Improving Accuracy of Respiratory Rate Estimation by Restoring High Resolution Features with Transformers and Recursive Convolutional Models

Alicja Kwasniewska
SiMa Technologies, Inc.
226 Airport Pkwy Suite 550, San Jose, CA 95110
alicja.kwasniewska@sima.ai

Maciej Szankin
Intel Corporation
16409 W Bernardo Dr, San Diego, CA 92127
maciej.szankin@intel.com

Jacek Ruminski
Gdansk University of Technology
Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland
jacek.ruminski@pg.edu.pl

Anthony Sarah
Intel Corporation
16409 W Bernardo Dr, San Diego, CA 92127
anthony.sarah@intel.com

David Gamba
SiMa Technologies, Inc.
226 Airport Pkwy 550, San Jose, CA 95110
david.gamba@sima.ai

Abstract

Non-contact evaluation of vital signs has been becoming increasingly important, especially in light of the COVID-19 pandemic, which is causing the whole world to examine people's interactions in public places at a scale never seen before. However, evaluating one's vital signs can be a relatively complex procedure, which requires both time and physical contact between examiner and examinee. These requirements limit the number of people who can be efficiently checked, either due to the medical station throughput, patients' remote locations or the need for social distancing. This study is a first step to increasing the accuracy of computer vision-based respiratory rate estimation by transferring texture information from images acquired in different domains. Experiments conducted with two deep neural network topologies, a recursive convolutional model and transformers, proved their robustness in the analyzed scenario by reducing estimation error by 50% compared to low resolution sequences. All resources used in this research, including links to the dataset and code, have been made publicly available.

1. Introduction

Non-contact evaluation of breathing anomalies and, in general, interest in the state of human health has gained in significance due to the recent pandemic and the increased need for social distancing. Researchers from MIT, Boston Dynamics, and Brigham & Women's Hospital proposed a method to reduce the risk of contracting viruses by using their Spot robot to remotely measure patients' vital signs [1]. Artificial Intelligence (AI) research has also shifted towards studies of vital signs, resulting in novel architectures focused on real-time person monitoring, *e.g.*, as presented by Liu X. et al. [2] in their Multi-task Temporal Shift Attention Network. Nevertheless, the need for delivering non-contact solutions emerged a decade ago catalyzed by the rapidly increasing number of aging societies and requirements needed to support at-home medicine [3].

All these factors are contributing to the rapidly increasing research interest in camera-based evaluation of physiological signals. Innovations in this area will bring enormous advantages to society, not only by enhancing the comfort of humans but also significantly reducing latency of obtained responses, enabling simultaneous reception of signals from multiple people and allowing for continuous evaluation of individuals in order to provide them precise medical solutions. Video-based analysis of physiological signals and high-level image semantics can greatly benefit from the use of data obtained in different ranges of the electromagnetic spectrum. Visible light sequences can provide us with details about heart rate [4], emotional status [5], quantitative movement analysis [6], and many other things as well. Yet, processing of images acquired in other ranges of the electromagnetic spectrum, *e.g.*, infrared image sequences, can be

even more powerful due to revealing information not visible to the naked eye, such as body temperature [7] or respiratory data [8]. Yet, there are some factors which make processing of infrared data challenging that should be carefully weighed and mitigated if necessary. One of the main limitations is the relatively small spatial resolution of thermal image sequences compared to the visible light cameras. This results in a high level of blurriness and a low contrast between image regions, what has already been proven to negatively affect accuracy of facial area detection [9]. This concern is especially valid in the IoT and embedded edge markets, where the device footprint influences the size of the imaging sensors and thus its resolution. By using certain image processing techniques, specifically AI-based super resolution, this issue can be resolved without increasing the cost or size of the platform. A very interesting research question is whether accuracy of the non-contact vital signs estimation can be also improved by restoring or transferring high resolution details to lower quality input sequences. The main motivation behind this hypothesis is the estimators utilize temporal changes of pixel values within specific facial regions. If the images have low resolution, these differences may become indistinguishable due to the smoothed transitions, blurriness and low contrast between adjacent image regions.

In light of this, the main contribution of this study lies in evaluating whether the performance of the non-contact respiratory rate (RR) estimation benefits from the thermal data enhancement with two different deep neural network architectures: TTSR [10] aimed at solving the RefSR task (Reference-based Super Resolution which transfers high resolution details from the reference data to the low resolution input) and the DRESNet [11] model designed for the SISR problem (Single Image Super Resolution - image enhancement learned using a pair of high and low resolution images). To the best of our knowledge, this is the first work focusing on texture restoration with transformers in order to improve accuracy of vital signs estimation. In addition, the possibility of transferring features and textures between different image domains is verified by using models trained on visible light data for enhancement of thermal sequences. Finally, quality performance metrics produced by both architectures are compared against results obtained by other conventional image processing techniques commonly used for magnification of color changes related to the physiological signals, i.e., Eulerian Video Magnification (EVM) [12].

2. Related Work

The respiratory rate is a very important vital sign [13] that is used to evaluate patients with different health problems, including COVID-19 (*e.g.*, respiratory rate as a leading indicator of SARS-CoV-2 [14]). The respiratory activity and its parameters (including rate) is measured using accelerometers or gyroscope sensors [15], oxygen masks [16], bioacoustic sensors [17], inductive plethysmographs or thoracic impedance systems [18], and thermal imaging [19]. The use of thermal cameras is very important for the remote measurement procedure. The sequence of face images is recorded and the temperature changes are observed in areas of the increased air flow (nostrils, mouth). As a result of data processing, a signal containing information on the temperature changes is extracted that represents breathing activities of a person. In [20, 21] the authors used a thermal camera in a study of healthy volunteers and pathological subjects (suffering from sleep apnea). The source of the respiratory-related signal was the set of Region Of Interests (ROIs) extracted from images of the nostrils. The extracted signals contained the mean temperature in each ROI and were first normalized and then processed with wavelet analysis to estimate the respiratory rate. A thermistor was used to obtain the reference signals. A high correlation was observed between the thermal-based respiratory rate values and the reference values. Abbas et al. [22] proposed a method based on the temperature difference between two respiration phases: inspiration and expiration. The selected thermal frames and related ROIs were processed using the continuous wavelet transform (Debauchies wavelet). The validation performed on five babies showed high correlation between the proposed method and the reference values (mean respiratory rate difference below 1.2 breaths per minute (BPM)). The authors of [23, 24] proposed a method to detect the area of the nostrils that could be used as an ROI for further processing of thermal face images. Each ROI was divided into 8 smaller regions and the values within these regions were averaged. Extracted signals (including 5th-order Butterworth filtration) for each region were processed to obtain peak-to-peak time intervals and to calculate respiratory rates. The method was validated for measurements performed on 20 children demonstrating high correlation (R2 = 0.994) between the thermal-based method and reference methods (thermistor and chest belt). Later, many papers were focused on the investigation of different measurement conditions (*e.g.*, measurement during the movement of a person [25, 26] or during speech [27]), different quality of images (*e.g.*, [28]), automatic face/ROI detection and tracking ([29, 30]), improving the quality of images (*e.g.*, [11], etc. Technological progress led to the availability of small and thermal camera modules that could be used for the estimation of respiratory rates in many applications (including smart cars to monitor drivers). However, these modules are usually characterized by the smaller spatial resolution of images so the detection of facial regions can be difficult due to the blurring effects (small resolution, out of focus, etc.).

3. Problem Statement

Image enhancement based on increasing image resolution, known as super resolution (SR), is a well-known technique. In general, the goal of SR is to restore high resolution (HR) outputs from corresponding low resolution (LR) sequences. If only a single image is used for the image restoration, the approach is known as single image super resolution (SISR). The SR task aims at restoring the HR data $\widehat{Y}$ to be as close as possible to the original HR input ($\widehat{Y} = Y$) by solving the inverse problem:

$$X = (Y \circledast K) \downarrow_s + n \tag{1}$$

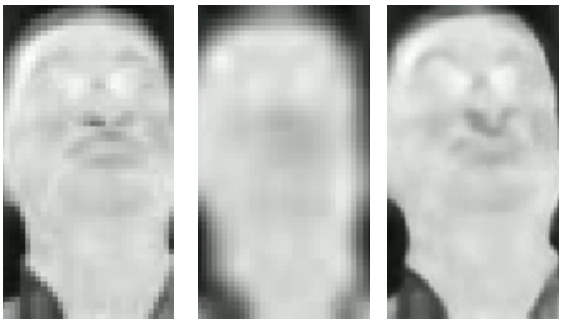
$$\widehat{Y} = SR(X) \tag{2}$$

where K is the degradation operator applied to the original HR input Y, $\downarrow_s$ is the down-scaling operation with a scale s and n is noise. Since a single LR input image X can result in various model outputs, the SR problem is very challenging. To alleviate this challenge, the solution space is constrained by the use of the structure or color information of an image occurring in input sub-parts [31, 32, 33, 34] or correspondence between LR and HR inputs[35, 36, 37, 38, 39]. The fundamental interpolation SISR uses bicubic scaling [32]. However, this and similar interpolation techniques lack the ability to discriminate between edges and centers of image regions which leads to significant image blur [39]. To mitigate this limitation, example-based solutions are often proposed. Such studies focus on preserving consistency between LR and HR data pairs by applying learning to restore detailed features. Thus, better results are produced than for random variables which are usually poorer representations of real images due to their higher variability [37].

Recent progress in artificial intelligence research has resulted in the development of a wide range of various topologies producing state-of-the-art image quality metrics, specifically Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM). The pioneering work in AI-based SISR uses Convolutional Neural Networks (SRCNN [40]) and their enhanced versions, including deeper topologies [41], skip connections combined with gradient clipping [42], recursive connections [43], residual blocks [44, 45, 46], and other approaches. Later studies focused on exploring additional modifications to the overall residual CNN-based SR idea, such as the use of multiscale input data [47]. Simultaneously, research on deconvolution-based SR networks has also been conducted. In 2016, Dong C. et al. introduced Fast SRCNN [48] - a model aimed at accelerating the breakthrough CNN SR architecture [40]. Due to the use of the transposed convolution layer at the end of the model, the need for the image to be interpolated before being fed into the network was eliminated. Another successful image upscaling-based architecture was proposed by Shi et al. [49]. The scheme of the introduced ESPCN model was motivated by the assumption that the standard deconvolution layer lead to redundant information due to repeated pixel values interpolated with the nearest neighbor approach. Thus, the alternative solution proposed in ESPCN was based on a novel subpixel convolutional layer, which stores additional pixel values in the expanded depth of feature maps. Some architectures combined both deconvolution operation and residuals introduced by ResNet. Many of these effective skip connection-based models are inspired by DenseNet [50], as it turns out that mapping between each layer and all preceding representations is very efficient for new feature exploration and as a result image enhancement [51]. Subsequently, recursive usage of residual units instead of standard convolutions led to further performance gains as shown in MemNet - the Persistent Memory Network for image restoration [52]. Applied local residual connections allow for preserving short-term memory information, which is beneficial for feature restoration in the SR task. A separate group of SR networks make use of generative models [53], *e.g.*, a topology proposed by Ledig C. et al. [45], which allows for synthesizing detailed components of the HR image and outperforming previous solutions even when using bigger scaling factors (4x). Another GAN-based SR model is based on the idea of the texture synthesis instead of image manipulation at pixel levels [54].

Although various deep neural networks have been proposed and proven to restore accurate HR representations, two architectures are of a particular interest to us. The first one, DRESNet [9], is based on the idea of increasing the receptive field while keeping the number of network parameters constant by using recursions and residuals with shared weights. The increased size of the receptive field has turned out to be crucial for the thermal data enhancement task due to the characteristics of thermal imagery. The heat flow between facial regions leads to a much lower contrast between image parts that is not correctly captured by smaller kernel sizes. DRESNet has also already demonstrated superior performance in remote medical diagnostic studies, allowing for estimation of vital signs from sequences as small as 15x20 pixels [9]. However, SISR solutions, such as DRESNet, can cause shape and structure deformations [55] as shown in the nostril area depicted in Fig. 1. We believe that this problem may lead to distortion in the physiological signals extracted as pixel values change over time, negatively affecting estimation accuracy. Taking this into account, we propose to apply the Reference-based Super Resolution (RefSR) technique for improving the quality of thermal sequences. Compared to the SISR problem solved by DRESNet, RefSR uses an additional HR reference image with textures that are helpful in super-resolving image components. Specifically, we evaluate different types of reference data in the thermal image enhancement pipeline based

on the Texture Transformer Super Resolution (TTSR) network [10], one of the first transformer architectures applied to the image generation task, achieving significant gain in image quality metrics over previous models. Based on extensive benchmarking analysis, we define what texture details lead to the best RR estimation accuracy and compare the achieved performance with other types of SR solutions, such as CNN-based SISR.



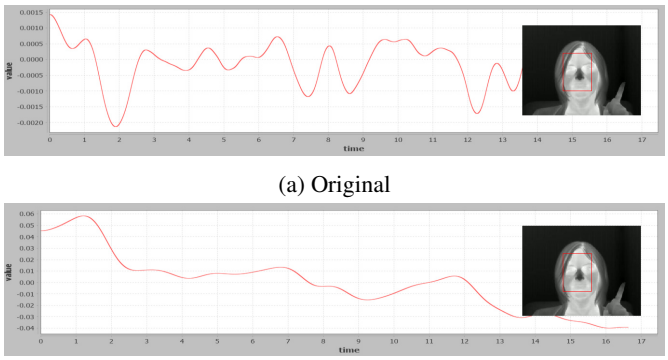
(a) Original (b) Bicubic (c) DRESNet

Figure 1: Shape and structure deformations produced by the SISR model, scale 4

4. Proposed Methods for Texture Restoration

Camera-based measurements of breathing signals bring a lot of advantages compared to traditional estimation techniques (*e.g.*, by using respiratory belts) such as the possibility to maintain social distancing, analyze RR of multiple people at once, or capture more reliable data by monitoring people without imposing any special behavior. However, low spatial resolution of the sensors leads to significant blurriness and as a result the lack of clear boundaries between facial regions (Fig. 4). At the same time, the precise localization of the ROI used for signal extraction is crucial for the accurate estimation of physiological signals [56]. In addition, reduced spatial resolution causes significant smoothing of data that might result in the high similarity of sequence frames and an inability to capture pixel value changes associated with inhaling and exhaling events. Fig. 2 shows a comparison of signals constructed by taking the skew of the pixel values within the facial area over time for the same ROI used in low resolution and high resolution sequences.

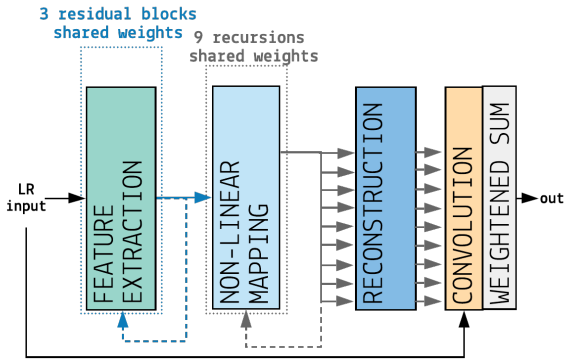
To mitigate this issue, we restore information about the image texture and high frequency features (such as edges and contours) in low resolution thermal sequences using deep neural networks: DRESNet and TTSR. Such a comparison allows for determining which topologies have a positive impact on the accuracy of the non-contact estimation of physiological signals in order to enable a new remote diagnostic solution without increasing the cost or size of acquisition devices. An overview of both architectures is shown in



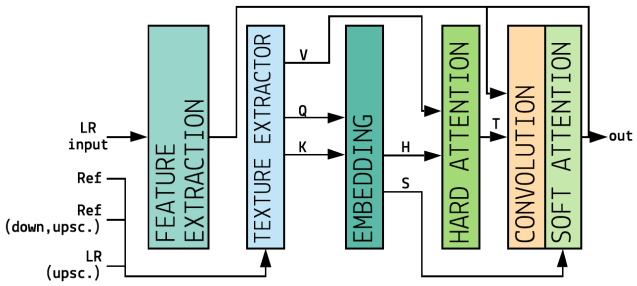
(a) Original

(b) Bicubic

Figure 2: While using the bicubic algorithm to upscale the LR image may generate a visually appealing HR image, it may not preserve the quality of other signals encoded in the original image, such as pixel value dynamics associated with physiological signs. (a) and (b) depict a signal constructed using skewness within the area over time.



(a) DRESNet Architecture



(b) TTSR Architecture

Figure 3: Overview of deep neural network topologies used for texture restoration and image resolution enhancement

Fig. 3. The DRESnet model [11] specifically addresses data acquired in the infrared range of the electromagnetic spectrum. The core idea of this model is based on the widening of the receptive field in order to take into account the more

distant relationship between image components caused by the heat flow. Because simple stacking of multiple layers is inefficient, the authors proposed to use residuals and recursions with weights shared at each step of the SISR pipeline, defined as:

$$\widehat{Y} = F_r(F_{nlm}(F_{fe}(X))) \tag{3}$$

where $\widehat{Y}$ is the restored HR data, X is the LR input, $F_{fe/nlm/r}$ are sub-networks responsible for feature extraction, non-linear mapping and reconstruction tasks. In the simplest setting, assuming a single convolution per layer, and similarly to SRCNN [40], all these steps can be solved by using convolutions (symbol $\circledast$) with trainable weights $W_{fe/nlm/r}$ and biases $B_{fe/nlm/r}$:

$$\widehat{Y} = W_r \circledast (\sigma(W_{nlm} \circledast (\sigma(W_{fe} \circledast X + B_{fe})) + B_{nlm})) + B_r \tag{4}$$

where σ is a non-linear activation function.

Due to the increased interest of transformer networks in computer vision studies and the very promising results achieved by TTSR for image generation tasks, we apply this architecture in our non-contact RR estimation pipeline.

The TTSR model consists of 4 blocks: texture extractor, embedding module, feature transfer and feature synthesis modules. Simultaneously, the LR data is fed through the backbone model to produce low resolution features. The inputs to the texture extractor are the upscaled low resolution image, the reference image which has reduced quality (by scaling with inverse scale factors) and the original reference image to produce texture components. After that, the embedding module calculates the inner product between the textures extracted from the upscaled LR image and the reference image which has reduced quality. Next, using the attention mechanism, the texture is transferred from the original reference input. Finally, the synthesis of the LR embedding and produced textures is performed using the soft-attention block to enhance relevant features and drop the noisy ones. In a simplified form, the TTSR can be denoted as:

$$\widehat{Y} = F_{fe}(X) + (F_{fe}(X) \& T) \odot S \tag{5}$$

where $\widehat{Y}$ is the restored HR data, X is the LR input, F_{fe} is the network used for extracting embeddings from the LR data, S is the output from the soft attention block, T represents transferred texture and the $\odot$ symbol denotes the element wise multiplication and the $\&$ symbol denotes the concatenation operation.

Because the choice of the reference image is arbitrary, we are particularly interested whether the texture can be transferred between images acquired in different domains, i.e., visible light details used to restore features of low resolution thermal data. Thermal image enhancement is solved by us by using three elements: single HR thermal image used across all LR thermal inputs (referred hereafter as TTSR-singleT); HR thermal images of each volunteer used for enhancing his/her LR thermal inputs (referred hereafter as TTSR-mulT); single HR visible light image used across all LR thermal inputs (referred hereafter as TTSR-singleVL).

For both architectures, the available pre-trained checkpoints are directly used for image enhancement to verify their generalization abilities and provide accurate remote diagnostic solutions across different data sets and image domains even if huge training data sets are not available for the tuning of the models.

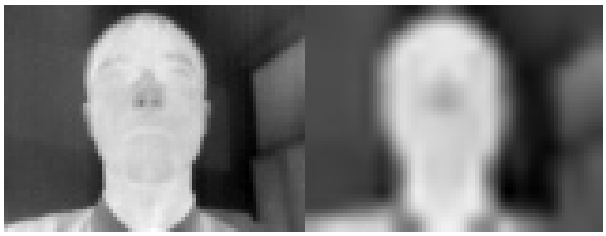
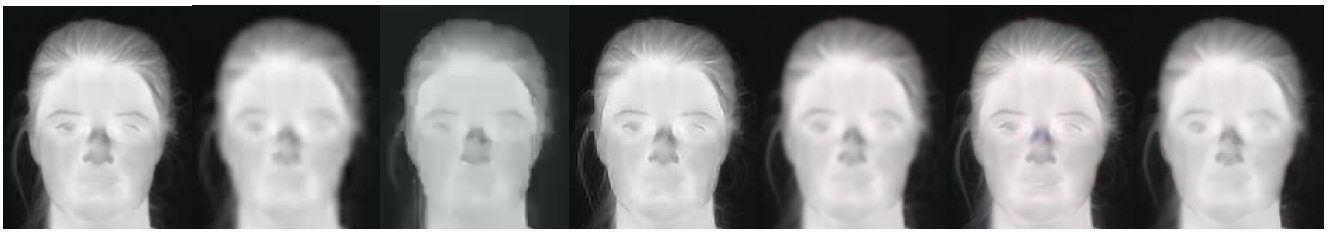


Figure 4: HR image and synthetically generated LR image

5. Datasets

For the purpose of this research, we prepared a test set consisting of thermal sequences saved in a raw format annotated with reference breathing rates. Data was collected with the help of 25 volunteers (age 34.11 $\pm$ 12). Participants were seated in front of the FLIR® SC3000 thermal camera and asked to breathe normally for a period of 2 minutes. The sensor was set up on a tripod at a distance of 1.2m from the volunteer's face, approximately 1m above the ground. This particular model of the thermal camera is capable of capturing a raw 14-bit radiometric infrared digital image in 320x240 spatial resolution in temperatures ranging from -20°C to 2000°C with 20 mK at 30°C sensitivity at up to 900 Frames Per Second (FPS). In the process of data acquisition, the frame rate was capped at 30 FPS in order to ensure that the temporal resolution is sufficient to capture breathing episodes. Due to the average respiratory rate of an adult human (10-14 BPM (Breaths Per Minute) [57]), using higher frame rates would not contribute significantly to make the measurement more precise. Given the FPS cap, an annotated set of 3600 thermal images per volunteer was generated (90000 frames in total). A simple yet effective method was used to obtain a ground truth frequency of breathing to evaluate the respiratory rate. During the data acquisition, all volunteers were instructed to signal exhaling by bending a finger and straighten it when taking a breath in. Excerpt frames from one of the gathered sequences showing this procedure are presented in Fig. 5.

Figure 5: Reference breathing signals were based on finger bending (exhaling) and straightening (inhaling)



(a) Original (b) Bicubic (c) EVM (d) DRESNet (e) TTSR-singleT (f) TTSR-mulT (g) TTSR-singleVL

Figure 6: Visualization of respective frames extracted from original HR sequence (a), sequence (c) with magnified changes using the EVM method, and HR sequences restored from LR ones using the Bicubic algorithm (b) and Super Resolution models - Convolution-based DRESNet (d) and Transformer-based TTSR (e, f, g) with various reference image setting.

The data has been linearly scaled down to 8-bit and saved as an image in a lossless grey-scale PNG format in post-processing. To accommodate for the decrease of contrast caused by down-scaling, the data set also offers raw data up-scaled to 16-bit. We plan to investigate the impact of using data of higher bit-resolution on respiratory rate estimation using Transformer-based SR in the future work.

Because sequences in this data set were captured using a single camera at a fixed resolution, LR inputs have to be generated synthetically. For this purpose, a downscaling operation ($\downarrow$) with a scale S_{down} was applied to the original HR frames, producing image sequences with a decreased quality. In particular, a scale of $S_{down} = \frac{1}{4}$ was used to verify that the RR can be estimated from extremely small thermal sequences (for the whole image measuring 80x60, a face occupies approximately 30% of a frame, which translates to 25x20 pixels). Later in the data pipeline the LR images were upscaled back to their original spatial resolution with a scale $S_{up} = \frac{1}{S_{down}}$ and fed to both DRESNet and TTSR neural networks in order to restore texture information. After super resolving all inputs, they were combined back into sequences with the same frame per second (FPS) as the source data and processed to extract the RR.

At the same time, acquired sequences were also processed with the Eulerian Video Magnification (EVM) algorithm, usually used for enhancing color changes associated with the presence of the physiological signals [12]. In the EVM algorithm, recorded video sequences are first decomposed into different frequency bands and then filtered separately to reveal changes corresponding to blood flow, vein pulsation or other information associated with physiological signals. After this, the extracted signals are magnified and added back to the original recording for visualization. In our case, the filtering frequency used in the EVM procedure was set to 0.16-0.33 Hz, given the fact that a normal RR of an adult ranges between 10-20 breaths per minute. The magnification factor was set to 20 as this value proved to have the highest estimation accuracy. Fig. 6 shows the original, magnified with EVM and super-resolved sequences.

6. Evaluation of Respiratory Rate

Non-contact estimation of respiratory rate has been already studied in-depth and various methods have been proposed to address this problem, *e.g.*, using chest motion [58], or COTS Wi-Fi devices [59]. Yet, it's important to note that this study doesn't aim at proposing better RR evaluation techniques. Instead, the main goal of our research was to evaluate the possibility of improving accuracy of vital signs extracted in a non-contact way by enhancing the texture and details of low resolution thermal sequences. Thus, we don't compare different respiration monitoring systems, but analyze how various resolution enhancement techniques affect the accuracy of the exemplary RR evaluation method, previously verified in the literature to produce satisfactory estimation results [27].

Specifically, the analysis was conducted for pixel value changes associated with temperature differences occurring during inhalation and exhalation events within facial areas. To avoid the results being influenced by incorrectly marked ROIs, areas were marked manually by an expert. However, to provide a fully automated solution, we want to combine the proposed technique with object detection models in fu-

ture studies. Ground truth annotations of the ROI used for signal extraction were done in the original HR sequences, as the location of facial features (mainly the nostril area, where the signal is visible) is much more clear and precise than in LR inputs or SR outputs that may suffer from shape distortions. For a fair comparison, the same ROI was used for extracting signals in all inputs, i.e., original HR, bicubic, EVM, DRESNet SISR, TTSR RefSR (singleT, mulT, singleVL).

We were mainly interested whether the accuracy of the RR estimation increases if proper texture and high resolution details are restored in LR sequences with the means of Deep Neural Networks. Solutions targeted by us include non-contact medical diagnostics performed in emergency rooms, principal care doctors' offices, autonomous vehicles, smart homes, etc. For such applications, latency of response is very crucial, thus the data fragments shouldn't be very long. Taking this into account, signals were extracted using up to 300 frames (10 sec.). The first few inputs in each sequence were removed to eliminate possible motion artifacts usually occurring during position adjustment at the beginning of the data acquisition process [56].

The raw respiratory signal was produced by aggregating values of pixels present in the marked ROI using a skewness operator in each frame over time. The choice of the skewness operator is motivated by the fact that it's less sensitive to specific ROI locations than the averaging method which may cause too much smoothing of important color changes. Constructed signals were filtered with a moving average and 4^{th} order Butterworth filter with the cut-off frequency set to 0.125Hz. RR values were estimated by obtaining the frequency value of the dominating peak in the signal spectrum (estimator eRR_{sp}[56]). Because the ROI position was constant across all sequences, we were able to precisely evaluate how the presence of texture information (and thus different representations of pixels) affects computer vision based estimation of physiological signals. All obtained RR values were verified against the ground truth measurement (self evaluation of volunteers) using Root Mean Square Error (RMSE).

7. Results

HR	Bicubic	DRESNet	EVM	TTSR singleT	TTSR mulT	TTSR singleVL
1.68 $\pm$1.71	3.60 $\pm$4.83	1.76 $\pm$1.90	3.42 $\pm$4.47	2.54 $\pm$3.69	2.40 $\pm$3.56	1.91 $\pm$2.16

Table 1: Mean values of the RMSE between estimated and ground truth values of RR for the original sequence and sequences enhanced with EVM and super-resolution methods

Accuracy of the RR estimation has been verified based on the root mean squared error between the ground truth measurement and the output from the eRR_{sp} estimator for each sequence separately. Table 1 presents the mean value of the error across all processed test sequences (25 volunteers). In addition, the robustness of the SR techniques was evaluated by calculating PSNR and SSIM between the original HR frames and the outputs from image enhancement algorithms, i.e., SISR DRESNet, RefSR TTSR and bicubic interpolation. In this way we were able to determine whether there is a potential gain of the RR estimation accuracy after improving data quality, restoring high frequency details and transferring image textures. PSNR and SSIM values are presented in Table 2.

	Bicubic	DRESNet	TTSR singleT	TTSR mulT	TTSR singleVL
PSNR	33.49 $\pm$1.73	43.97 $\pm$0.22	37.60 $\pm$1.52	38.94 $\pm$0.93	37.44 $\pm$1.54
SSIM	0.94 $\pm$0.01	0.96 $\pm$0.01	0.96 $\pm$0.008	0.98 $\pm$0.004	0.96 $\pm$0.008

Table 2: Mean values of the PSNR and SSIM metrics calculated across volunteers for sequences with restored spatial resolution calculated against the original sequence

8. Discussion

The presented research focused on verifying the effectiveness of various image processing techniques in improving accuracy of non-contact estimation of physiological signals. Resolution enhancement has been previously proved to bring benefits in the healthcare industry due to the possibility of restoring components important for making diagnostic decisions that are usually not visible in lower quality data [60]. Yet, only a limited number of studies were conducted for usage of AI-based SR in remote diagnostic solutions. Such analysis is very important, especially in situations like the world has been facing recently. The global pandemic has significantly increased a demand for tools allowing for self monitoring at home, or evaluation of a person's state of health while maintaining social distancing. Inspired by promising findings of McDuff D. [61] showing that convolutional models can improve accuracy of video-based photoplethysmography, as well as other vital signs, *e.g.*, respiratory rate [11], we decided to go one step further and evaluate the robustness of other recent architectures proven to be very successful in computer vision studies.

Our study is a novel contribution to this area of research, proving that transformer-based image enhancement can be used in non-contact vital sign estimation systems leading to improved accuracy without introducing more expensive and larger imaging sensors. Specifically, to the best of our knowledge, this work is a first attempt to provide thermal

sequence enhancements with attention-based RefSR in order to improve accuracy of non-contact measurements by transferring textures between imaging domains.

Based on this analysis, it has been proven that the proposed TTSR-based approach allows for decreasing the estimation error by 50% compared to LR data and achieving almost the same accuracy as in the case of the original HR inputs. Although both PSNR and RMSE of the SISR-based solution are slightly better than for TTSR, it's important to note that DRESNet was trained using data from the same imaging domain, while the TTSR was optimized using visible light images. We believe that even better results can be achieved after fine-tuning the RefSR model. In the future studies, we are planning to build bigger data sets of thermal images to be able to retrain tested architectures. In addition, the SSIM metric, which corresponds to the perceived quality of the image and thus should be more intuitive and easier to interpret turned out to be better for the TTSR pipeline, what might indicate that this model allows for producing outputs which are more pleasant for the human eye.

On the other hand, one should be aware of TTSR limitations, i.e., a need for providing reference high resolution data, which frequently might be not available. On the other hand, our study proposed to transfer textures from images representing other objects, acquired in a different domain (visible light). As presented in Table 1, such an approach led to the second best RMSE result, what might allow for eliminating the need for acquiring HR data and using other images as a reference instead. This is a very interesting outcome of the study which proves the importance of the presence of texture components in physiological signs estimation using color information. Because visible light frames usually contain more high frequency features, such as edges and contours, texture transferred from them to thermal data allowed for better restoration of color changes associated with vital signs. The estimation error in this case was approximately 25% lower than for sequences enhanced using thermal data as the reference, even though the PSNR was lower as well. This also shows a weakness of the PSNR metric in its applicability to determining accuracy of vital signs estimation. When thermal images are used as a reference texture, the restored images might be closer to the original data (higher PSNR) than when the visible image is used. However, texture information transferred from thermal images restores original representation of sequences, which is very blurry. As a result, accuracy of RR estimation might be lower due to smoother transitions between adjacent frames, lower dynamics of pixel values changes and thus higher RMSE, as it was shown in our study. It has been also proven that much better estimation accuracy can be achieved by using any of the SR methods instead of color magnification algorithms.

Although the results are very promising, they are preliminary and should be further verified in the future work. First of all, it's very important to perform similar analysis in less controlled environments, as various factors can influence the reliability of the estimation, i.e. camera angle, body position, environment conditions, etc. Secondly, the presented study addresses only a single person setting, at a close proximity to a sensor, due to target applications, such as vital signs deployed at the border control, computer stations, etc. However, real-life scenarios would require less strict restrictions on the user, what should be further analysed. Moreover, in order to develop a solution which could become an industry standard, the proposed methods have to be first verified against professional vital signs estimation devices, *e.g.*, respiratory belts. This study aimed at verifying the performance gain, which could be obtained by utilization of additional resolution enhancement techniques, but it's important to focus on verification of the method against other ground truth measurements as well, and it will be a subject of our future experiments.

The experiments conducted show the importance of SR in the application of video-based vital signs estimation. We believe that image enhancement can enable new applications, where higher resolution devices are simply not available, *e.g.*, due to the footprint of the sensor, *e.g.*, in embedded edge solutions in autonomous vehicles, telemedicine, robotics and other markets. In future work we would like to evaluate other RR estimators as well, because even for signals other than vital sings, such as noise, a maximum value in the frequency domain can be retrieved leading to false outcomes. We will also explore other image enhancement techniques using different scaling factors. Moreover, the proposed methods can be combined with other deep topologies that will also benefit from image super resolution, *e.g.*, face detection and recognition [62, 63, 64]. Such multi-model pipelines will allow for the development of fully automated systems for patient monitoring.

9. Conclusion

We have performed a benchmark evaluation aimed at determining what image domain contains textures that enable improved estimation accuracy and whether the estimation error can be reduced by super resolving source sequences. Transfer of detailed components from the visible light image domain to low resolution thermal sequences using an attention-based transformer has been shown to reduce the RR estimation error by half. This is a very important finding for many non-contact monitoring solutions. Evaluated algorithms can be deployed as a part of specialized ML-powered embedded systems to promote a better patient experience and higher accuracy of measurements without increasing the cost or size of the device. In this way, adoption of automated medicine can be accelerated ensuring that AI can really become a transformational force in healthcare.

References

[1] Hen-Wei Huang, Claas Ehmke, Gene Merewether, Fara Dadabhoy, Annie Feng, Akhil John Thomas, Canchen Li, Marco da Silva, Marc H Raibert, Edward W Boyer, et al. Agile mobile robotic platform for contactless vital signs monitoring. 2020.

[2] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[3] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011.

[4] Jure Kranjec, S Beguš, G Geršak, and J Drnovšek. Non-contact heart rate and heart rate variability measurements: A review. *Biomedical signal processing and control*, 13:102–112, 2014.

[5] Alex D Torres, Hao Yan, Armin Haj Aboutalebi, Arun Das, Lide Duan, and Paul Rad. Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. In *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*, pages 61–89. Elsevier, 2018.

[6] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):1–10, 2020.

[7] Jia-Wei Lin, Ming-Hung Lu, and Yuan-Hsiang Lin. A thermal camera based continuous body temperature measurement system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[8] Carina Barbosa Pereira, Xinchi Yu, Tom Goos, Irwin Reiss, Thorsten Orlikowsky, Konrad Heimann, Boudewijn Venema, Vladimir Blazek, Steffen Leonhardt, and Daniel Teichmann. Noncontact monitoring of respiratory rate in newborn infants using thermal imaging. *IEEE transactions on Biomedical Engineering*, 66(4):1105–1114, 2018.

[9] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Mariusz Kaczmarek. Super-resolved thermal imagery for high-accuracy facial areas detection and analysis. *Engineering Applications of Artificial Intelligence*, 87:103263, 2020.

[10] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[11] Alicja Kwasniewska, Jacek Ruminski, and Maciej Szankin. Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. *Applied Sciences*, 9(20):4405, 2019.

[12] Weixuan Chen and Daniel McDuff. Deepmag: Source specific motion magnification using gradient ascent. *arXiv preprint arXiv:1808.03338*, 2018.

[13] Michelle A Cretikos, Rinaldo Bellomo, Ken Hillman, Jack Chen, Simon Finfer, and Arthas Flabouris. Respiratory rate: the neglected vital sign. *Medical Journal of Australia*, 188(11):657–659, 2008.

[14] Dean J Miller, John V Capodilupo, Michele Lastella, Charli Sargent, Gregory D Roach, Victoria H Lee, and Emily R Capodilupo. Analyzing changes in respiratory rate to predict the risk of covid-19 infection. *PloS one*, 15(12):e0243693, 2020.

[15] Javier Hernandez, Daniel McDuff, and Rosalind W Picard. Biowatch: estimation of heart and breathing rates from wrist motions. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 169–176. IEEE, 2015.

[16] Anaxsys technology ltd, respir8,. http://www.respir8.com/about_respiR8.html. Accessed: 2021-03-08.

[17] MR Macknet, PL Kimball-Jones, RL Applegate, RD Martin, and MW Allard. Accuracy and tolerance of a novel bioacoustic respiratory sensor in pediatric patients. *Anesthesiology*, 107:A84, 2007.

[18] Robert T Brouillette, Anna S Morrow, Debra E Weese-Mayer, and Carl E Hunt. Comparison of respiratory inductive plethysmography and thoracic impedance for apnea monitoring. *The Journal of pediatrics*, 111(3):377–383, 1987.

[19] Jin Fei, Zhen Zhu, and Ioannis Pavlidis. Imaging breathing rate in the co 2 absorption band. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 700–705. IEEE, 2006.

[20] Jayasimha N Murthy, Johan Van Jaarsveld, Jin Fei, Ioannis Pavlidis, Rajesh I Harrykissoon, Joseph F Lucke, Saadia Faiz, and Richard J Castriotta. Thermal infrared imaging: a novel method to monitor airflow during polysomnography. *Sleep*, 32(11):1521–1527, 2009.

[21] Jin Fei and Ioannis Pavlidis. Thermistor at a distance: unobtrusive measurement of breathing. *IEEE Transactions on Biomedical Engineering*, 57(4):988–998, 2009.

[22] Abbas K Abbas, Konrad Heimann, Katrin Jergus, Thorsten Orlikowsky, and Steffen Leonhardt. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *Biomedical engineering online*, 10(1):93, 2011.

[23] Farah Al-Kalidi, Heather Elphick, Reza Saatchi, and Derek Burke. Respiratory rate measurement in children using a thermal camera. *International Journal of Scientific and Engineering Research*, 6(4):1748–1756, 2015.

[24] Farah Q Al-Khalidi, Reza Saatchi, Derek Burke, and Heather Elphick. Tracking human face features in thermal images for respiration monitoring. In *ACS/IEEE International Conference on Computer Systems and Applications-AICCSA 2010*, pages 1–6. IEEE, 2010.

[25] Yan Zhou, Panagiotis Tsiamyrtzis, Peggy Lindner, Ilya Timofeyev, and Ioannis Pavlidis. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5):1280–1289, 2012.

[26] Marcin Kopaczka, Jan Nestler, and Dorit Merhof. Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 518–529. Springer, 2017.

[27] Jacek Rumiński. Evaluation of the respiration rate and pattern using a portable thermal camera. In *Proc. Of the 13th Quantitative Infrared Thermography Conference*, 2016.

[28] Jacek Rumiński. Analysis of the parameters of respiration patterns extracted from thermal image sequences. *Biocybernetics and biomedical engineering*, 36(4):731–741, 2016.

[29] Youngjun Cho, Simon J Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical optics express*, 8(10):4480–4503, 2017.

[30] Prasara Jakkaew and Takao Onoye. Non-contact respiration monitoring and body movements detection for sleep using thermal imaging. *Sensors*, 20(21):6307, 2020.

[31] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

[32] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

[33] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing*, 15(8):2226–2238, 2006.

[34] Yaniv Romano, Matan Protter, and Michael Elad. Single image interpolation via adaptive nonlocal sparsity-based modeling. *IEEE Transactions on Image Processing*, 23(7):3085–3098, 2014.

[35] William T Freeman and Egon C Pasztor. Markov networks for super-resolution. In *Proc. 34th Annual Conf. on Information Sciences and Systems (CISS 2000)*, 2000.

[36] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[37] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[38] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[39] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[40] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[43] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[44] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[45] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[46] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in neural information processing systems*, pages 550–558, 2016.

[47] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[48] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016.

[49] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[50] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[51] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4799–4807, 2017.

[52] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[53] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[54] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[55] Wei-Yu Lee, Po-Yu Chuang, and Yu-Chiang Frank Wang. Perceptual quality preserving image super-resolution via channel attention. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1737–1741. IEEE, 2019.

[56] Jacek Ruminski and Alicja Kwasniewska. Evaluation of respiration rate using thermal imaging in mobile conditions. In *Application of Infrared to Biomedical Sciences*, pages 311–346. Springer, 2017.

[57] Chen Chen, Yi Han, Yan Chen, Hung-Quoc Lai, Feng Zhang, Beibei Wang, and KJ Ray Liu. Tr-breath: Time-reversal breathing rate estimation and detection. *IEEE Transactions on Biomedical Engineering*, 65(3):489–501, 2017.

[58] Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard De Haan. Video-based respiration monitoring with automatic region of interest detection. *Physiological measurement*, 37(1):100, 2015.

[59] Kai Niu, Fusang Zhang, Zhaoxin Chang, and Daqing Zhang. A fresnel diffraction model based human respiration detection system using cots wi-fi devices. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 416–419, 2018.

[60] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[61] Daniel McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018.

[62] Maciej Szankin, Alicja Kwasniewska, and Jacek Ruminski. Influence of thermal imagery resolution on accuracy of deep learning based face recognition. In *2019 12th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2019.

[63] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019.

[64] Angelo G Menezes. Analysis and evaluation of deep learning based super-resolution algorithms to improve performance in low-resolution face recognition. *arXiv preprint arXiv:2101.10845*, 2021.