Unknown Face Presentation Attack Detection via Localized Learning of Multiple Kernels

Shervin Rahimzadeh Arashloo

Abstract—The paper studies face spoofing, a.k.a. presentation attack detection (PAD) in the demanding scenarios of unknown attacks. While earlier studies have revealed the benefits of ensemble methods, and in particular, a multiple kernel learning (MKL) approach to the problem, one limitation of such techniques is that they treat the entire observation space similarly and ignore any variability and local structure inherent to the data. This work studies this aspect of face presentation attack detection with regards to one-class multiple kernel learning to benefit from the intrinsic local structure in bona fide samples to adaptively weight each representation in the composite kernel. More concretely, drawing on the one-class Fisher null formalism, we formulate a convex localized multiple kernel learning algorithm by regularising the collection of local kernel weights via a joint matrix-norm constraint and infer locally adaptive kernel weights for zero-shot one-class unseen attack detection. We present a theoretical study of the proposed localized MKL algorithm using Rademacher complexities to characterise its generalisation capability and demonstrate its advantages over some other options. An assessment of the proposed approach on general object image datasets illustrates its efficacy for anomaly and novelty detection while the results of the experiments on face PAD datasets verify its potential in detecting unknown/unseen face presentation attacks.

Index Terms—Unknown face presentation attack/spoofing detection, one-class multiple kernel learning, localised learning, generalisation analysis.

I. INTRODUCTION

WHEN exposed to sensor-level attacks, known as presentation/spoofing attacks, the trustworthiness of face recognition systems [1], [2] is significantly compromised, necessitating the development of robust presentation attack detection algorithms. Despite big advances, the current face PAD approaches have their own limitations. One source of difficulty in detecting presentation attacks is the diversity in attack samples which may be attributed to using a variety of different presentation attack mechanisms such as print, replay, mask, and make-up, or to inconsistent environmental imaging conditions resulting in a significant within-class variance that may erode the separability between classes. Unseen/unknown attacks that were not accessible during the system’s training phase represent another, an arguably severe challenge in face PAD, rendering it as an open-set classification problem [3]. In this respect, when the test-phase attacks are fundamentally different from those observed during the training phase, the performance of the classical two-class closed-set systems may be seriously compromised. Due to its significant practical importance, the detection of unseen presentation attacks has received considerable attention both in face [4], [5], [6], [7], [8] as well as in other biometric modalities [9], [10], [11], [12].

Although other alternatives may exist, the one-class classification methodology [13], [14] presents a promising strategy for the unknown face presentation attack detection problem. Specifically, the undesired training bias towards seen attacks may be removed by utilising only genuine samples for one-class training and solely trying to capture the essence of the bona fide observations. Various studies have looked into the functionality of this approach to detect unknown/unseen face presentation attacks [4], [15], [16] where it has been observed that a pure one-class learning approach trained solely on bona fide samples may yield better generalisation capability against unseen attacks compared to the closed-set two-class formulations. Although single-learner one-class face PAD systems have been developed [7], [16], [17], currently, their capabilities may be insufficient to meet the stringent requirements of a reliable and secure real-world system.

Among others, ensemble learning serves as a well-known strategy for improving pattern classification system performance in which several knowledge sources for a pattern are merged for the final decision. Different variations of this generic methodology can be related to the level at which the combination is realised, such as data level, feature level, soft-, or hard-decision level, etc. In the case of face PAD, this approach has been effectively implemented in a pure one-class setting using a classifier combination [18], [19], [20] or a multiple kernel fusion [4], [5] scheme with the ensemble system being observed to outperform each individual learner/representation enrolled in the ensemble. In particular, a reproducing kernel Hilbert space (RKHS) [21] fusion facilitates benefiting from learning strategies with a strong basis in statistical learning theory [22], [23] while allowing the learning of the classifier to be separated from the representations. Furthermore, by constructing each individual kernel based on deep representations, a multiple kernel system can benefit from
deep convolutional representations for improved detection [4], [5]. As an instance of the kernel-based ensemble methods, the approach in [5], seeks to infer one-class kernel weights to be fused for a Fisher null-space classifier in the presence of a matrix-norm regularisation on kernel weights. An experimental evaluation of the method led to promising results for detecting unseen attacks. In spite of the relative success of the one-class ensemble-based techniques for face PAD, one limitation of the existing methods is that they ignore any inherent variation in the data and treat the entire observation space similarity by associating global weights to each representation/learner irrespective of the local region where the pattern resides. This is an unrealistic assumption since, compared to other representations/learners, each particular representation/learner may be more effective in certain regions of the observation space while providing less discriminative capability in others. As such, a local strategy that considers potential variability in genuine observations may be advantageous. This aspect of the problem is overlooked in the existing ensemble one-class face PAD methods.

In the context of multiple kernel learning (MKL), while the vast majority of existing algorithms are global in nature, meaning the kernel weights are shared across the whole observation space, the issue was initially identified in [24] and has received considerable attention since then [25], [26], [27]. However, the existing localised multiple kernel learning methods use non-convex objective functions which raises questions with regard to their generalisation capacities. Another potential drawback of a non-convex formulation is the computational difficulty in obtaining a good local optimum, as well as the ensuing hurdles in reproducibility due to variable local optima. Furthermore, the majority of existing local MKL techniques’ theoretical performance guarantees are poorly understood. An exception to these is the study in [28] which formulates the MKL problem as a convex optimisation task. Yet, certain limitations apply to this method too. First, the MKL problem is formulated in the presence of an \( \ell_p \)-norm constraint on kernel weights. More recent studies have demonstrated that a more general matrix \( \ell_{p,q} \)-norm regularisation on kernel weights may yield better performance in different applications settings [5], [29], [30]. Second, the localised MKL method in [28] decouples the learning task into a set of disjoint MKL problems each associated with a different cluster in the data. Such an approach completely ignores any dependencies which might exist between kernel weights corresponding to different clusters. Third, the work in [28] focuses on a multi-/binary-class SVM-based objective function and does not provide a straightforward extension for the one-class Fisher null-space classification which has demonstrated promising performance for detecting unknown face spoofing attacks.

In this paper, we study the locality information for unknown face PAD in the context of ensemble techniques and try to capture and use the inherent local structure in bona fide samples for improved detection performance. More specifically, drawing on the one-class MKL method of [5], we formulate a local multiple kernel learning one-class method addressing the aforementioned limitations of the existing localised MKL approaches and illustrate the utility of the proposed technique in detecting unknown/unseen face presentation attacks. In particular, in the proposed method, we regularise kernel weights using a joint matrix-norm constraint as compared with the existing local MKL methods using disjoint vector \( \ell_p \)-norm regularisation. By imposing a joint norm constraint on the collection of kernel weights from all the existing clusters one enables an active interaction between kernel weights corresponding to different clusters which is in sharp contrast with the earlier studies where disjoint regularisation constraints were considered in the localised methods. Mathematically, we illustrate that incorporating locality into the MKL framework amounts to increasing the number of effective kernels (by a factor equal to the number of clusters) in the ensemble which provides further explanation for its superior performance as compared with its non-localised variant. We theoretically study the proposed approach and illustrate its merits in terms of the empirical Rademacher complexities compared to some other alternatives. Experimentally, we evaluate the proposed localised one-class MKL algorithm on different datasets and provide a comparison to the state-of-the-art techniques.

A. Main Contributions

A summary of the major contributions of this work is as follows:

- we propose a localised multiple kernel learning method for pure one-class classification of unknown/unseen face presentation attacks. By virtue of a localised formulation, the proposed approach is able to capture and benefit from the intrinsic variability and local structure inherent to bona fide samples for improved detection performance;
- we formulate the learning problem of the proposed localised MKL approach as a convex optimisation problem and present an effective method to solve it;
- a theoretical study of the generalisation performance of the proposed method based on Rademacher complexities is conducted and a comparison against other alternatives is provided;
- we evaluate the proposed localised MKL one-class approach for general object image anomaly and novelty detection as well as unseen face PAD and compare it against state-of-the-art approaches from the literature.

The rest of the paper is structured as detailed next. In Section II, we review the relevant literature on face PAD with particular attention to the unknown/unseen and ensemble face presentation attack detection methods. Section III briefly reviews the Fisher null projection for one-class classification and its extension to an MKL setting as presented in [5]. Section IV presents a new localised one-class Fisher null approach and then extends it to a multiple kernel learning setting followed by an optimisation approach to solve the corresponding convex optimisation task. Section V theoretically studies the proposed approach in terms of Rademacher complexities to derive a generalisation error bound for the proposed method and theoretically compares the proposed algorithm against some other alternatives. In Section VI,
we provide the experimental results of evaluating the proposed approach on multiple datasets. Finally, Section VII brings the paper to a conclusion.

II. PRIOR WORK

Challenge-response, hardware-, and software-based strategies [31], [32], [33] have all been put into practice to address the face PAD problem, with software-based alternatives gaining the most traction. Software-based approaches detect a presentation attack in a video/image by utilising various intrinsic features of signal content. This work uses a face PAD software formalism using RGB photos. Regarding cues employed in the software-based category, texture [34], [35] may be considered as the most extensively deployed cue for presentation attack detection, but there are other motion-based methods [36], [37], [38]. Some other studies [39], [40], [41] employ shape or colour information or spoofing medium contours [42] for face spoofing detection, while by using frequency information content in the Fourier domain, frequency-based approaches [38], [43] try to detect presentation attacks. Another study [44] has employed a statistical method to model noise in the facial PAD while deep convolutional neural networks serve as a well-known tool for detecting presentation attacks [45], [46], [47].

From a classification viewpoint, the classical method to detect face presentation attacks is the closed-set two-class formulation where both genuine and attack samples are utilised to train the system. Linear discriminant analysis [39], SVMs [48], neural networks [37] and CNNs [45], [46], Bayesian networks [49] and Adaboost techniques [50], constitute some of the most extensively used two-class methods. Regression-based procedures [51], in contrast, attempt to map the input features to the associated labels.

One-class classification provides a promising alternative to the classical two-class formulation for face presentation attack detection and has been found to be especially useful in the case of unseen attacks [15]. As an instance, the work in [6] employs a Gaussian mixture model to estimate the distribution of bona fide samples using quality metrics of images. Auto-Encoder and One-Class SVM are used to detect unknown attacks in [17] while [16] and [7] propose a subject-specific modelling strategy to construct separate one-class classifiers for each person. Other work [52] considers the unseen zero-shot face PAD using a tree network while [8] uses a triplet focal objective function for metric learning.

Ensemble methods, in which the final decision for an object is generated as a consensus of multiple learners, constitute a relatively successful subclass of one-class algorithms for detecting unknown face presentation attacks. As an instance of the work in this group, the study in [18] considers a weighted linear combination of normalised classifier output scores where the combination weights are inferred using both negative and positive observations. However, such a weight-adjusting technique raises issues about the generalisability of the learned weights to unknown attacks. The work in [19], considers an ensemble of one-class learners fused via a stacking method and presents a two-stage Genetic Algorithm to improve the generalisation error by using only genuine samples for training. In a subsequent study [53], the authors propose to optimise the parameters of a weighted averaging of multiple classifiers through a two-stage optimisation approach composed of Particle Swarm and Pattern Search algorithms to minimise the possibility of local minimum in weight learning. A new classifier score normalisation method for heavy-tailed distributions is presented to facilitate an effective combination of multiple learners. Evaluation of the algorithm on multiple datasets illustrated the promising performance of this method. However, since the fusion mechanism in a classifier combination scheme is based solely on the scores provided by individual classifiers, any extra information that could be potentially beneficial is omitted. Another study [4] considers a multiple kernel fusion approach, revealing that a kernel fusion strategy can lead to substantial improvements in performance in comparison to each base kernel. Unlike the classifier fusion paradigm, in a kernel combination method, the information content related to each representation is not summarised merely in terms of a single score, allowing the ensemble system to access a more informative representation of the data. The relative discriminative potential of each kernel, however, is not taken into account in [4], and all representations in the composite kernel are weighted equally. Another study [5] considers a multiple kernel learning algorithm that could capture the inherent sparsity within a problem and take into account potential discrepancies in the discriminative content of different kernels and infer kernel weights in a one-class learning setting to address this shortcoming. The study in [5] presented a number of desired attributes to build an ensemble face PAD system. First, instead of classifier scores, it operates on kernel-space representations, and thus, could access richer information content for classification. Second, each representation is weighted in the composite kernel where the weights are learned using genuine samples only. Third, the multiple kernel learning approach could minimise the contribution of a potentially less useful representation to the composite kernel, thanks to deploying a matrix-norm regularisation constraint inducing sparsity. And last but not the least, it is formulated as a convex optimisation problem, facilitating optimisation.

Despite their advantages with respect to the single-learner PAD systems, the existing one-class ensemble face PAD techniques have the limitation in that they ignore any inherent variation in the bona fide data and treat the entire observation space similarly by assigning a global weight to each representation/learner, regardless of the locale where the model operates. This is quite a restrictive assumption since each representation/learner may be more effective in some regions of the observation space while providing less discriminative capability in others. A local learning strategy may thus be helpful as it takes into account the potential variation in genuine observations, which is generally neglected in the existing ensemble one-class face PAD methods. This work addresses this aspect of the ensemble one-class face presentation attack detection in the context of multiple kernel learning. By virtue of using local structure, the proposed approach assigns locally adaptive weights to different representations in distinct regions
of the reproducing kernel Hilbert space and learns the relative importance of the incorporated representations using a convex formulation.

III. PRELIMINARIES

The learning problem for the regularised null-space one-class kernel Fisher method in the primal space is [54]

$$\min_{\eta} \frac{\theta}{n} \sum_{i=1}^{n} (1 - \psi(x_i)^T \eta)^2 + \|\eta\|_2^2$$

(1)

where $x_i$ denotes the set of positive training observations and $\psi(x_i)$ stands for the feature associated with $x_i$ in the Hilbert space. The optimal solution to the problem above yields a linear model in the Hilbert space, i.e. $f(.) = \psi(.)^T \eta$, that maps observations onto the regularised Fisher null space for classification. In particular, positive samples are encouraged to be mapped closer to point “1” while non-target test observations are expected to be projected further away. In practice, however, it is more convenient to consider the dual of the problem in Eq. 1 given as [5] and [54]:

$$\max_{\lambda} -\lambda^T K \lambda - \delta \lambda^T \lambda + 2 \lambda^T 1$$

(2)

where $K$ stands for the kernel matrix, the element of which at the $j^{th}$ column and $i^{th}$ row is given as $K_{ij} = \psi(x_i)^T \psi(x_j) = \kappa(x_i, x_j)$ and $\kappa(\cdot)$ stands for the kernel and $\delta = n/\theta$. The optimal solution to Eq. 2 may be determined by setting the gradient of the objective function w.r.t. $\lambda$ to zero to obtain $\lambda = (K + \delta I)^{-1} 1$, where $I$ stands for the identity matrix, and $1$ is a vector of 1’s of dimension $n$. Once the optimal $\lambda$ is computed, an observation $y$ is mapped onto the Fisher-nullspace as

$$f(y) = \left[ \kappa(y, x_1), \ldots, \kappa(y, x_n) \right] \lambda$$

(3)

The work in [5] extends the one-class kernel Fisher null method to an MKL setting by imposing a matrix $\ell_{p,q}$-norm constraint ($p, q \geq 1$) on kernel weights. Assuming $G$ base kernels $K_g$, $g = 1, \ldots, G$ which are to be fused linearly using the non-negative weights $\mu_g$’s, the MKL problem considered in [5] is

$$\min_{\mu} \max_{\lambda} -\lambda^T \left( \sum_{g=1}^{G} \mu_g K_g \right) \lambda - \delta \lambda^T \lambda + 2 \lambda^T 1$$

s.t. $\mu \geq 0$, $\|\mu\|_{p,q} \leq 1$

(4)

where the non-negativity constraint on kernel weights $\mu_g$’s, collectively represented as vector $\mu$, ensures that the combined kernel matrix is valid while $\|\|_{p,q}$ denotes a matrix $\ell_{p,q}$-norm ($p, q \geq 1$) defined as

$$\|\|_{p,q} = \left( \sum_s \left( \sum_t |u_{st}|^p \right)^{q/p} \right)^{1/q}$$

(5)

where $s$ and $t$ index the columns and the rows of the matrix, respectively, and $u_{st}$ stands for the element at the $i^{th}$ row and $j^{th}$ column of the matrix $Y$. A matrix mixed-norm constraint on kernel weights not only provides a controlling mechanism over the sparsity of kernel weights but also enables interactions between kernels [30]. On contrary, the more commonly used vector $\ell_p$-norm regularisation [23] lacks an explicit mechanism to capture inter-kernel interactions. As such, the matrix-norm regularisation has been observed to outperform its vector-norm counterpart in different settings [5], [29], [30], [55]. Note that the vector-norm constraint may be considered as a special case of the matrix-norm constraint as setting $p = q$ would reduce the matrix-norm into a vector-norm.

In [5], for solving Eq. 4, $u$ is assumed as a $G$-dimensional vector the $g^{th}$ element of which is defined as $u_g := \lambda^T K_g \lambda$. By applying the minimax theorem [56], [57] and switching the order of optimisation, the problem in Eq. 4 is expressed as

$$\max_{\mu} \left\{ -\delta \lambda^T \lambda + 2 \lambda^T 1 + \min_{\mu} \mu^T u \right\}$$

s.t. $\mu \geq 0$, $\|\mu\|_{p,q} \leq 1$

(6)

Next, an iterative two-step method based on an analytical solution in kernel weights is proposed in [5] to determine the optimisers of the saddle point MKL problem in Eq. 6.

IV. THE PROPOSED LOCALISED ONE-CLASS MKL APPROACH

In this section, first, we present a novel localised variant of the regularised null-space Fisher approach, and then extend the proposed localised method to a multiple kernel learning setting.

A. Localised Fisher Null Projection

In order to incorporate localised information into the one-class kernel Fisher null method, let us assume that the target/positive training set $x_i$ is probabilistically grouped into $C$ clusters where the probability of observation $x_i$ being assigned to the cluster indexed by $c$ is denoted by $p_c(x_i)$. Inspired by [28], we consider a convex combination of $C$ local models each associated with a different cluster and use cluster likelihoods to form the projection function. More explicitly, the projection function in the localised model is defined as

$$f(x_i) = \sum_{c=1}^{C} p_c(x_i) f_c(x_i)$$

(7)

where each local model $f_c(.)$ is characterised as $f_c(x_i) = \psi(x_i)^T \eta_c$. Considering $\psi_c(x_i) := \psi(x_i) \psi(x_i)$, the primal problem for the proposed localised one-class kernel Fisher null approach then reads

$$\min_{\eta_c, s} \frac{\theta}{n} \sum_{i=1}^{n} [1 - \sum_{c=1}^{C} \psi_c(x_i)^T \eta_c]^2 + \sum_{c=1}^{C} \|\eta_c\|_2^2$$

(8)

where the first term captures the empirical loss while the second term imposes a Tikhonov regularisation on the parameters of the localised model.
B. Dualisation

As it is common in kernel-based methods, we consider the dual of the problem in Eq. 8. For this purpose, the problem in Eq. 8 can be expressed as

\[
\begin{align*}
\min_{\eta_c} & \quad \frac{1}{\delta} \| \tau \|_2^2 + \sum_{c=1}^{C} \| \eta_c \|_2^2 \\
\text{s.t.} & \quad \tau = 1 - \sum_{c=1}^{C} \Psi_c(X)^T \eta_c
\end{align*}
\]

where \( \Psi_c(X) \) denotes a matrix collection of \( \psi_c(x_i) \)'s for \( x_i \) and \( \tau \) stands for an \( n \)-vector of one's and \( \delta = n/\theta \). The Lagrangian of the constrained problem above is

\[
\mathcal{L} = \frac{1}{\delta} \| \tau \|_2^2 + \sum_{c=1}^{C} \eta_c^T \eta_c + \omega^T (1 - \sum_{c=1}^{C} \Psi_c(X)^T \eta_c - \tau)
\]

where \( \omega \) denotes the vector of Lagrange multipliers. Next, one has to minimise \( \mathcal{L} \) in \( \eta_c \)'s and \( \tau \) and then maximise it w.r.t. \( \omega \). For the minimisation of \( \mathcal{L} \) w.r.t. \( \eta_c \)'s and \( \tau \), the derivatives with respect to these variables are set to zero. Omitting the intermediate steps, the Lagrangian is obtained as

\[
\mathcal{L} = -\frac{1}{4} \omega^T \left( \sum_{c=1}^{C} K_c \right) \omega - \frac{\delta}{4} \omega^T \omega + \omega^T 1
\]

where \( K_c \) denotes the kernel matrix associated with cluster \( c \) whose \((i,j)^{th}\) element is \( K_{ij}^{c} := \psi_c(x_i)^T \psi_c(x_j) = p_c(x_i) \kappa(x_i, x_j) p_c(x_j) \). Applying a variable change as \( \lambda = \omega/2 \), the dual problem for the proposed localised one-class null-space Fisher approach is expressed as

\[
\max_{\lambda} -\lambda^T \left( \sum_{c=1}^{C} K_c \right) \lambda - \delta \lambda^T \lambda + 2 \lambda^T 1
\]

Setting the gradient of the function above w.r.t. \( \lambda \) to zero yields the discriminant in the Hilbert space as

\[
\lambda = \left( \sum_{c=1}^{C} K_c \right)^{-1} 1
\]

After computing \( \lambda \), an observation \( y \) would be mapped onto the localised Fisher null-space as

\[
f(y) = \sum_{c=1}^{C} p_c(y) \left[ \kappa(y, x_1)p_c(x_1), \ldots, \kappa(y, x_n)p_c(x_n) \right] \lambda
\]

As may be observed from Eq. 12, the optimisation problem associated with the localised approach resembles that of its non-localised counterpart (see Eq. 2) except for the fact that the effective kernel matrix in the localised model corresponds to a sum of local kernel matrices. As such, the extensions and optimisation procedures corresponding to the non-localised approach may be adapted and applied to the new localised model. In this context, of particular interest to this study is a multiple kernel learning extension, discussed next.

C. Multiple-Kernel Localised Fisher Null Projection

For the extension of the proposed one-class localised Fisher null approach to a multiple kernel learning setting, let us assume there exist \( G \) base kernels, \( \kappa_g(., .), g = 1, \ldots, G \) which are to be combined linearly using non-negative weights.
functions of $\mu, \mu^T$, and consequently, give rise to a set which is convex [30], [58]. Furthermore, when $\mu$ is fixed, the loss is concave w.r.t. $\lambda$. Hence, the minimax theorem [56], [57] can be utilised to change the optimisation order as
\[
\max_{\lambda} \left\{ -\delta \lambda^T \lambda + 2\lambda^T \mathbf{1} + \min_{\mu} \| \mu \mu^T \|_{p,q} \right\}
\]
s.t. $\mu \geq 0, \| \mu \mu^T \|_{p,q} \leq 1$
(17)

The resulting optimisation problem above has thus become compatible with that of Eq. 6 of the non-localised MKL approach with the difference that in the proposed localised method the dimensionality of the kernel weight vector, i.e., $\mu$, is increased by a factor of the number of clusters, i.e., $C$. Since the structure of the learning problem corresponding to the proposed localised approach in Eq. 17 matches that of Eq. 6, following a similar optimisation approach as that advocated in [5], $\mu$ is derived as (a complete derivation of $\mu$ is provided in the supplementary material, Section I.)
\[
\mu = \bar{u} / \sqrt{\| \bar{u} \|_p \| \bar{u} \|_q}
\]
where
\[
\bar{u} = u \odot \left( \mu^{p-2} / \| \mu \|_p^p + \mu^{q-2} / \| \mu \|_q^q \right)^{-1}
\]
(18)
where $\odot$ denotes a Hadamard (element-wise) multiplication. Once $\mu$ is calculated, the next step is to maximise the objective function w.r.t. $\lambda$. This may be readily realised by following the standard approach to solve for the regularised Fisher null projection [54] with an effective kernel matrix of $K_c \sum_{c=1}^C \sum_{g=1}^G \mu g \circ K_g$. Requiring the partial derivative of the cost function w.r.t. $\lambda$ to vanish yields
\[
\lambda = \left( \delta \mathbf{I} + \sum_{c=1}^C \sum_{g=1}^G \mu g \circ K_g \right)^{-1} \mathbf{1}
\]
(19)

Note that the relation above gives $\lambda$ in terms of $\mu$ which is itself a function of $\lambda$. Denoting $\left( \delta \mathbf{I} + \sum_{c=1}^C \sum_{g=1}^G \mu g \circ K_g \right)^{-1} \mathbf{1} := (\lambda),$ the optimal $\lambda$ should then fulfil $\lambda = h(\lambda)$. In order to determine $\lambda$ satisfying $\lambda = h(\lambda)$, a fixed-point iteration method [59] is applied. Algorithm 1 summarises the approach described above where $\lambda$ is initialised using uniform kernel weights with a unit $\ell_{p,q}$-norm. Once $\lambda$ and $\mu$ are computed, an observation $y$ can be mapped onto the localised Fisher null subspace as
\[
f(y) = \sum_{c=1}^C p_c(y) \sum_{g=1}^G \nu g \circ K_g \left( \kappa_g(y, x_1) p_c(x_1), ..., \kappa_g(y, x_n) p_c(x_n) \right) \lambda
\]
(20)

The following proposition characterises the convergence properties of the proposed localised matrix-norm multiple kernel learning approach summarised in Algorithm 1.

**Proposition 1:** For an adequately large $\delta$, the proposed approach will converge to a single point independent of a particular initialisation for $\lambda$.

Consult the supplementary material, Section II for proof.

**V. Theoretical Analysis**

In this section, a theoretical analysis of the proposed localised $\ell_{p,q}$-norm one-class MKL approach based on Rademacher complexities is conducted which facilitates deriving a generalisation error bound for the proposed method. Furthermore, the derived Rademacher complexity bound enables a theoretical comparison between the proposed method and some other alternatives.

The hypothesis set for the proposed localised $\ell_{p,q}$-norm one-class MKL method with $C$ clusters and $G$ kernels may be expressed as
\[
H_{CG}^{p,q} = \left\{ f : \mathcal{X} \rightarrow \mathbb{R} \mid f(y) = \sum_{c=1}^C p_c(y) \sum_{g=1}^G \nu g \circ K_g \left( \kappa_g(y, x_1), ..., \kappa_g(y, x_n) p_c(x_n) \right) \lambda, \lambda \leq \Lambda, \| \mu \mu^T \|_{p,q} \leq 1 \right\}
\]
(21)

where the constraint $\lambda^T \left( \sum_{c=1}^C \sum_{g=1}^G \mu g \circ K_g \right) \lambda \leq \Lambda^2$ follows from the equivalency between a Tikhonov and an Ivanov regularisation for some $\Lambda$ [23]. The empirical Rademacher complexity [60] of the hypothesis class $H_{CG}^{p,q}$ for samples $x_i$’s is
\[
\mathcal{R}(H_{CG}^{p,q}) = \frac{1}{n} \mathbb{E}_{\sigma} \left[ \sup_{f \in H_{CG}^{p,q}} \sum_{i=1}^n \sigma_i f(x_i) \right]
\]
(22)

where $\sigma_i$’s are uniform independent random $\{-1, +1\}$ variables called Rademacher variables and $\sigma$ is a vector collection of $\sigma_i$’s. The Rademacher complexity measures the richness of the function class $H_{CG}^{p,q}$ by quantifying how well the class of functions under analysis is capable of correlating with random labels.

**Proposition 2:** The empirical Rademacher complexity of the proposed approach is bounded as
\[
\mathcal{R}(H_{CG}^{p,q}) \leq \frac{\Lambda^2}{n} \left( \mathbb{E}_{\sigma} \left[ \sum_{i=1}^n \sigma_i^2 \right] \right)^{1/2}
\]
(23)
A proof is provided in the supplementary material, Section III.

Using the Rademacher complexity, one may characterise the generalisation performance of the proposed algorithm as follows.

**Proposition 3 (Generalisation Error Bound):** Assuming that the kernel function is bounded by \( r^2 \) and the squared error loss is bounded by \( B_1 \) and is \( \mu \)-Lipschitz for some \( \mu \), then the following inequality holds with a confidence larger than \( 1 - \Delta \) over samples of size \( n \) for all classifiers \( f \in \mathcal{H}_{CG}^{p,q} \)

\[
\epsilon(f) \leq \frac{1}{n} \sum_i (f(x_i) - 1)^2 + \frac{n \mu}{n} (CG)^{2-1-1} G^{1/2} \left( \sum_c \sum_i p_c^2(x_i) \right)^{1/2} + 3B_1 \sqrt{\frac{\ln(2/\Delta)}{2n}} \tag{24}
\]

where \( \epsilon(f) := E\left[ (\hat{f}(y) - f(y))^2 \right] \), and \( \hat{f}(\cdot) \) denotes the true label of an observation.

**Proof:** The proof follows by plugging in the Rademacher complexity bound and the empirical loss function of the proposed approach into Theorem 11.3 in [60].\( \square \)

The generalisation error bound in Eq. 24 suggests a trade-off between reducing the empirical error over the training set and controlling the Rademacher complexity of \( \mathcal{H}_{CG}^{p,q} \). As suggested by Eq. 23, the empirical Rademacher complexity for the hypothesis set \( \mathcal{H}_{CG}^{p,q} \) may be controlled using \( \Delta \) which depends on \( \delta \) (see the supplementary material, Section IV for a proof).

As a result, in order to optimise the generalisation error bound, one needs to set \( \theta \) (or equivalently \( \delta \)) so as to maintain a trade-off between different terms in the upper bound given in Eq. 24. In practice, a validation set independent from the training set may be used to tune \( \theta \) so that a performance metric is optimised.

### A. Other Regularisation Constraints

In this study, we considered a joint \( \ell_{p,q} \)-norm constraint on the collection of kernel weights from all the existing clusters. Other closely related alternatives for norm regularisation include: (1) a joint vector \( \ell_p \)-norm; (2) a disjoint vector \( \ell_p \)-norm; or (3) a disjoint matrix \( \ell_{p,q} \)-norm. For the joint vector \( \ell_p \)-norm regularisation, the constraint is expressed as \( \| \mu \|_p \leq 1 \) while for the disjoint vector-norm regularisation, the constraint set is defined as \( \| \mu_c \|_p \leq 1 \), for \( c = 1, \ldots, C \). Regarding the disjoint \( \ell_{p,q} \)-norm regularisation, the constraints are \( \| \mu_c, \mu_c^\top \|_{p,q} \leq 1 \), for \( c = 1, \ldots, C \). As formally stated in the following proposition, the empirical Rademacher complexity bound of the proposed joint matrix-norm approach is lower than those of the joint vector-norm, disjoint vector-norm, and the disjoint matrix-norm class.

**Proposition 4:** The empirical Rademacher complexity of the proposed localised joint \( \ell_{p,q} \)-norm (1 \( \leq q \leq p \)) one-class MKL approach is

(i) lower than the localised joint vector-norm function class by a factor of \( (CG)^{-1/2} \)

(ii) lower than the localised disjoint vector-norm function class by a factor of \( C^{-1/2} G^{-1} \)

(iii) lower than the localised disjoint matrix-norm function class by a factor of \( C^{-1/2} G^{-1/2} \)

A proof is provided in the supplementary material, Section V.

### B. Runtime Complexity

During training, we incur \( O(CGn^2) \) for computing kernel matrices, and during each training iteration \( O(Cn^3) \) for computing vector \( u \) while for solving the linear equation \( (\mathbf{I} + \sum_{c=1}^C \sum_{g=1}^G \mu_{cg} K_{cg}) \lambda = \mathbf{1} \) via incremental Cholesky factorisation using the Sherman’s March algorithm [61] and forward-back substitution one needs \( O(n^2) \). The complexity at the test phase is \( O(CGn) \).

### VI. Experiments and Results

In this section, the results of evaluating the proposed approach on different datasets are presented and compared against existing methods. The remainder of this section is structured as detailed below.

- **Section VI-A,** provides the details of our implementation.
- In **Section VI-B,** we visualise the local kernel weights inferred for some sample classes to confirm the possibility of inferring locally adaptive kernel weights using the proposed method.
- In **Section VI-C,** the results of evaluating the proposed approach on one-class general object classification for anomaly and novelty detection on the Caltech256 [62] and the CIFAR10 [63] datasets are presented and compared against other methods.
- **Section VI-D** presents the results of the unseen face PAD on four widely used datasets in a zero-shot evaluation scenario and compares the proposed approach with the state-of-the-art techniques.

#### A. Details of Implementation

In the following experiments, we use an RBF kernel function and set its width to half of the mean pairwise Euclidean distance over training observations. \( \delta \) is selected from \( \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\} \times n \) while \( p \) and \( q \) are selected from \( \{32/31, 16/15, 8/7, 4/3, 2, 4, 8, 10\} \). On all datasets, the parameters above are set using the data from classes other than the target class under consideration. The number of genuine data clusters \( C \) is set to 3 in all of the following experiments. We form the clusters using the kernel k-means algorithm and derive \( p_c(x_i) \)'s using a softmax function operating on the distances from cluster centres.

#### B. Effect of Locality

The objective of the current study is to infer locally adaptive kernel weights. In order to visualise and confirm that different weights can be inferred for different data clusters, we randomly pick two classes from the Caltech256 dataset [62],
and following [5], construct seven kernels using deep representations derived from the pre-trained convolutional neural networks of Resnet50 [64], Googlenet [65], Alexnet [66], Vgg16 [67], Densenet201 [68], Mobilenetv2 [69], and Nasnetlarge [70]. We set the number of clusters to three and run the proposed algorithm to infer local kernel weights. The weights inferred for the seven kernel matrices used are visualised in Fig. 1. As can be seen from the figure, for both object classes, a specific kernel can be weighted differently in different clusters. As such, the proposed approach is effective in inferring locally adaptive kernel weights. The utility of the inferred local weights for one-class classification is analysed in the following sections.

C. One-Class General Object Classification

In the first set of experiments, the proposed approach is evaluated for one-class classification of general object images. For this purpose, two experiments for novelty and anomaly detection are conducted on the Caltech256 and CIFAR10 datasets, discussed next.

1) The Caltech 256 Dataset [62]: is commonly used to test novelty detection algorithms. The dataset incorporates images of objects from 256 different classes giving rise to 30607 images in total. For novelty detection, different protocols have been introduced to convert the multi-class dataset into one suitable for one-class novelty detection. This first experimental set-up is introduced in [71] where a single class is presumed as the target/normal class and the learned model is then evaluated by considering objects from all the remaining 255 classes as novel objects. The experiment is repeated for the first 40 classes of the database. The second experimental set-up is the one followed in [72]. In this case, each of the five randomly chosen object classes is considered as the positive class and the samples from the “clutter” class are presumed as outliers. Samples from the clutter category are randomly selected in a way that each experiment incorporates 50% outlier objects.

2) The CIFAR10 Dataset [63]: is one of the frequently used datasets for anomaly detection that includes 60000, 32 × 32 colour images from 10 classes. The dataset includes 50000 images for training and 10000 samples for testing purposes. Each category incorporates 6000 images from which 5000 images are employed for training and 1000 samples for testing. In this experiment, ten anomaly detection problems are generated where in each problem a single category is assumed to be the normal class while the others are anomalies.

On the Caltech256 and CIFAR10 databases, we use the generally known AUC performance metric to compare the performance of the proposed method to those of existing methods. The AUC provides a performance statistic independent of a particular threshold by measuring the area under the Receiver Operating Characteristic curve (ROC), which contrasts the true positive rate against the false positive rate at varying thresholds. In order to specifically gauge the performance benefits brought by the proposed localised MKL approach compared to the non-localised matrix-regularised MKL [5], we use the features obtained from the seven deep pre-trained CNN’s of Resnet50 [64], Googlenet [65], Alexnet [66], Vgg16 [67], Densenet201 [68], Mobilenetv2 [69], and Nasnetlarge [70] as suggested in [5] to construct 21 localised kernel matrices (7 representations x 3 clusters) and compare the performance of the proposed method to its non-localised variant as well as other methods from the literature. The results of this set of experiments are tabulated in Table I, II, and Table III for the Caltech256 and CIFAR0 datasets, respectively. From the tables, the following observations may be made. On the Caltech256 dataset, while the non-localised MKL approach achieves an average AUC of 99.6% for the first experimental set-up, the proposed localisation improves this performance and yields a perfect classification for novelty detection, ranking the best for novelty detection in the first protocol for the Caltech256 database. In the second experimental set-up on the Caltech256 dataset (Table II) the proposed approach achieves the best-reported performance. On the CIFAR10 database, while the average AUC of the non-localised method is 87.1%,
the proposed localised MKL approach obtains an average AUC of 87.9%, performing better than other alternatives.

D. One-Class Unseen Face PAD

In this set of experiments, the proposed approach is evaluated for the unknown face PAD and compared to other methods that follow a similar zero-shot unseen PA detection setting. For this purpose, we use four commonly used face PAD databases briefly introduced below.

1) The Replay-Mobile Database [87]: incorporates 1190 video sequences of genuine and attack observations from 40 subjects that are captured by two different video cameras under various lighting conditions. In this dataset, three disjoint subsets exist which are to be used for training, development, and testing purposes as well as an additional enrolment subset.

2) The OULU-NPU Database [88]: includes 4950 attack and genuine videos corresponding to 55 individuals recorded with 6 different cameras in 3 sessions subject to various lighting conditions and background settings. The dataset evaluation protocol requires PA detection subject to previously unseen video acquisition settings, unseen attack types and also unseen camera sensors. The videos are separated into three subject-disjoint groups as train, development, and test sets. For testing, 4 evaluation protocols are considered among which Protocol IV is the most difficult one and is followed in this work.

3) The MSU-MFSD Dataset [89]: incorporates 440 videos recorded from either video or photo attack trials corresponding to 55 individuals which are captured using 2 different devices. The subset of the dataset which is publicly available includes videos corresponding to 35 subjects. The dataset is partitioned into 2 sets for training and testing purposes with no overlap between the subjects from the two sets.

4) The Replay-Attack Dataset [90]: incorporates 1300 videos of genuine and attack data from fifty subjects. Attack data are generated by means of a mobile phone, an HD iPad screen, or using a printed image. The database is randomly divided into three subject-disjoint partitions for training, development, and testing purposes.

The ISO metrics [91] used to gauge presentation attack detection system performance are 1) BPCER: bona fide presentation classification error rate which measures the ratio of genuine samples which are incorrectly classified; and 2) APCER: attack presentation classification error rate that reflects the ratio of incorrectly classified PAs realised with a similar PA instrument species. Based on the maximum APCER over all attack instruments, the presentation attack detection system performance may be summarised in terms of ACER, i.e. the Average Classification Error Rate as

$$ACER = \left( \frac{BPCER + \max_{PA} APCER_{PA}}{2} \right)$$

In addition to the aforementioned metrics and to enable benchmarking with other approaches, we also report the performance of the proposed approach in terms of the Area Under the ROC Curve (AUC) and Half Total Error Rate (HTER). The kernels used in this set of experiments are those suggested in [5] which have been observed to be effective for face presentation attack detection. These kernels correspond to deep CNN features extracted using Resnet50 [64], Googlenet [65], and Vgg16 [67] pre-trained models from four different facial regions, giving rise to 36 kernels in the proposed localised approach: $3^\text{CNN’s} \times 4\text{(regions)} \times 3\text{(clusters)}$. The facial regions from which features are extracted roughly represent the nose and the surrounding areas, the two eyes along with the nose region, the mouth and nose regions in addition to the whole face image. All deep pre-trained networks are applied to the above four regions. Using similar deep representations as those suggested in [5] enables a direct comparison between the proposed localised MKL method and its non-localised variant. We follow the methodology in [16] and build subject-specific classifiers for each client enrolled on the system. The results of this set of experiments are provided in Table IV, Table V.

### Table III

| Method              | AUC %  |
|---------------------|--------|
| LSA [79]            | 64.1   |
| Skip-冈Ganomaly [80]| 73.2   |
| DAA [81]            | 75.3   |
| HRN [82]            | 71.3   |
| CAVGA-Da [83]       | 73.7   |
| OCGAN [84]          | 65.7   |
| DeepSVD [85]        | 64.8   |
| MKD [86]            | 84.5   |
| Localised MKL (this work) | **87.9** |

### Table IV

| Approach                                   | HTER (%) |
|--------------------------------------------|----------|
| The work in [16]                          | 8.5      |
| Weighted-Fusion [18]                       | 9.9      |
| GoogleNet-MD [7]                          | 13.7     |
| The work in [4]                           | 11.8     |
| Multiple Kernel-SVDD-Slim [5]              | 7.6      |
| Multiple Kernel-OC3VM-Slim [5]             | 8.1      |
| $\ell_p$, MKL [5]                         | 6.7      |
| Localised MKL (this work)                 | 5.6      |

### Table V

| Approach                                   | BPCER | APCER | ACER   |
|--------------------------------------------|-------|-------|--------|
| SAPLC [92]                                 | 6.6 $\pm$ 5.5 | 11.9 $\pm$ 6.9 | 9.3 $\pm$ 4.4 |
| OCA-PAS [93]                               | 5.9 $\pm$ 4.5 | 2.3 $\pm$ 2.3 | 4.1 $\pm$ 2.7 |
| The works in [94]                          | 1.7 $\pm$ 2.6 | 5.8 $\pm$ 4.9 | 3.7 $\pm$ 2.1 |
| The works in [95]                          | 9.7 $\pm$ 4.8 | 11.3 $\pm$ 3.9 | 9.8 $\pm$ 4.2 |
| The work in [5]                            | 0.6 $\pm$ 2.0 | 11.6 $\pm$ 13.6 | 6.2 $\pm$ 6.8 |
| Multiple Kernel-SVDD-Slim [5]              | 1.7 $\pm$ 4.1 | 10.8 $\pm$ 8.0 | 6.2 $\pm$ 4.4 |
| Multiple Kernel-OC3VM-Slim [5]             | 1.7 $\pm$ 4.1 | 10.8 $\pm$ 8.0 | 6.2 $\pm$ 4.4 |
| $\ell_p$, MKL [5]                          | 0 $\pm$ 0 | 5.0 $\pm$ 4.5 | 2.5 $\pm$ 2.2 |
| Localised MKL (this work)                 | 0 $\pm$ 0 | 3.3 $\pm$ 4.1 | 1.6 $\pm$ 2.0 |
the average APCER of the proposed approach is 3.

improves upon its non-localised variant. In particular, while
be made. That is, the proposed localised learning approach
On the Oulu-NPU protocol IV, a similar observation may
presented in [18] and also better than the method of [16].
formulations. The proposed method also yields better perfor-
other multiple kernel learning systems operating on SVM
of HTER. The proposed localised approach also outperforms
5
localised MKL method of [5] obtains an HTER of 6.

Mobile database, in an unknown face PAD setting, the non-
observations from the tables are in order. On the Replay-
Table VI, and Table VII, for the Replay-Mobile [87], Oulu-
NPU [88], MSU-MFSD [89], and

Replay-Attack [90] databases, respectively. The following observations from the tables are in order. On the Replay-
Mobile database, in an unknown face PAD setting, the non-
localised MKL method of [5] obtains an HTER of 6.7%. The proposed localised MKL approach reduces the error rate to
5.6% accounting for more than 15% improvement in terms of HTER. The proposed localised approach also outperforms other multiple kernel learning systems operating on SVM formulations. The proposed method also yields better performance in comparison to the weighted classifier fusion method presented in [18] and also better than the method of [16]. On the Oulu-NPU protocol IV, a similar observation may be made. That is, the proposed localised learning approach improves upon its non-localised variant. In particular, while the average APCER of the proposed approach is 3.3%, for the non-localised version the corresponding average APCER is 5.0 which amounts to more than 30% improvement in terms of APCER. The overall performance of the proposed method on the Oulu-NPU dataset measured as ACER is better than other methods evaluated in an unknown face spoofing detection scenario. For the MSU-MFSD and the Replay-Attack databases, the proposed approach provides a perfect separation between bona fide and attack samples. The outstanding performance of the proposed approach as compared with other alternatives may be attributed to 1) an optimal fusion of multiple deep CNN representations in the RKHS which is in contrast to the majority of existing work that tries to design and optimise a single CNN for face PAD; and 2) utilisation of the locality information in the kernel feature space to adaptively weight kernels which is in contrast to other ensemble-based face PAD methods that ignore the local structure in bona fide samples.

VII. CONCLUSION

We considered the unseen face PAD problem in the framework of one-class multiple kernel learning. While earlier MKL studies failed to explicitly take into the variability in bona fide samples in learning kernel weights, we presented a localised one-class MKL algorithm which could be utilised to learn locally adaptive kernel weights. We posed the corresponding learning problem as a convex optimisation task and presented a method for its optimisation. The results of evaluating the proposed method on general object image and face spoofing data verified its efficacy in detecting presentation attacks by solely training on genuine samples while a theoretical analysis of the proposed localised multiple kernel learning approach based on Rademacher complexities illustrated its benefits compared to some other alternatives.

REFERENCES

[1] F. Mokhayeri, E. Granger, and G. A. Bilodeau, “Domain-specific face synthesis for video face recognition from a single sample per person,” IEEE Trans. Inf. Forensics Security, vol. 14, no. 3, pp. 757–772, Mar. 2019.
[2] B. Meden et al., “Privacy–enhancing face biometrics: A comprehensive survey,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4147–4183, 2021.
[3] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, “Toward open set recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 7, pp. 1757–1772, Jul. 2013.
[4] S. R. Arashloo, “Unseen face presentation attack detection using sparse multiple kernel Fisher null-space,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 4084–4095, Oct. 2021.
[5] S. R. Arashloo, “Matrix-regularized one-class multiple kernel learning for unseen face presentation attack detection,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4635–4647, 2021.
[6] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, “On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing,” in Proc. Int. Conf. Biometrics (ICB), Feb. 2018, pp. 75–81.
[7] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, “Spoofing attack detection by anomaly detection,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 8464–8468.
[8] D. Perez-Cabo, D. Jimenez-Cabello, A. Costa-Pazo, and R. J. Lopez-Sastre, “Deep anomaly detection for generalized face anti-spoofing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 1–10.
[9] A. Rattani, W. J. Scheirer, and A. Ross, “Open set fingerprint spoof detection across novel fabrication materials,” IEEE Trans. Inf. Forensics Security, vol. 10, no. 11, pp. 2447–2460, Nov. 2015.
[10] Y. Ding and A. Ross, “An ensemble of one-class SVMs for fingerprint spoof detection across different fabrication materials,” in Proc. IEEE WIFS, Dec. 2016, pp. 1–6.
[11] P. M. Ferreira, A. F. Sequeira, D. Peroes, A. Rebelo, and J. S. Cardoso, “Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations,” in Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), Sep. 2019, pp. 1–7.
