Engineering Applications of Artificial Intelligence 127 (2024) 107255

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Domain-knowledge Inspired Pseudo Supervision (DIPS) for unsupervised image-to-image translation models to support cross-domain classification

Firas Al-Hindawi [a,*], Md Mahfuzur Rahman Siddiquee [a], Teresa Wu [a], Han Hu [b], Ying Sun [c]

[a] Arizona State University, 699 Mill Avenue, Tempe, AZ 85281, United States of America
[b] University of Arkansas, 1 University of Arkansas, Fayetteville, AR 72701, United States of America
[c] University of Cincinnati, 600 Clifton Ave, Cincinnati, OH 45221, United States of America

ARTICLE INFO

Keywords:
Unsupervised metrics
Cross-domain classification
Critical heat flux
Domain adaptation
Generative adversarial networks
Image-to-image translation
Pool boiling
Unsupervised machine learning

ABSTRACT

The ability to classify images is dependent on having access to large labeled datasets and testing on data from the same domain of which the model was trained on. Classification becomes more challenging when dealing with new data from a different domain, where gathering and especially labeling a larger image dataset for retraining a classification model requires a labor-intensive human effort. Cross-domain classification frameworks were developed to handle this data domain shift problem by utilizing unsupervised image-to-image translation models to translate an input image from the unlabeled domain to the labeled domain. The problem with these unsupervised models lies in their unsupervised nature. For lack of annotations, it is not possible to use the traditional supervised metrics to evaluate these translation models to pick the best-saved checkpoint model. This paper introduces a new method called Domain-knowledge Inspired Pseudo Supervision (DIPS) which utilizes Gaussian Mixture Models and domain knowledge to generate pseudo annotations to enable the use of traditional supervised metrics. This method was designed specifically to support cross-domain classification applications contrary to other typically used metrics such as the Fréchet Inception Distance (FID) which were designed to evaluate the model in terms of the quality of the generated image from a human-eye perspective. DIPS outperforms state-of-the-art GAN evaluation metrics when selecting the optimal saved checkpoint. Furthermore, DIPS showcases its robustness and interpretability by demonstrating a strong correlation with truly supervised metrics, highlighting its superiority over existing state-of-the-art alternatives The boiling crisis problem has been approached as a case study. The code and data to replicate the results can be found on the official GitHub-repository[1].

1. Introduction

Machine learning prediction and classification algorithms have been a rapidly growing field in recent years, particularly with the significant advancements in deep learning and computer vision technologies. These advancements enabled prediction algorithms to become highly efficient and accurate, making them applicable to a wide range of applications in various domains such as tensile strength prediction of polymers (Alhindawi and Altarazi, 2018; Altarazi et al., 2019), Critical heat flux detection (Rassoulinejad-Mousavi et al., 2021; Al-Hindawi et al., 2023), soil fertility classification (Padmapriya and Sasilatha, 2023), and skin lesion classification (Omeroglu et al., 2023). For classification models to work optimally, there are several contingencies to consider, one of which is having access to large, balanced, and accurately labeled datasets.

The limitations of these classification models become evident when dealing with new data from a different domain. In such situations, gathering a substantial dataset with labels and creating a new classifier from the beginning may require significant time and resources which may not always be practical, this problem is known in the field as the data domain shift problem. This prompted the development of a branch of machine learning called unsupervised domain adaptation (UDA), which deals with the data domain shift problem in an unsupervised manner. The specific classification problem described earlier falls under the UDA umbrella and is generally referred to in the field of machine learning as unsupervised cross-domain classification, which is the problem of training a classifier on a dataset from one domain and using it to predict a dataset from a different domain without the labeling information. UDA could be categorized into discrepancy-based, reconstruction-based, or

* Corresponding author.
E-mail addresses: falhinda@asu.edu (F. Al-Hindawi), mrahmans@asu.edu (M.M. Rahman Siddiquee), teresa.wu@asu.edu (T. Wu), hanhu@uark.edu (H. Hu), sunyg@ucmail.uc.edu (Y. Sun).

[1] https://github.com/Hindawi91/DIPS

https://doi.org/10.1016/j.engappai.2023.107255
Received 1 March 2023; Received in revised form 23 September 2023; Accepted 2 October 2023
Available online 11 October 2023

adversarial-based UDA depending on the domain adaptation approach used (Wang and Deng, 2018). Researchers used a variety of these UDA approaches to tackle the cross-domain classification problem, such as transforming the original image into analogs in multiple related domains, thereby learning features that are robust to variations across domains (Ghifary et al., 2015), and aligning the distributions of the domains using extracted features (Tzeng et al., 2017), or multiple features representations (Zhu et al., 2019) to capture the information from different aspects. Such methods can handle domains with big shifts between them, but in the case of domains with small shifts, the lost spatial information during transformation might affect the results negatively.

Recently, with the rise of Generative Adversarial Networks (GANs) and Unsupervised Image to Image (UI2I) translation models, researchers started investigating their utility in solving the cross-domain classification problem. Image-to-image translation models, in general, are used to convert an input image from one domain (e.g., horse images) to another domain (e.g., zebra images) using a generative model. Once trained, UI2I translation models can be used to transform an input image from the source domain to the target domain by inputting the image into the model and generating a synthetic output image in the target domain. This translation capability inspired a number of approaches that leverage GANs and UI2I models to address the cross-domain classification problem.

For example, Deng et al. (2018) translated the source dataset to the target domain and then trained a new model on the features of the translated images. Xiang et al. (2020) synthesized a dataset and fine-tuned the model using the synthesized dataset. Li et al. (2021) utilized GANs and attention to leverage the information available from the source domain to the target domain. Al-Hindawi et al. (2023) developed a framework using UI2I translation models that expanded the classification capability to include the target domain. Goel and Ganatra (2023) Proposed a framework that utilizes a guided transfer learning approach to select layers for fine-tuning, enhancing feature transferability, and minimizing domain discrepancies using JS-Divergence.

One of the main challenges in such unsupervised frameworks is knowing which UI2I model to select/save during training (when to stop the training). If the model was not trained long enough, it will underfit the training data and produce poor results. In contrast, if the model was overtrained, it will overfit the training data and generate poor results as well. Because of the unsupervised nature of UI2I models, the standard supervised metrics that are used to validate supervised models during training (e.g., accuracy, AUC) are not available to be used without access to the labeling information. To overcome this issue, such frameworks depend on one of the most used metrics in evaluating unsupervised image-to-image translation models, the Frechet Inception Distance (FID) (Heusel et al., 2017). FID is a popular metric for evaluating the quality of generated images in the context of Generative Adversarial Networks (GANs) and other image synthesis models. However, like any metric, FID has its drawbacks and may not always be the best choice for evaluating the performance of a particular model (Borji, 2022). Some of the general disadvantages of FID are that the Gaussian assumption in FID calculation might not hold in practice, the FID has high bias, and the sample size to calculate FID has to be large enough (usually above 50k), otherwise it would lead to an over-estimation of the actual FID (Borji, 2022). Moreover, it is computationally expensive (Mathiasen and Hvilshøj, 2020). For the specific application of unsupervised cross-domain classification such as the framework mentioned in Al-Hindawi et al. (2023), the FID was helpful in selecting a relatively good model, but it had problems. Most of the time it was not able to pick the best possible model; moreover, the FID model ranking was not correlated with the true model ranking based on the true supervised classification metrics, making the model selection choice seem random and unexplainable. It was stressed in Al-Hindawi et al. (2023) that there is a need for a framework to properly assess the validation datasets in order to improve the UI2I translation model selection criteria in unsupervised cross-domain classification models. The reason why traditional GAN evaluation metrics such as the FID are not suitable to support cross-domain classification is that they were not designed for that purpose, but rather were designed to evaluate the model in terms of the quality of the generated image from a human-eye perspective, and its ability to generate diverse results (not falling into mode collapse). Moreover, they do not take advantage of prior domain knowledge available in classification tasks such as the number of classes expected.

This work proposes a framework for evaluating UI2I translation models designed to support cross-domain classification applications using pseudo-supervised metrics. In this proposed approach, the Inception model is utilized to extract the features from the unlabeled target Dataset, but unlike other methods such as the Inception Score (IS) or the FID, this approach utilizes an unsupervised clustering technique known as the Gaussian Mixture Models to create pseudo labels which enable the use of standard supervised metrics. This methodology is shown not only to outperform unsupervised metrics such as the FID, but also is highly correlated with the true supervised metrics and mimics the monotonically decreasing behavior of their model ranking. This demonstrates the robustness and explainability of the metric, unlike the FID which is poorly correlated with the true supervised metrics and has inconsistent ranking order that is neither robust nor explainable. To showcase the efficiency of the methodology, the boiling crisis detection problem was used as an example. The boiling crisis, also known as critical heat flux (CHF), represents a formidable challenge in the field of thermal engineering and is of paramount importance to various industrial processes, particularly in nuclear reactors, electronics cooling, and power generation systems (Rassoulinejad-Mousavi et al., 2021). This phenomenon occurs when the rate of heat transfer from a heated surface to a boiling liquid abruptly deteriorates, leading to a sudden increase in surface temperatures and potentially catastrophic consequences (Rassoulinejad-Mousavi et al., 2021). Resolving the boiling crisis is critical as it can prevent the formation of vapor film insulating the surface, thereby averting equipment damage, reactor meltdowns, and ensuring the safety and efficiency of numerous technological applications (Rassoulinejad-Mousavi et al., 2021), making it a vital and urgent research frontier. Recent efforts were dedicated to alleviating this problem using machine learning techniques (Rassoulinejad-Mousavi et al., 2021; Rokoni et al., 2022) and most recently (Al-Hindawi et al., 2023) developed a cross-domain classification framework using UI2I translation models to tackle this problem. Their work serves as a suitable case study for the method proposed in this manuscript, especially since using an unreliable metric such as FID was one of the limitations. The same datasets used by the authors in Al-Hindawi et al. (2023) were also used in this work. Two experiments were conducted using the two publicly available datasets ($DS1$ and $DS2$) by alternating the target and source datasets for each experiment ($DS1 \rightarrow DS2$ and $DS2 \rightarrow DS1$).

To summarize the contribution of this manuscript:

- This work introduces a new framework for evaluating UI2I translation models using pseudo-supervised metrics. The framework was designed specifically to support cross-domain classification.
- The introduced framework utilizes an unsupervised clustering technique (GMM) to cluster the extracted features into N clusters, where N is the number of classes known as a prior from domain knowledge, and use these clusters as pseudo labels to enable the use of standard supervised metrics.
- The framework not only outperforms unsupervised metrics such as the FID, but is also highly correlated with the true supervised metrics, robust, and explainable.

The rest of the manuscript is organized as follows, Section 2 discusses relevant GAN-based I2I translation studies and related evaluation frameworks. Section 3 describes the proposed framework and the

Table 1
Strengths and weaknesses of common GAN evaluation metrics.

Metric	Strengths	Weaknesses
MMD	• Versatile metric applicable to various data distributions. • Not dependent on Inception model. • Captures mode collapse and diversity.	• Choice of kernel function and hyperparameters influence performance. • Computationally expensive. • Not designed for cross-domain classification.
KID	• Captures quality and diversity. • Less sensitive to Inception model. • More interpretable than FID.	• Depends on Inception model. • Computationally expensive. • Not designed for cross-domain classification.
IS	• Encourages visually appealing and diverse images. • Captures quality and diversity.	• Biased towards ImageNet. • Sensitive to model parameters. • Requires a large sample size. • Insensitive to intra-class diversity • Not interpretable. • Does not compare source images with target images. • Not designed for cross-domain classification.
FID	• Correlates with human judgment. • Captures quality and diversity. • Relatively stable and consistent. • Can detect intra-class mode collapse.	• Depends on Inception model. • Not interpretabile. • Not designed for cross-domain classification. • Requires a large sample size. • Biased towards ImageNet.

analysis procedures used in the study. Section 4 details the conducted experiments and Section 5 showcases and discusses the results. Finally, the manuscript is concluded in Section 6, followed by a listing of the sources cited afterward.

2. Related work

This section is composed of two parts. The first part will briefly discuss the evolution of GANs in the literature and expand on their utilization in both supervised and unsupervised image-to-image translation. The second part discusses the most common GAN evaluation measures and compares their advantages and disadvantages.

2.1. GANs and image-to-image translation

First introduced by Goodfellow et al. (2014), GANs consist of two networks, a generator and a discriminator that are trained together for an overall objective of generating realistic synthetic data that resembles a given dataset. The generator takes a random noise vector as input and transforms it into synthetic data that aims to resemble the real dataset. The discriminator evaluates both the real and the fake generated data samples and assigns scores indicating their authenticity. During training, the generator seeks to maximize these scores, while the discriminator strives to correctly distinguish between real and fake data. This adversarial process continues iteratively until the generator generates data that is indistinguishable from real data, resulting in GANs capable of generating high-quality synthetic data samples (Goodfellow et al., 2014).

The development of GANs has been the subject of numerous research papers and has led to the introduction of various variations and extensions of the original GAN architecture, such as Conditional GANs (Mirza and Osindero, 2014), InfoGANs (Chen et al., 2016), Progressive GANs (Karras et al., 2017), and Wasserstein GANs (Arjovsky et al., 2017). The first work to utilize GANs to solve the I2I translation problem was (Isola et al., 2017). In their pix2pix model, they used conditional adversarial networks to learn the mapping from an input image to an output image, where the networks learn a loss function to train this mapping. This method uses a "U-Net" based architecture for the generator and a "PatchGAN" classifier for the discriminator. Multiple efforts were spent to improve and build upon the pix2pix model and overcome its weaknesses. The major pitfall of the pix2pix model was that it was supervised. The training process required paired images in the training set for the model to learn the mapping $G : X \rightarrow Y$. Thus, to solve this problem, Zhu et al. (2017) introduced CycleGAN. The authors coupled the adversarial loss with an inverse mapping $F : Y \rightarrow X$ and introduced a cycle consistency loss. The objective of this loss is to enforce $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. A similar approach was performed by Yi et al. (2017) and Kim et al. (2017) concurrently with cycleGAN. Building on these works, Choi et al. (2018) introduced their StarGAN framework that simultaneously trains multiple datasets with different domains using a single generator and discriminator pair. However, StarGAN tends to change the images unnecessarily during image-to-image translation even when no translation is required (Rahman Siddiquee et al., 2019). To address this issue, Rahman Siddiquee et al. (2019) proposed the Fixed-Point GAN (FP-GAN) framework. This framework focused on identifying a minimal subset of pixels for domain translation and introduced fixed-point translation by supervising same-domain translation through a conditional identity loss and regularizing cross-domain translation through revised adversarial, domain classification, and cycle consistency losses.

2.2. GANs evaluation metrics

With the rapid rise of GANs and their use in various applications, the need for evaluation metrics to assess these models became increasingly critical. Traditional image evaluation measures such as the peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) were mainly designed to support tasks related to image compression and restoration where the image-to-image similarity was of utmost importance. Thus, they focused on measuring the similarity between images and were not suitable for image synthesis tasks. Depending on the application, there are two main types of GAN evaluation measures currently used, qualitative and quantitative measures. In this section, the focus will be on quantitative measures since the proposed work falls under that category. Currently, the most common quantitative GAN evaluation measures are the IS (Salimans et al., 2016), the FID (Heusel et al., 2017), the Maximum Mean Discrepancy (MMD) (Gretton et al., 2008) and the Kernel Inception Distance (KID) (Bińkowski et al., 2018). This section discusses these metrics and Table 1 summarizes their strengths and weaknesses.

2.2.1. Maximum Mean Discrepancy (MMD)

The Maximum Mean Discrepancy (MMD) is a statistical measure used to quantify the discrepancy between two probability distributions (Wynne and Duncan, 2022). It provides a way to assess the dissimilarity between samples drawn from different distributions and is commonly used in machine learning and generative modeling to compare the distributions of real and generated data samples (Wilson

et al., 2016; Li et al., 2015). The MMD measure is based on the idea of comparing the means of feature representations of samples from each distribution (Wynne and Duncan, 2022). By calculating the MMD, it can be determined whether two sets of data come from the same distribution or differ significantly (Wynne and Duncan, 2022). The general formula for MMD can be found in Table 2. In this equation, $MMD_u^2(X,Y)$ represents the squared unbiased MMD statistic between X and Y. The variables m and n represent the numbers of samples in distributions X and Y, respectively. The variables x_i and y_i represent individual samples from distributions X and Y, respectively. k denotes the kernel function applied to samples from two distributions. The equation consists of three main terms. The first term calculates the average kernel similarity between all pairs of samples within distribution X. The second term calculates the average kernel similarity between all pairs of samples within distribution Y. The third term calculates the average kernel similarity between samples from distribution X and distribution Y. By comparing these three terms, the MMD quantifies the discrepancy or difference between the distributions X and Y.

Main advantages of MMD is the ease of implementation and the rich kernel based theory behind it, making it a versatile metric that can be applied to various types of data distributions. It is less dependent on the specific architecture of the feature extractor and can also capture both the mode collapse and diversity of generated samples. It is however sensitive to the choice of kernel function and can be computationally expensive for large datasets.

2.2.2. Kernel Inception Distance (KID)

The KID metric is a method used for evaluating the quality and diversity of generated images in the field of generative adversarial networks (GANs). The KID utilizes the MMD metric introduced earlier to measure the discrepancy between the features extracted from real and generated images using the InceptionV3 neural network. The general mathematical expression of the KID is listed in Table 2. where f_{real} and f_{fake} represent the extracted features from real and fake images using the inception model.

KID has multiple advantages. It enables a more objective evaluation of GAN performance and it considers both the intra-class and inter-class variations of features, capturing the high-level semantics of images. Moreover, KID is model-agnostic and can be applied to evaluate GANs trained with different architectures and datasets. However, the KID also suffers from a number of disadvantages. It assumes that the Inception network's feature representations are sufficient to capture the quality and diversity of images even though it may not capture all aspects of image quality. It also does not capture lower-level image characteristics, such as pixel-level details and spatial coherence; because it used the extracted features in its calculations. Furthermore, the choice of the kernel function in KID can heavily influence the metric.

2.2.3. Inception Scores (IS)

The IS Eq. listed in Table 2. is a measure of the quality and diversity of generated images, based on the KL divergence between the predicted and true class distributions of a pre-trained Inception network (Borji, 2022; Treder et al., 2022). where $\mathbb{E}_{x\sim p_{\text{gen}}}$ is the expectation over the images sampled from the generator, $\mathbb{KL}$ refers to the Kullback–Leibler divergence, $p(y)$ is the marginal distribution of class labels, and $p(y \mid x)$ represents the conditional distribution of class labels y given an input image x.

Although used commonly, The IS has some downsides. The IS does not capture intra-class diversity and lacks correlation with the human judgment of image quality and may produce inconsistent results when compared to human evaluations (Barratt and Sharma, 2018). Moreover, it is insensitive to the prior distribution over labels (hence is biased towards ImageNet dataset and Inception model), and is very sensitive to model parameters and implementations (Borji, 2022). The IS also requires a large sample size to be reliable.

2.2.4. Fréchet Inception Distance (FID)

Similar to IS, the FID depends on the Inception model to generate its value, the difference, however, is that the FID calculates the Wasserstein-2 (a.k.a Fréchet) distance between multivariate Gaussians fitted to the embedding space of the Inception-v3 network of generated and real images (Borji, 2022; Treder et al., 2022). The general mathematical expression of the FID is listed in Table 2. In this equation, μ_g and μ_r represent the means of the feature maps of the generated images and real images, respectively, and C_g and C_r represent the covariance matrices of the feature maps of the generated images and real images, respectively. Tr represents the trace operator, which sums the diagonal elements of a matrix. Unlike IS, the FID is more consistent with human inspection, is sensitive to minimal changes in the real distribution, and can detect intra-class mode collapse (Borji, 2022). That being said, the FID has shortcomings as well. For example, the Gaussian assumption in FID calculation is not always valid, the FID has high bias, and requires a large sample size ($\geq 50k$ images) to be efficient (Chong and Forsyth, 2020).

Rahman Siddiquee et al. (2023) showed that the I2I model selected by FID has a weak correlation with the target classification task; therefore, the I2I model selected by FID performs poorly on the classification task. As a result, they proposed a pseudo-AUC metric for their anomaly detection task. Although this work was inspired by their work, their proposed pseudo-AUC metric cannot be applied directly to our task as they had images partially annotated in their problem setting.

Despite these efforts and those of others, these metrics either focus solely on the diversity of the results, the image quality from a human-eye perspective or require a portion of the target domain to be partially annotated (as the case with the pseudo-AUC). A metric or framework designed to support fully unsupervised cross-domain classification frameworks has yet to emerge.

3. Methodology

This section describes the methodology behind the proposed metric. The entire framework is summarized in both Fig. 1 and Algorithm 1. As shown, the methodology is broken down into smaller parts and described in detail in each sub-section as follows. Step 1 in Algorithm 1 (depicted in Part A) in Fig. 1 represents the source classification model training, which is further explained in 3.1. Steps 2 and 3 in Algorithm 1 (depicted in Part B) in Fig. 1 represent the UI2I model training and the cross-domain image translation process. Detailed explanation in Sections 3.2 and 3.3 respectively. Finally, Steps 4 and 5 in Algorithm 1 (depicted in Parts C and D) in Fig. 1 explain how the pseudo labels were created and incorporated to generate the proposed pseudo-supervised metrics. Details are documented in Section 3.4. The demonstrated figures in this section show only one of the experiments, which is when $DS1$ is used as a source dataset and $DS2$ as a target dataset ($DS1 \rightarrow DS2$). The same logic applies to the other experiment.

3.1. Source classification model training

The source dataset is split into three subsets, training, validation, and testing. A classification model is then trained on the training set for a pre-set number of iterations and the model is validated after every epoch using the validation dataset. The model that scores the best on the validation dataset is saved. Afterward, the best-saved model is tested on the testing set for final evaluation. For the purpose of these experiments, a convolutional neural network (CNN) was used as a classifier, but the methodology is agnostic to the type of classification model used.

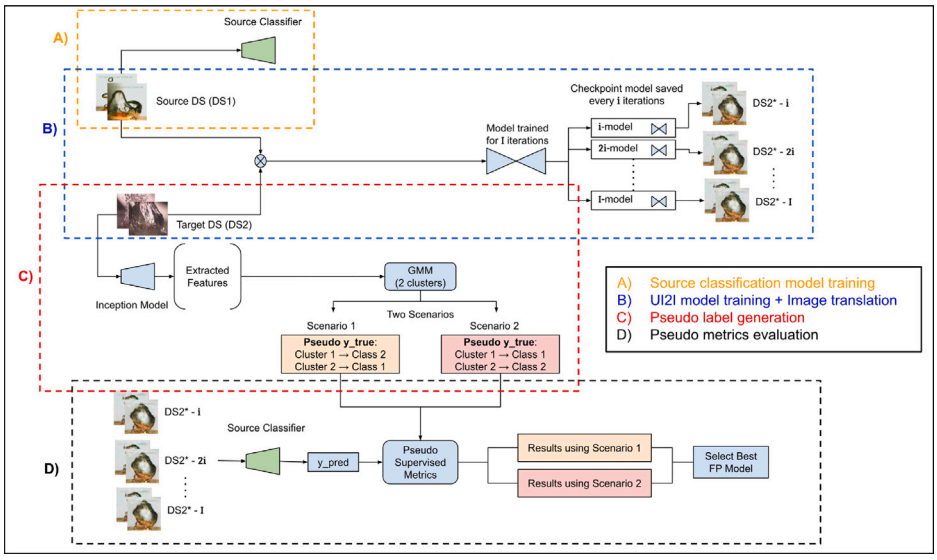

Fig. 1. DIPS framework: part (A) in orange shows the classification model training, part (B) shows the UI2I model training and the image translation, part (C) shows the pseudo label generation and part (D) shows the final pseudo metrics evaluation.

Algorithm 1: DIPS Algorithm

Step 1: Function `train_source_classifier(`*source_DS*`)`:
    train model
    save best model
    source-DS classifier = best saved model
    return *source-DS classifier*

Step 2: Function `train_UI2I_model(`*source_DS,target_DS*`)`:
    initialize models_list = [ ]
    use source_DS and target_DS to train UI2I translation model
    run training for I iterations
    save a checkpoint model every i iterations
    append saved checkpoint model to models_list
    return *models_list [i-model, 2i-model, ... , I-model]*

Step 3: Function `translate_target(`*target_DS,models_list*`)`:
    initialize translated_sets_list = [ ]
    For each model in models_list:
        translate target_DS to source domain using model
        append translated images set to translated_sets_list
    return *translated_sets_list* $[source\_DS^* - i,\ source\_DS^* - 2i,\ \dots\ ,\ source\_DS^* - I]$

Step 4: Function `generate_pseudo_labels(`*target_DS,GMM,Inception_model*`)`:
    extract features from the target_DS
    use GMM to cluster the data into 2 clusters
    assign labels to clusters covering both possible scenarios:
        Scenario 1: {cluster 1 = label 1, cluster 2 = label 2 }
        Scenario 2: {cluster 1 = label 2, cluster 2 = label 1 }
    return *Pseudo_*y_{true}*(scenario 1), Pseudo_*y_{true} *(scenario 2)*

Step 5: Function `pseudo_supervised_metrics(`*translated_sets_list,Pseudo*y_{true}*,Source-DS classifier*`)`:
    initialize models_results_list = [ ]
    For each translated_set in translated_sets_list:
        y_{pred} = Source-DS_classifier(translated_set)
        generate Pseudo metrics using y_{pred} and pseudo y_{true} from scenario 1
        generate Pseudo metrics using y_{pred} and pseudo y_{true} from scenario 2
        append best Pseudo metrics results to models_results_list
    best_UI2I_model = best(models_results_list)
    return *best_UI2I_model*

3.2. UI2I model training process

Fig. 2 summarizes the UI2I translation training process. In order for enabling the source classification model to correctly classify images from the target dataset; which is an unlabeled dataset coming from a different domain that the source classification model has not seen before, an unsupervised Image-to-Image (UI2I) translation Generative Adversarial network (GAN) was employed to translate the images in the target dataset from their domain to the source domain, so that they look like images familiar to the source classification model. For the purpose of demonstrating the methodology Fixed-Point GAN (FP-GAN) (Rahman Siddiquee et al., 2019) was employed as the UI2I translation model. FP-GAN was designed to support domain adaptation by identifying a minimal subset of pixels for domain translation and has shown superiority over other models in domain translation tasks. This being said, the methodology is agnostic to the type of UI2I translation model used. The UI2I translation model is trained for I number of iterations and a model is saved every i iterations during training, making a total of I/i checkpoint models saved. Both I and i are hyper-parameters that need to be tuned.

3.3. Cross-domain image translation process

Once the UI2I translation model training is complete, the framework uses each of the saved i/I checkpoint models to translate the target validation set from the target domain to the source domain as shown in Fig. 3. This is done in order to evaluate which model is the best one to be used in deployment.

3.4. Pseudo supervised metrics

The problem now is knowing which of these I/i UI2I translation models to use for model deployment. Since the target dataset is unlabeled, supervised metrics cannot be used to evaluate the validation set. The go-to metric in I2I translation is either the Inception Score (IS) or the Frechet Inception Distance (FID). The problem with these metrics is that they do not guarantee the best model to be selected. To combat this problem, this novel framework is introduced to generate pseudo-labels for the target dataset before translation, which will eventually allow the use of the traditional supervised metrics to evaluate the best model, as demonstrated in Fig. 4. Any of the standard supervised metrics used for classification could be used as a pseudo metric once the pseudo labels were generated, but to demonstrate the methodology, the results are showcased using both the balanced accuracy and the AUC metrics and then compared against the IS, FID, KID, and MMD. The metrics used in our experiments and their mathematical definition are listed in Table 2.

The method starts by using the pre-trained inception model to generate features from the validation set of the target dataset prior to translation. Once the features were extracted, the idea is to use a clustering technique to separate the two classes of data, thus creating pseudo labels which will allow using the traditional supervised metrics to evaluate the models. The clustering technique adopted to perform this task was GMMs. Using GMMs to cluster images and image features has been widely adopted in the literature (Bakheet et al., 2023; Pu et al., 2023; Hou et al., 2014; Kermani et al., 2015) for multiple reasons: (1) GMMs are suitable for clustering tasks where the underlying data

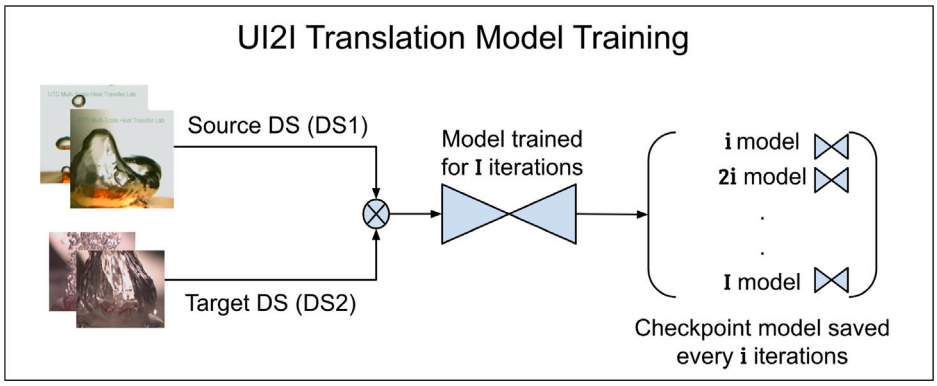

Fig. 2. UI2I training process.

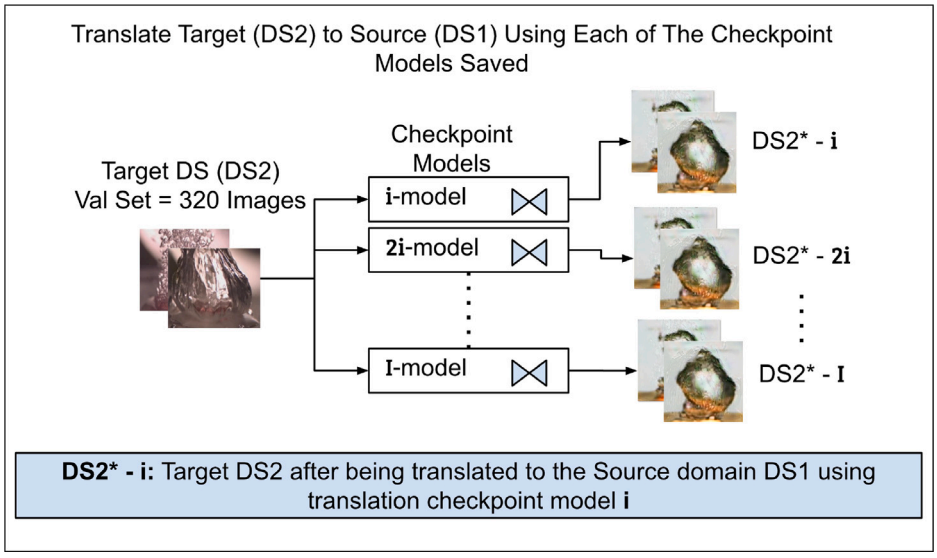

Fig. 3. Image translation process.

Table 2
Metrics definitions.

Metric name	Mathematical formulation
AUC	$= \int_{-\infty}^{\infty} \mathrm{TPR}(t) \cdot \mathrm{FPR}'(t)\, dt$
Balanced accuracy	$= \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP+FN}} + \frac{\mathrm{TN}}{\mathrm{TN+FP}}\right)$
FID	$= \lVert\mu_P - \mu_Q\rVert^2 + \mathrm{Tr}(C_P + C_Q - 2(C_P \cdot C_Q)^{1/2})$
Inception score	$= \exp\left(\mathbb{E}_x\left[D_{KL}(P(y\|x)\|P(y))\right]\right)$
MMD	$= \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k\left(x_i, x_j\right) + \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k\left(y_i, y_j\right) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k\left(x_i, y_j\right)$
KID metric	$= MMD_{poly}\left(f_{\text{real}}, f_{\text{fake}}\right)^2$

distribution is not well-defined or contains multiple subpopulations such as the case with boiling images. (2) GMMs can provide robust clustering results even in data consisting of high-dimensional features such as image features. (3) GMMs allow the incorporation of prior knowledge about the data distribution through the initialization of the expected number of clusters. such as in the case of the CHF detection problem. Thus, after extracting the features, a Gaussian Mixture Model is used with a pre-set number of N clusters, where N is the number of classes known from prior domain knowledge. This will group the images into N clusters which are used as pseudo labels (or pseudo-classes) for the unlabeled data. For the purpose of the experiments conducted in this work, the number of classes set by prior domain knowledge is equal to two ($N = 2$).

Now that the pseudo labels are obtained, they could be used to generate the pseudo-supervised metrics. For each translated validation set of the target dataset, the source classification model is used to generate predictions. The generated predictions are then compared with the pseudo labels using the traditional supervised metrics equations, thus providing "pseudo" supervised metrics. Note that since it is unknown which cluster represents which actual label (class), all possible scenarios are explored. The average of the pseudo metric for all models is then calculated for all scenarios, and the best-scoring scenario is adopted. Finally, the best scoring model from the best scoring scenario is selected for deployment. The pseudo metrics evaluation is described in Fig. 5.

It is worth mentioning that employing a pre-trained Inception model for feature extraction has the potential of introducing bias towards the ImageNET dataset. However, it is important to note that In contrast to methods like FID which uses summarized statistics of the features in their metric calculation making it more susceptible to the bias present in the pre-trained model. DIPS does not use neither the features nor their summarized statistics directly in metrics calculations. In DIPS the features obtained from the Inception model are subjected to GMM clustering. The objective here is to group similar features together based on their underlying distribution in the data. GMM clustering operates independently of the original dataset's bias because it seeks to identify patterns and relationships within the feature space, not the dataset from which the features were extracted. Once the clusters

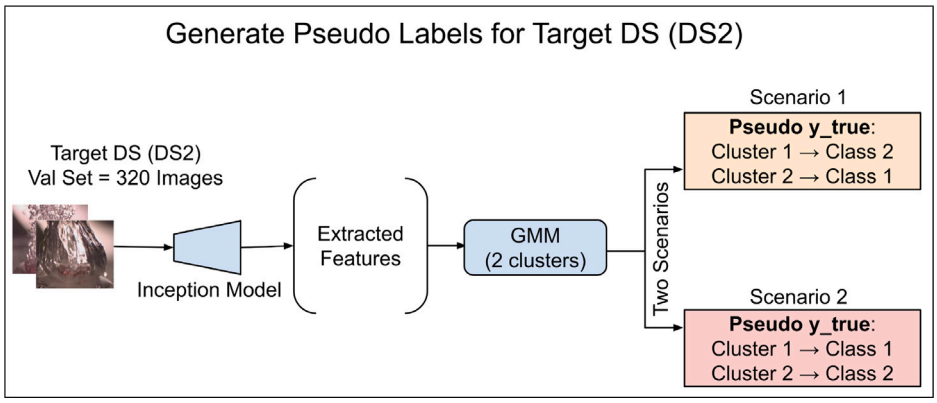

Fig. 4. Pseudo labels generation.

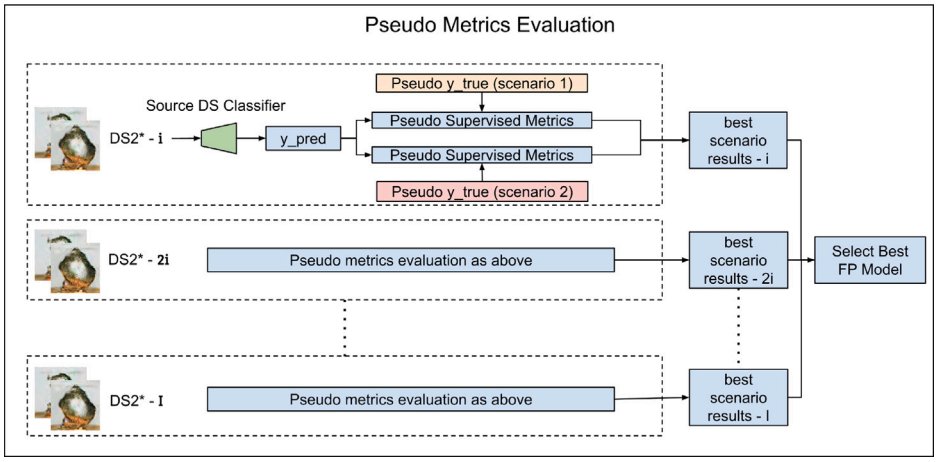

Fig. 5. Pseudo metrics evaluation.

are obtained through GMM, the Inception-extracted features are no longer used for metric calculation. Instead, the clusters are relied on to generate pseudo-labels. The subsequent metric calculations are based on these pseudo-labels, not the original features. This step ensures that the bias from the pre-trained Inception model is mitigated and does not directly influence the generated metrics.

4. Experiment

4.1. The boiling crisis detection problem

Over the past few decades, the study of heat transfer mechanisms has become a focal topic for researchers around the world. Heat transfer mechanisms are critical in various industrial applications (Dirker et al., 2019; Birbarah et al., 2020; El-Genk, 2012; Kandlikar, 2014; Fenech, 2013). One of the widely implemented heat transfer mechanisms is boiling heat transfer, a mechanism that utilizes the latent heat of the working fluid to dissipate a large amount of heat with minimal temperature increase (Rassoulinejad-Mousavi et al., 2021). Despite its wide application and the amount of effort spent studying boiling heat transfer, this mechanism comes with a dangerous drawback known as the boiling crisis. The boiling crisis is the phenomenon where the heat flux of boiling reaches a critical bound known as the Critical Heat Flux (CHF), after which the heating surface will be covered by a blanket of continuous vapor layer that adversely affects heat dissipation by depreciating the heat transfer coefficient (Al-Hindawi et al., 2023). This is critically dangerous because the improper heat dissipation will lead to a quick temperature upraise on the heater surface beyond its capability and eventually cause it to break down. Many efforts were dedicated to investigating the applicability of machine learning algorithms in CHF detection using a variety of techniques and data types. Whether it was acoustic emissions (Sinha et al., 2021), optical images (Rokoni et al., 2022), thermographs (Ravichandran et al., 2021) or whether it was using a variety of supervised learning algorithms, including support vector machine (Hobold and Silva, 2018), multilayer perceptron (MLP) neural networks (Hobold and Silva, 2018), transfer learning (Rassoulinejad-Mousavi et al., 2021), and most recently researchers (Al-Hindawi et al., 2023) started using frameworks supported by UI2I translation models to solve the cross-domain classification problem in boiling crisis detection such as the example used in this work to showcase our methodology.

4.2. Dataset

In this work, two different pool boiling experimental image datasets (DS-1 and DS-2) were prepared, where both DS-1 and DS-2 were generated using publicly available videos (You, 2014; Minseok et al., 2014). Specifically, the video from which DS-1 was prepared shows a pool boiling experiment performed using a square heater made of high-temperature, thermally-conductive microporous coated copper where the surface was fabricated by sintering copper powder. The square heater had a surface area of $\approx$100 mm^2 and the working fluid used was water. All experiments were performed at a steady-state under an ambient pressure of 1 atm. A T-type thermocouple was used for temperature measurements. The resolution of the video frames was 512 $\times$ 480 pixels. The YouTube video from which DS-2 was prepared shows a pool boiling experiment performed using a circular heater made of microporous-coated copper where the surface was fabricated

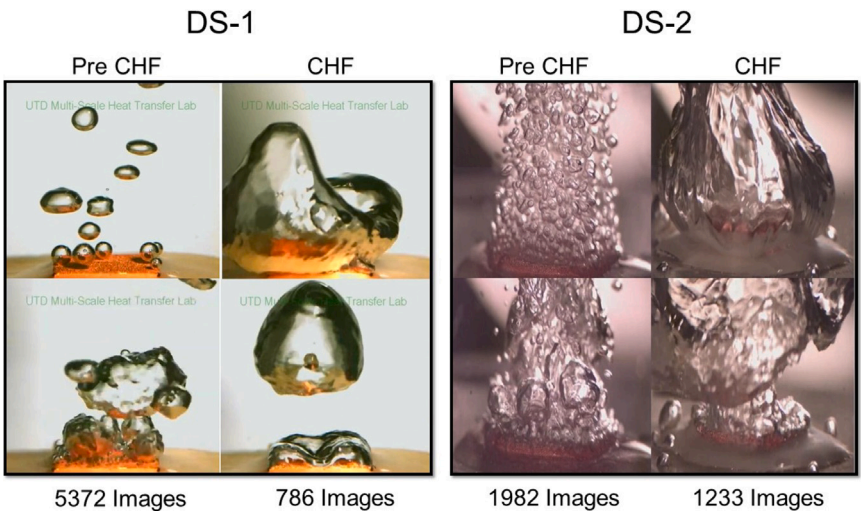

Fig. 6. Representative images of bubble dynamics from source videos.

Table 3
Datasets summary.

DS	Pre-CHF	CHF
DS-1	5372	786
DS-2	1982	1233

by sintering copper powder. The circular heater had a diameter of ≈16 mm and the working fluid used was DI water. All experiments were performed at a steady state under an ambient pressure of 50 kPa. A T-type thermocouple was used for temperature measurements. The resolution of the video frames was 1280 × 720 pixels.

Images for DS-1 and DS-2 were prepared by downloading the videos from YouTube and extracting individual frames using a MATLAB code via the VideoReader and imwrite functions. Recognizing duplicate frames extracted from the YouTube videos, quality control was conducted to remove the repeated images by calculating the relative difference using the Structural Similarity Index (SSIM) value (Gao et al., 2020) between two consecutive images where images with a relative difference less than 0.03% were removed. This pre-processing is important to ensure DL models were not biased by identical image frames.

The images were categorized into two boiling regimes: (1) The critical heat flux regime (CHF), where a significant drop in the heat transfer coefficient is observed due to a continuous vapor layer blanketing the heater surface and (2) pre-CHF regime, where optimal heat transfer coefficient is obtained and only discrete bubbles or frequent bubble coalescence is observed before departure. Originally, DS-1 had a total of 6158 images (786 CHF versus 5372 pre-CHF) and DS-2 had a total of 3215 (1233 CHF versus 1982 pre-CHF). As seen, both data sets were unbalanced. In each of the two experiments, only the training data for the source dataset was balanced before use. The target dataset was not balanced since the objective of this study is to introduce a framework that utilizes unsupervised learning, that is, the labeling information of the target datasets are assumed to be unavailable; Thus, impossible to balance using traditional oversampling or undersampling techniques. Table 3 shows the original number of images in each regime for each dataset and Fig. 6 shows a visual representation of the images for each dataset. The pixel intensity values in each image were normalized to fit in the range [0,1] to ensure uniformity over multiple datasets during deep learning training.

4.3. Framework pipeline

As depicted in Fig. 1, the proposed framework consists of four main parts.

Table 4
Experiment settings.

Parameter	Value
Image size	256 × 256
C_{dim}	1
Batch size	8
Number of workers	4
λ_{id}	0.1
Number of iterations (I)	300,000
Checkpoint saving frequency (i)	10,000
Learning rate for G and D	0.0001
Number of D updates per each G update	5
β_1 for Adam optimizer	0.5
β_2 for Adam optimizer	0.999
Data augmentation used	Random horizontal flips

4.3.1. Source classification model training

To diversify the training, different architectures for the classification model were used in each experiment. For the $DS1 \rightarrow DS2$ experiment, the same architecture in Al-Hindawi et al. (2023) was employed to train the model. The architecture for that model is summarized in Fig. 7. For the $DS2 \rightarrow DS1$ experiment, the ResNet50 model architecture (He et al., 2015) was employed. In both experiments, the data was split into three subsets: a training set (80%), a validation set (10%), and a testing set (10%). The models were trained for a total of 100 epochs and the best model was saved and used in our pipeline.

4.3.2. UI2I translation model training

For the UI2I translation, the FP-GAN model was employed. The model was trained using the same architectures for both the generator and the discriminator as implemented by the authors (Rahman Siddiquee et al., 2019). Table 4 summarizes the hyperparameters chosen for training the model and Table 5 summarizes the loss functions used in the model.

4.3.3. Pseudo labels generation

The pseudo labels generation process utilizes the InceptionV3 model which was pre-trained on the ImageNET dataset to extract the features from the target DS images after being resized to the dimensions expected by the InceptionV3 model (299 × 299). Specifically, the features were extracted using the "conv2d_93" intermediate layer of the model. The extracted features were then clustered into two clusters using the GMM provided by the Scikit-learn library. Other than the number of clusters, the default settings set by the library were used for GMM. Two possible scenarios are then formulated for the true labels (pseudo y_{true}):

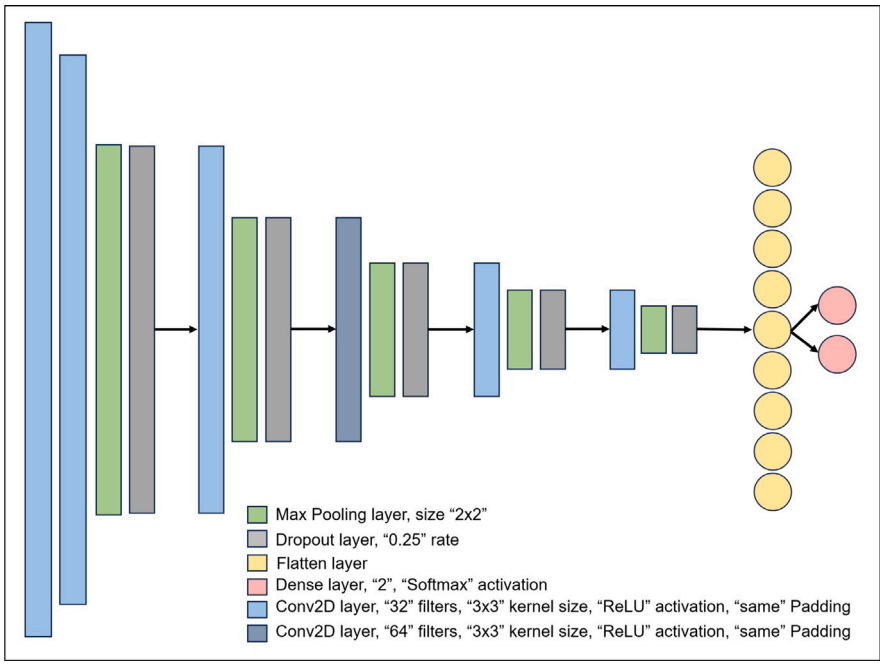

Fig. 7. Architecture for the source classification model for the $DS1 \rightarrow DS2$ experiment.

Table 5
Summary of the loss functions used in FP-GAN.

Equation	Loss	Definition
Eq. 1	L_{adv}	$= \sum_{c\in\{cx,cy\}} E_{x,c}[\log(1 - D_{r/f}(G(x,c)))] + E_x[\log D_{r/f}(x)]$
Eq. 2	L_{domain}	$= E_{x,cx}[\log D_{domain}(cx\|x)]$
Eq. 3	$L_{f_{domain}}$	$= \sum_{c\in\{cx,cy\}} E_{x,c}[-\log D_{domain}(c\|G(x,c))]$
Eq. 4	L_{cyc}	$= \sum_{c\in\{cx,cy\}} E_{x,cx}[\|\|G(G(x,c),cx) - x\|\|_1]$
Eq. 5	L_{id}	$= E_{x,c}[\|\|G(x,c) - x\|\|_1]$ if $c = cx$; 0 otherwise
Eq. 6	L_D	$= -L_{adv} + \lambda_{domain} L_{r_{domain}}$
Eq. 7	L_G	$= L_{adv} + \lambda_{domain} L_{f_{domain}} + \lambda_{cyc} L_{cyc} + \lambda_{id} L_{id}$

(a) cluster 1 = "Pre CHF" and cluster 2 = "CHF" and (b) cluster 1 = "CHF" and cluster 2 = "Pre CHF". The two scenarios are then used in the final step.

4.3.4. Pseudo metrics evaluation

The labels for each of the translated datasets from the UI2I translation step are predicted (y_{pred} using the same classification model from Section 4.3.1. The "balanced accuracy score" and the "roc auc score" from the standard supervised metrics provided by Scikit-learn were used to evaluate y_{pred} against the pseudo y_{true} generated from each scenario. The average of the pseudo metric for all 30 models is then calculated for both scenarios, and the best-scoring scenario is adopted. Finally, the best scoring model from the best scoring scenario is selected for deployment.

5. Results and discussion

Once the pseudo metrics were generated, the models generated by the FP-GAN training process can now be ranked according to these metrics and the best model can now be selected. Any of the regular metrics used for classification could be used as a pseudo metric once the pseudo labels were generated, but to demonstrate the methodology, the results are showcased using both the balanced accuracy and the AUC metrics.

To demonstrate the effectiveness of the framework, three key aspects must be addressed. Firstly, it must be assessed whether the proposed metric performs better than the current state-of-the-art GAN evaluation metrics in the field, including FID, KID, IS, and MMD. Secondly, it is essential to compare the proposed metric with the ideal scenario, where true supervised metrics are calculated using the actual labels. Finally, it should be examined how the proposed metric correlates with the actual true metric when compared to how the existing state-of-the-art metrics correlate with the actual true metric. This section addresses these aspects in two ways. Firstly, each of the saved checkpoint models is evaluated using the truly-supervised metric, pseudo-supervised metric, FID, KID, IS, MMD (linear Kernel), and MMD (Gaussian kernel). The models are then ranked according to each of the competing metrics and are plotted against the true actual supervised score of the ranked checkpoint models. In this comparison, each metric is judged by how close it is to mimicking the ranking behavior of the truly supervised metrics that had access to the annotations; thus allowing for an objective evaluation of both the proposed metric and the state-of-the-art metrics. The best metric is the one that could mimic the monotonically decreasing behavior of the true ranking line. Secondly, the true metric value is plotted against all the competing metrics and multiple linear and non-linear correlation values were generated for each plot, including R-squared, Pearson correlation, Spearman correlation, and Kendall rank correlation. The rationale here is to showcase the degree of correlation between the true supervised metric and all the competing metrics. The higher the correlation the better the metric is. The results in this section show that the proposed method not only outperforms the state-of-the-art but also heavily resembles the true ranking behavior of the true supervised metrics and is highly correlated with the true supervised metrics.

Fig. 8 shows the models' ranking results based on the true-supervised balanced accuracy for all competing metrics in the experiments ran with DS1 as a source Dataset ($DS1 \rightarrow DS2$). Fig. 9 shows the models' ranking results based on the true-supervised AUC for the same experiment. Figs. 10 and 11 show the same plots but for the ($DS2 \rightarrow DS1$) experiment where DS2 was used as the source dataset. The $x-axis$ represents the model ranking according to that metric from best to worst, while the $y-axis$ represents the real metric value of the ranked model. The ranking curve for the actual ranking will show a monotonically decreasing behavior. In all the mentioned plots, the true actual ranking is compared against (A) the pseudo-balanced accuracy

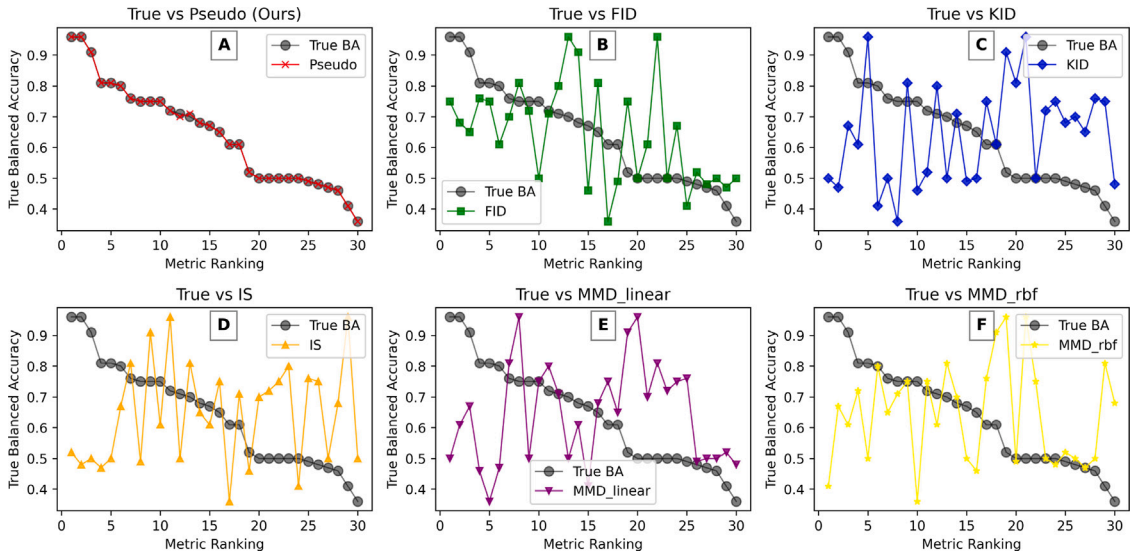

Fig. 8. Models ranking results vs. the true balanced accuracy for (A) pseudo-balanced accuracy (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS1 \rightarrow DS2$ experiment.

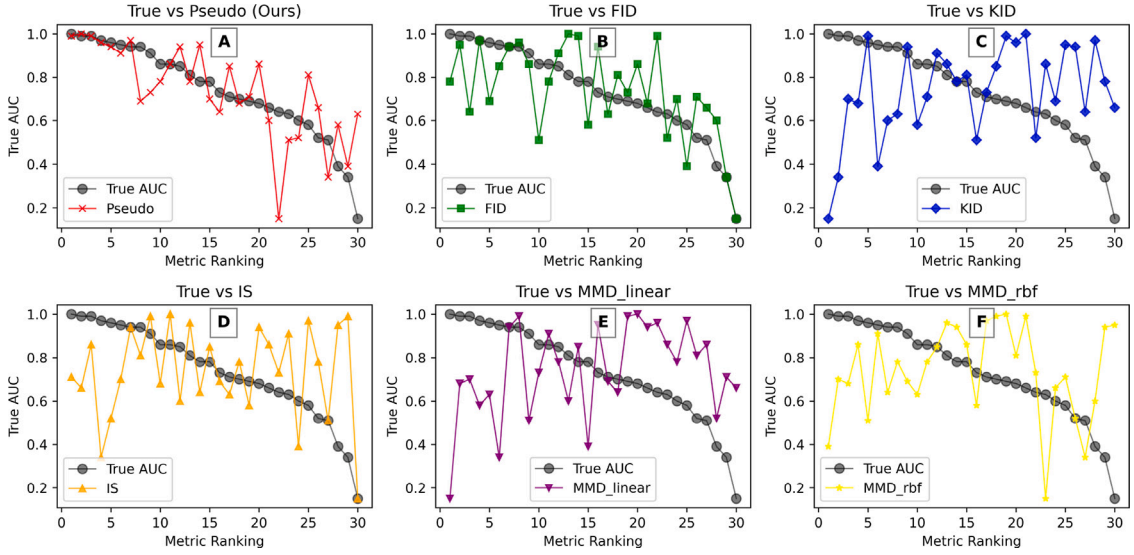

Fig. 9. Models ranking results vs. the true AUC for (A) pseudo-AUC (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS1 \rightarrow DS2$ experiment.

(ours), (B) the FID, (C) the KID, D) the IS, (E) the MMD (linear kernel), and (F) the MMD (gaussian kernel).

As seen from the figures, the pseudo metric ranking is consistent for both the balanced accuracy and the AUC metrics while all the other metrics show very inconsistent and random behavior and demonstrating that indeed none of these metrics are suitable for cross-domain classification. The robustness of the pseudo metrics is more evident when DS2 was used as a source dataset, even when the pseudo metrics deviate from the original ranking, they still pick a model that is close in terms of the actual metric value and returns to mimic the actual ranking behavior. Table 6 shows the true balanced accuracy and AUC results of the picked model using each competing metric for both experiments. Again, the proposed metric demonstrate both consistency and superiority over other metrics. Even when the metric comes second (only by a value of 0.01) as in the case of the balanced accuracy for the $DS2 \rightarrow DS1$ experiment, it still picks a very close model to the actual best. Moreover, it shows how inconsistent the state-of-the-art metrics are in cross-domain classification applications.

Fig. 12 shows the correlation results based on the true-supervised balanced accuracy vs all competing metrics in the experiments ran with DS1 as a source Dataset ($DS1 \rightarrow DS2$). Fig. 13 shows the same only using true-supervised AUC for the same experiment. Similarly, Figs. 14 and 15 show the correlations but for the ($DS2 \rightarrow DS1$) experiment where DS2 was used as the source dataset. The $x - axis$ represents the metric value while the $y - axis$ represents the real metric value. For each plot, the coefficient of determination (R^2), Pearson (r), Spearman (ρ), and Kendall (τ) correlations were calculated and presented on the plot.

The results from Figs. 12 and 13 show that for both the pseudo metrics there is an almost perfect correlation with the truly supervised metrics for the $DS1 \rightarrow DS2$ experiment. On contrast, all the other competing metrics show weak to none existing correlation with either the AUC or the balanced accuracy.

The results in Fig. 14 show that again the results for the pseudo metric is strongly correlated with the truly supervised balanced accuracy in the $DS2 \rightarrow DS1$ experiment, while again the other competing

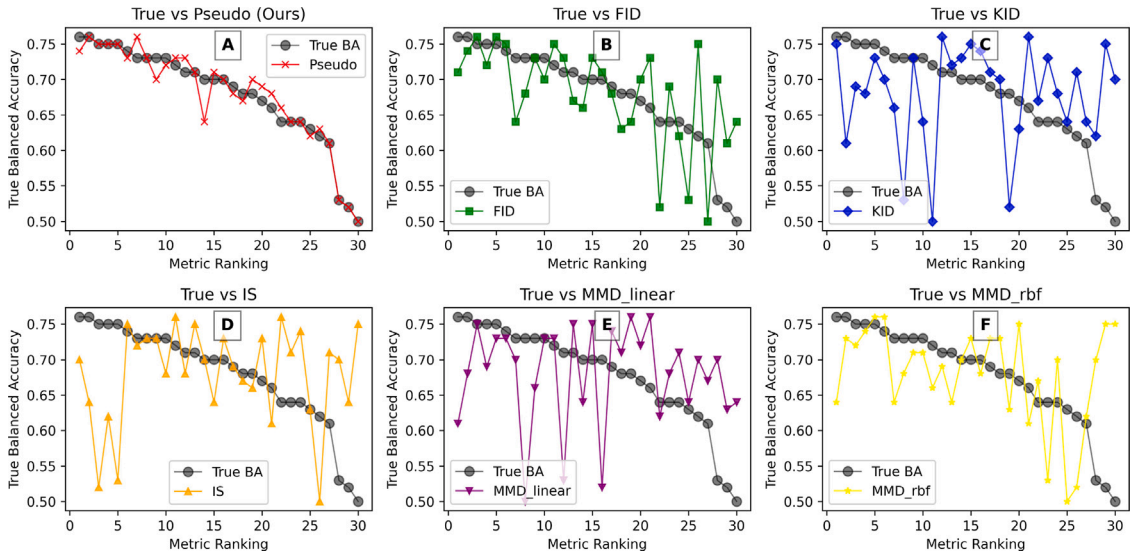

Fig. 10. Models ranking results vs. the true balanced accuracy for (A) pseudo-balanced accuracy (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS2 \rightarrow DS1$ experiment.

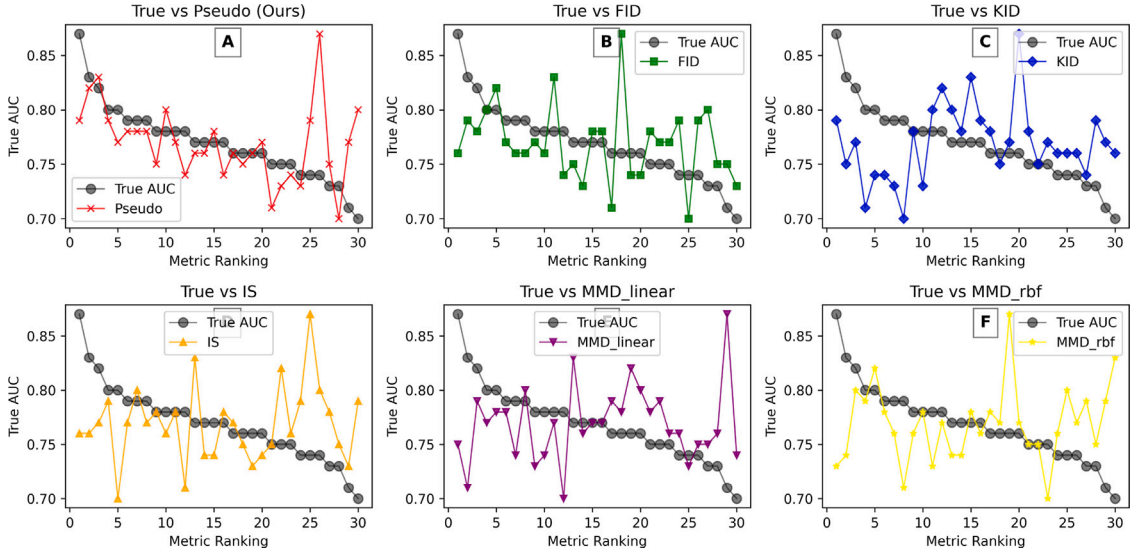

Fig. 11. Models ranking results vs. the true AUC for (A) pseudo-AUC (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS2 \rightarrow DS1$ experiment.

Table 6
Competing metrics comparison (the best is in bold and the second best is underlined).

Exp	Metric	True	Competing metrics					
			Pseudo (ours)	FID	KID	IS	MMD_L	MMD_G
$DS1 \rightarrow DS2$	Bal. Acc.	0.96	**0.96**	0.75	0.50	0.52	0.50	0.41
	AUC	1.00	**0.99**	0.78	0.15	0.71	0.15	0.39
$DS2 \rightarrow DS1$	Bal. Acc.	0.76	0.74	0.71	**0.75**	0.70	0.61	0.64
	AUC	0.87	**0.79**	0.76	**0.79**	0.76	0.75	0.73

metrics show weak to none existing correlation with the truly supervised balanced accuracy. Although both the FID and KID showed some improvement in the correlation values, this improvement is not to a level that would qualify them as a suitable candidate for cross-domain classification tasks and their performance remains far worse than that of the pseudo metric.

The correlation performance of all metrics against the truly supervised AUC for the $DS2 \rightarrow DS1$ experiment is shown to be poor in Fig. 15. However, the performance of the pseudo metric is still far better than all the other metrics on all of the correlation measures. One thing to note about this particular experiment is that while we acknowledge the presence of a weak correlation between our pseudo metric and the true metric, it is essential to highlight the uniqueness of the results in this specific experiment. The true metric results for all 30 models demonstrated an exceptionally tight cluster ($\mu = 0.77$ and $std = 0.035$). This consistency suggests that the models' performances were very close to one other according to the true metric. In such scenarios where the true metric values are essentially constant across all models,

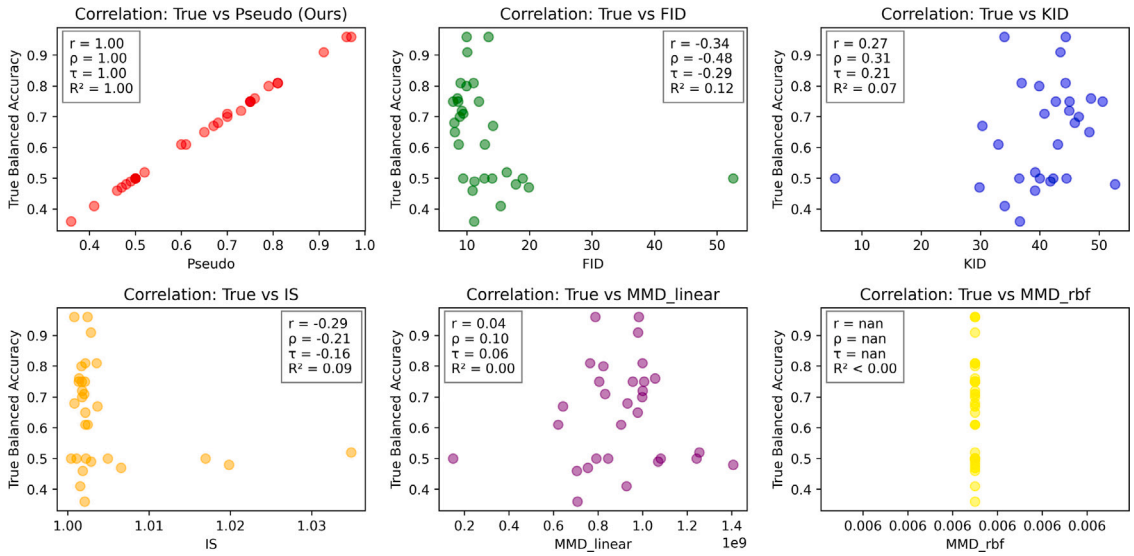

Fig. 12. Correlations results of true balanced accuracy vs. (A) pseudo-balanced accuracy (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS1 \rightarrow DS2$ experiment.

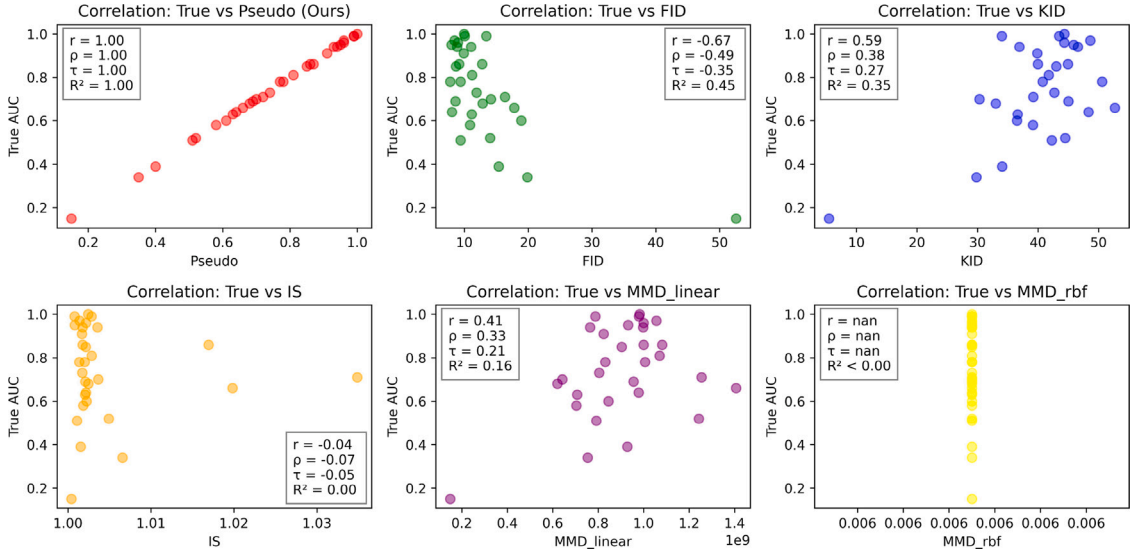

Fig. 13. Correlations results of true AUC vs. (A) pseudo-AUC (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS1 \rightarrow DS2$ experiment.

correlation measures may not be the most appropriate yardstick for assessing the worth of a metric; as in such cases, the performance remains consistently similar irrespective of which checkpoint model is chosen making the choice of the checkpoint model inconsequential in terms of the true metric. This explains why the pseudo metric exhibited an underperformance in this experiment.

The results from the correlation figures demonstrate yet again the superiority of the proposed metric over the state-of-the-art. Moreover, they also indicate that the proposed metrics are explainable since they are highly correlated with the explainable true supervised metrics. Moreover, the results also demonstrate that none of the current state-of-the-art metrics are consistent and they are uncorrelated with the true metrics, making their decision unexplainable and unreliable to assess UI2I translation models for unsupervised cross-domain classification tasks and should not be utilized in such applications.

The pseudo-supervised metric developed in this work will make a major impact on the boiling community. While many research groups have demonstrated visualization-based boiling regime classification or boiling crisis detection using AI models, most of the studies are based on single datasets, leaving model generalizability a major challenge (Hobold and Silva, 2018; Ravichandran and Bucci, 2019). As explored in the authors' previous work, a well-trained CNN model may only lead to an accuracy of 0.4–0.5 when classifying new data sets that are not included in the training (Rassoulinejad-Mousavi et al., 2021; Al-Hindawi et al., 2023). This accuracy can be improved using transfer learning, where a small amount of labeled data from the new data set are used to fine-tune pre-trained classifiers (Rassoulinejad-Mousavi et al., 2021). UI2I using GAN does not require labeled data from the new data sets but only leads to an accuracy of 0.75 when using state-of-the-art metric (FID) to select the best-performing generator (Al-Hindawi et al., 2023). The present work using the pseudo-supervised metric can lead to a significantly higher balanced accuracy of 0.96 and thus demonstrates that using UI2I with the pseudo-supervised metric, a pre-trained boiling regime classifier can be adapted to any new boiling data set without additional training or labeled data (see Fig. 16).

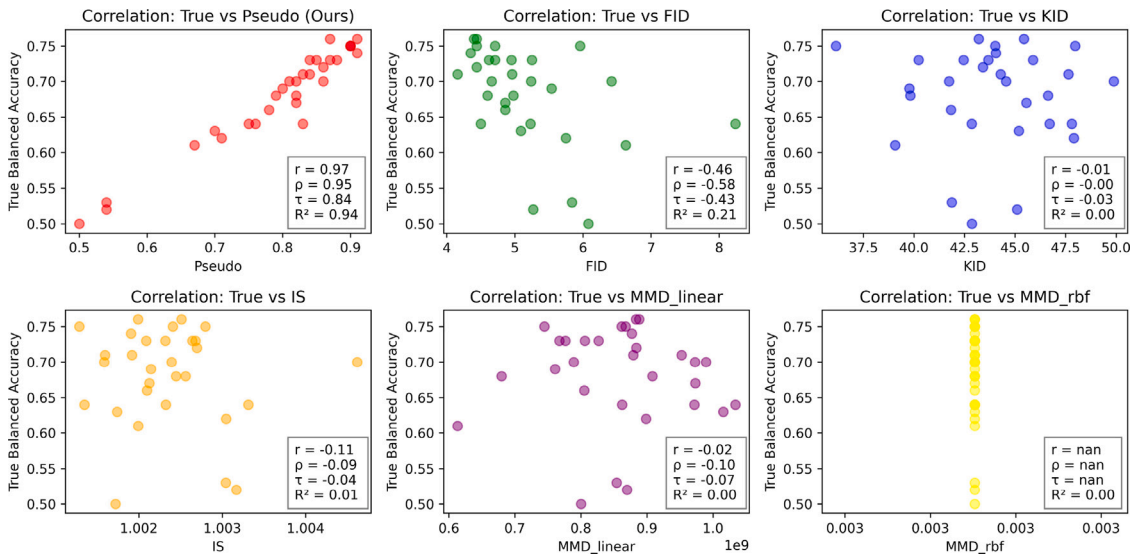

Fig. 14. Correlations results of true balanced accuracy vs. (A) pseudo-balanced accuracy (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS2 \rightarrow DS1$ experiment.

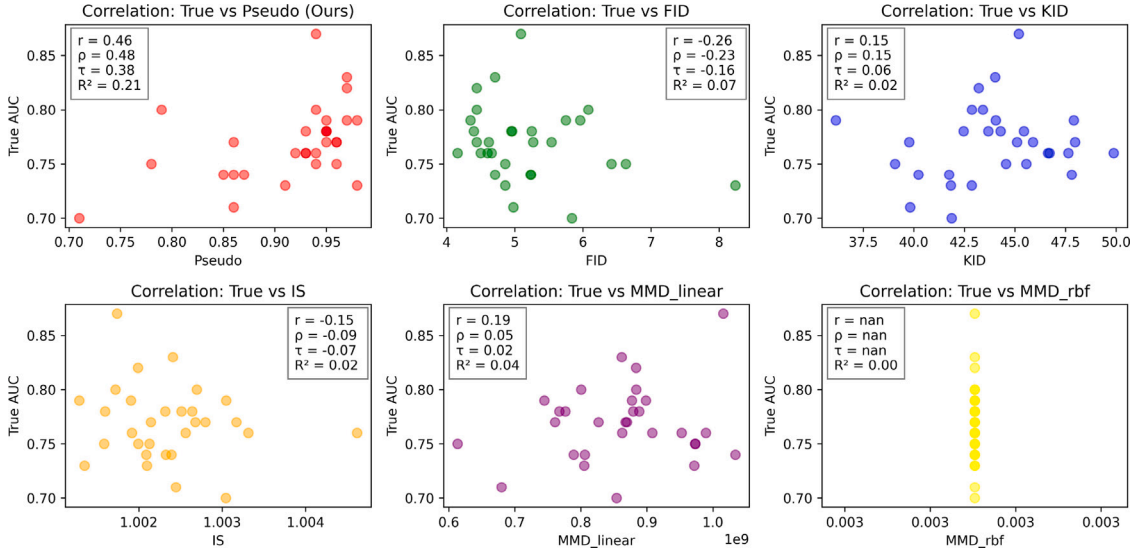

Fig. 15. Correlations results of true balanced accuracy vs. (A) pseudo-AUC (ours), (B) FID, (C) KID, (D) IS, (E) MMD (linear kernel), and (F) MMD (gaussian kernel) for the $DS2 \rightarrow DS1$ experiment.

6. Conclusion and future work

In this paper, a framework was introduced to evaluate UI2I translation models. The DIPS framework was designed specifically to support cross-domain classification applications using pseudo-supervised metrics. To showcase the efficiency of the framework, the boiling crisis detection problem was used as an example. The efficacy of the results was demonstrated by conducting two experiments using two publicly available datasets from different domains.

The proposed methodology was shown to not only outperform the state-of-the-art unsupervised metrics but was also shown to be highly correlated with the true supervised metrics, unlike the state-of-the-art which were poorly correlated with the true supervised metrics and were shown to be inconsistent. Moreover, it was displayed that typical state-of-the-art GAN evaluation metrics which were designed to evaluate models based on their ability to generate images that are both diverse and realistic to the human eye are not suitable to support cross-domain classification tasks as presented in the results section. In almost all comparisons, the ranking provided by the pseudo metric was superior to the state-of-the-art metrics and showed that it can mimic the monotonically decreasing behavior of the true metric; thus providing explainable and consistent results.

Although the proposed metric is showing great potential in evaluating UI2I translation models in cross-domain prediction frameworks, it still has room for improvement and future work. For starters, the proposed metric was only demonstrated to work in a binary classification setting ($N = 2$). An important factor to consider in the case of multi-class classification is that the problem becomes more computationally expensive as the number of clusters N increases. Specifically, in the cluster assignment step of the framework, the number of possible scenarios is expected to increase exponentially. For future work, we plan on addressing both of these issues by incorporating physics-assisted labeling extracted from either a physical source such as acoustic sensors mounted on the apparatus of the experiment, or extracted from the images themselves such as the count and size of the blobs in the image using computer vision segmentation techniques. Another issue is that the framework is dependent on having prior domain knowledge to set up the number of expected clusters N that are later used as pseudo

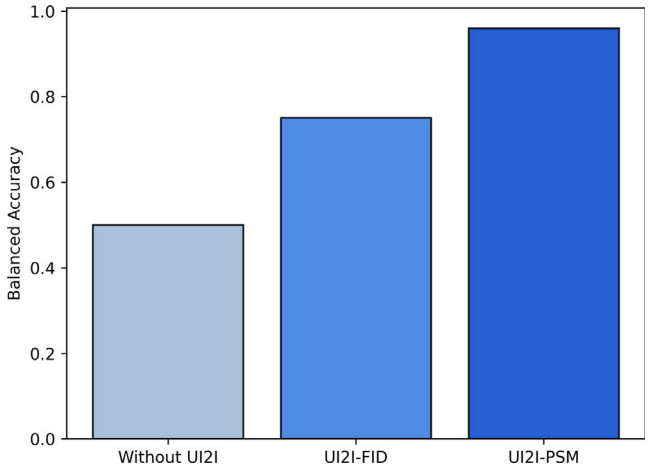

Fig. 16. Comparison of classification accuracy of CNN trained on DS-1 for classifying the boiling regime of DS-2 with no translation, UI2I with FID, and UI2I with the developed pseudo supervised metric (PSM).

labels. The method falls short in the case where no prior domain knowledge is available about the expected number of labels. Again, physics-assisted labeling extracted using segmentation could provide a solution to this issue. Furthermore, the use of a pre-trained Inception model for feature extraction introduces the potential for bias towards the ImageNET dataset. Although DIPS takes measures to mitigate this bias effect. it is important to acknowledge that there may still be some residual influence from the pre-trained model. Lastly, the method is strictly applicable to cross-domain classification problems and is not suitable to address cross-domain regression problems where the predicted value is continuous rather than discrete. If the data in the source DS is instead labeled with quantifiable heat flux value, then an interesting direction would be to explore the utilization of the temporal factor to relate each frame from the target DS with the source. Researchers are advised to study the mentioned issues and explore the capability of this metric to expand on its potential beyond this work.

CRediT authorship contribution statement

Firas Al-Hindawi: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Md Mahfuzur Rahman Siddiquee:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Teresa Wu:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Resources, Supervision, Project administration. **Han Hu:** Conceptualization, Writing – original draft, Writing – review & editing, Data curation. **Ying Sun:** Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Al-Hindawi, F., Soori, T., Hu, H., Rahman Siddiquee, M.M., Yoon, H., Wu, T., Sun, Y., 2023. A framework for generalizing critical heat flux detection models using unsupervised image-to-image translation. Expert Syst. Appl. 120265.

Alhindawi, F., Altarazi, S., 2018. Predicting the tensile strength of extrusion-blown high density polyethylene film using machine learning algorithms. In: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, pp. 715–719.

Altarazi, S., Allaf, R., Alhindawi, F., 2019. Machine learning models for predicting and classifying the tensile strength of polymeric films fabricated via different production processes. Materials 12 (9), 1475.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 214–223.

Bakheet, S., Mofaddel, M., Soliman, E., Heshmat, M., 2023. Content-based image retrieval using BRISK and SURF as bag-of-visual-words for naïve Bayes classifier. Sohag J. Sci. 8 (3), 329–335.

Barratt, S., Sharma, R., 2018. A note on the inception score. arXiv preprint arXiv:1801.01973.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Birbarah, P., Gebrael, T., Foulkes, T., Stillwell, A., Moore, A., Pilawa-Podgurski, R., Miljkovic, N., 2020. Water immersion cooling of high power density electronics. Int. J. Heat Mass Transfer 147, 118918.

Borji, A., 2022. Pros and cons of GAN evaluation measures: New developments. Comput. Vis. Image Underst. 215, 103329.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Adv. Neural Inf. Process. Syst. 29.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chong, M.J., Forsyth, D., 2020. Effectively unbiased fid and inception score and where to find them. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6070–6079.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J., 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 994–1003.

Dirker, J., Juggurnath, D., Kaya, A., Osowade, E.A., Simpson, M., Lecompte, S., Noori Rahim Abadi, S.M.A., Voulgaropoulos, V., Adelaja, A.O., Dauhoo, M.Z., et al., 2019. Thermal energy processes in direct steam generation solar systems: Boiling, condensation and energy storage–a review. Front. Energy Res. 6, 147.

El-Genk, M.S., 2012. Immersion cooling nucleate boiling of high power computer chips. Energy Convers. Manage. 53 (1), 205–218.

Fenech, H., 2013. Heat Transfer and Fluid Flow in Nuclear Systems. Elsevier.

Gao, F., Wu, T., Chu, X., Yoon, H., Xu, Y., Patel, B., 2020. Deep residual inception encoder-decoder network for medical imaging synthesis. IEEE J. Biomed. Health Inf. 24 (1), 39–49,. http://dx.doi.org/10.1109/JBHI.2019.2912659.

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., 2015. Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2551–2559.

Goel, P., Ganatra, A., 2023. Unsupervised domain adaptation for image classification and object detection using guided transfer learning approach and JS divergence. Sensors 23 (9), 4436.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets ian.

Gretton, A., Borgwardt, K., Rasch, M.J., Scholkopf, B., Smola, A.J., 2008. A kernel method for the two-sample problem. arXiv preprint arXiv:0805.2368.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. p. 30.

Hobold, G., Silva, A., 2018. Machine learning classification of boiling regimes with low speed, direct and indirect visualization. Int. J. Heat Mass Trans. 125, 1296–1309,. http://dx.doi.org/10.1016/j.ijheatmasstransfer.2018.04.156.

Hou, X., Zhang, T., Xiong, G., Lu, Z., Xie, K., 2014. A novel steganalysis framework of heterogeneous images based on gmm clustering. Signal Process., Image Commun. 29 (3), 385–399.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Kandlikar, S.G., 2014. Review and projections of integrated cooling systems for three-dimensional integrated circuits. J. Electron. Packag. 136 (2), 024001.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kermani, S., Samadzadehaghdam, N., EtehadTavakol, M., 2015. Automatic color segmentation of breast infrared images using a Gaussian mixture model. Optik 126 (21), 3288–3294.

Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 1857–1865.

Li, Y.-F., Lin, Y., Gao, Y., Khan, L., 2021. Cross-domain sentiment classification with attention-assisted GAN. In: 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI). IEEE, pp. 88–95.

Li, Y., Swersky, K., Zemel, R., 2015. Generative moment matching networks. In: International Conference on Machine Learning. PMLR, pp. 1718–1727.

Mathiasen, A., Hvilshøj, F., 2020. Backpropagating through fréchet inception distance. http://dx.doi.org/10.48550/ARXIV.2009.14075, URL https://arxiv.org/abs/2009.14075.

Minseok, H., Bertina, B., Graham, S., 2014. Pool boiling experiment.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Omeroglu, A.N., Mohammed, H.M., Oral, E.A., Aydin, S., 2023. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. Eng. Appl. Artif. Intell. 120, 105897.

Padmapriya, J., Sasilatha, T., 2023. Deep learning based multi-labelled soil classification and empirical estimation toward sustainable agriculture. Eng. Appl. Artif. Intell. 119, 105690.

Pu, Y., Sun, J., Tang, N., Xu, Z., 2023. Deep expectation-maximization network for unsupervised image segmentation and clustering. Image Vis. Comput. 104717.

Rahman Siddiquee, M.M., Shah, J., Wu, T., Chong, C., Schwedt, T.J., Dumkrieger, G., Nikolova, S., Li, B., 2023. Brainomaly: Unsupervised neurologic disease detection utilizing unannotated T1-weighted brain MR images. arXiv preprint arXiv:2302.09200.

Rahman Siddiquee, M.M., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J., 2019. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Rassoulinejad-Mousavi, S., Al-Hindawi, F., Soori, T., Rokoni, A., Yoon, H., Hu, H., Wu, T., Sun, Y., 2021. Deep learning strategies for critical heat flux detection in pool boiling. Appl. Therm. Eng. 190, 116940,. http://dx.doi.org/10.1016/j.applthermaleng.2021.116849.

Ravichandran, M., Bucci, M., 2019. Online, quasi-real-time analysis of high-resolution, infrared, boiling heat transfer investigations using artificial neural networks. Appl. Therm. Eng. 163 (September), 114357,. http://dx.doi.org/10.1016/j.applthermaleng.2019.114357.

Ravichandran, M., Su, G., Wang, C., Seong, J., Kossolapov, A., Phillips, B., Rahman, M., Bucci, M., 2021. Decrypting the boiling crisis through data-driven exploration of high-resolution infrared thermometry measurements. Appl. Phys. Lett. 118 (25), 253903,. http://dx.doi.org/10.1063/5.0048391.

Rokoni, A., Zhang, L., Soori, T., Hu, H., Wu, T., Sun, Y., 2022. Learning new physical descriptors from reduced-order analysis of bubble dynamics in boiling heat transfer. Int. J. Heat Mass Transfer 186, 122501.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. Adv. Neural Inf. Process. Syst. 29.

Sinha, K., Kumar, V., Kumar, N., Thakur, A., Raj, R., 2021. Deep learning the sound of boiling for advance prediction of boiling crisis. Cell Rep. Phys. Sci. 2 (3), 100382,. http://dx.doi.org/10.1016/j.xcrp.2021.100382.

Treder, M.S., Codrai, R., Tsvetanov, K.A., 2022. Quality assessment of anatomical MRI images from generative adversarial networks: human assessment and image quality metrics. J. Neurosci. Methods 374, 109579.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312, 135–153.

Wilson, A.G., Hu, Z., Salakhutdinov, R., Xing, E.P., 2016. Deep kernel learning. In: Artificial Intelligence and Statistics. PMLR, pp. 370–378.

Wynne, G., Duncan, A.B., 2022. A kernel two-sample test for functional data. J. Mach. Learn. Res. 23 (73), 1–51.

Xiang, S., Fu, Y., You, G., Liu, T., 2020. Unsupervised domain adaptation through synthesis for person re-identification. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1–6.

Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2849–2857.

You, S., 2014. Pool boiling. URL https://msht.utdallas.edu/.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017. Toward multimodal image-to-image translation. Adv. Neural Inf. Process. Syst. 30.

Zhu, Y., Zhuang, F., Wang, J., Chen, J., Shi, Z., Wu, W., He, Q., 2019. Multi-representation adaptation network for cross-domain image classification. Neural Netw. 119, 214–221.