2019)
50. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
51. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
52. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
53. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
54. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR), pp. 723–727. IEEE (2011)
55. Vu, C.T., Phan, T.D., Chandler, D.M.: A spectral and spatial measure of local perceived sharpness in natural images. IEEE transactions on image processing **21**(3), 934–945 (2011)
56. Liu, X., Tanaka, M., Okutomi, M.: Noise level estimation using weak textured patches of a single noisy image. In: 2012 19th IEEE International Conference on Image Processing, pp. 665–668. IEEE (2012)