Fully Unsupervised Domain-Agnostic Image Retrieval

Ziqiang Zheng#, Hao Ren#, *Member, IEEE*, Yang Wu, *Member, IEEE*, Weichuan Zhang, *Member, IEEE*, Hong Lu, *Member, IEEE*, Yang Yang*, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*—Recent research in cross-domain image retrieval has focused on addressing two challenging issues: handling domain variations in the data and dealing with the lack of sufficient training labels. However, these problems have often been studied separately, limiting the practicality and significance of the research outcomes. The existing cross-domain setting is also restricted to cases where domain labels are known during training, and all samples have semantic category information or instance correspondences. In this paper, we propose a novel approach to address a more general and practical problem: *fully unsupervised domain-agnostic image retrieval* under the domain-unknown setting, where no annotations are provided. Our approach tackles both the *domain variation* and *missing labels* challenges simultaneously. We introduce a new fully unsupervised One-Shot Synthesis-based Contrastive learning method (termed OSSCo) to project images from different data distributions into a shared feature space for similarity measurement. To handle the domain-unknown setting, we propose One-Shot unpaired image-to-image Translation (OST) between a randomly selected one-shot image and the rest of the training images. By minimizing the global distance between the original images and the generated images from OST, the model learns domain-agnostic representations. To address the label-unknown setting, we employ contrastive learning with a synthesis-based transform module from the OST training. This allows for effective representation learning without any annotations or external constraints. We evaluate our proposed method on diverse datasets, and the results demonstrate its effectiveness. Notably, our approach achieves comparable performance to current state-of-the-art supervised methods.**

***Index Terms*—One-shot image translation, Unsupervised learning, Image retrieval, Domain adaptation.**

I. Introduction

IMAGE retrieval has been a subject of extensive research for several decades [1]–[6]. The Content-Based Image Retrieval (CBIR) [7]–[9] aims to semantically match a query image with a large image gallery based on its visual content [2], and its applications span a wide range of domains, including visual place recognition [10], sketch-based image retrieval [11]–[14], person re-identification [15], and more [16], [17].

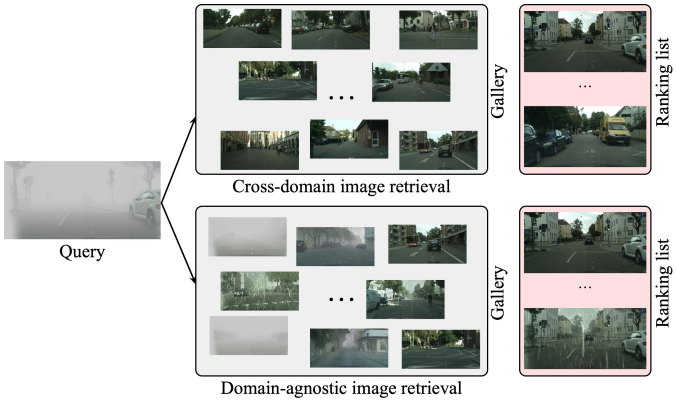


Fig. 1. Comparison of existing *cross-domain image retrieval* task and proposed *domain-agnostic image retrieval* task. In the cross-domain task (top), the goal is to match query images from a known source domain with gallery images from a different known target domain. Conversely, the domain-agnostic task (bottom) involves matching query and gallery images from an unspecified number of unknown domains. No domain-level or image-level labels are provided, requiring the model to learn domain-agnostic representations for handling diverse real data variations.

Recently, there has been growing attention given to a specific and practical case of image retrieval known as *cross-domain image retrieval*. This task involves matching query and gallery images from different domains, such as those affected by variations in illumination [18], weather conditions [19], or artistic style [20]. Cross-domain image retrieval is of paramount importance due to the widespread existence of domain variations in real-world data across various applications, including web data, which often spans open-ended domains.

The term "cross-domain" and its associated research works, such as [21], [22], have been primarily limited to scenarios where both the source and target domains are known beforehand. These studies typically involve a small number of domains, usually just two, and assume a large number of samples per domain. However, we have to acknowledge that these restrictions have limited the practical application and usability of the research outcomes.

Consider an image retrieval algorithm specifically trained for clear-to-foggy cross-domain place recognition. While it

Ziqiang Zheng, Yang Yang, and Heng Tao Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China. (e-mail: zhengziqiang1@gmail.com; dlyyang@gmail.com; shenhengtao@hotmail.com).

Hao Ren and Hong Lu are with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, 200438, China (email: {hren17, honglu}@fudan.edu.cn).

Yang Wu is with Tencent AI Lab, Shenzhen, 518100, China (email: dylan.yangwu@qq.com).

Weichuan Zhang is with the Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD 4222, Australia (email: zwc2003@163.com).

# Ziqiang Zheng and Hao Ren contributed equally to this research.

* Yang Yang is the corresponding author.

This work was partially supported by the National Natural Science Foundation of China under grant U20B2063, 62220106008 and 62306067.

may excel in that particular clear ↔ foggy domain shift, it becomes challenging to trust the same algorithm for deployment in an autonomous driving car, which might encounter various other weather conditions, such as rainy or low-light environments. Automatic extraction of domain information is often difficult in real-world scenarios, and relying on user input for domain labeling can be impractical and burdensome.

To address these limitations and develop a more versatile and practical approach, we propose a novel task named *domain-agnostic image retrieval*. This task does not require any domain information and makes no assumptions about the number of image domains available, as illustrated in Fig. 1. In addition to the domain variation issue, another long-standing challenge in image retrieval is the necessity for an ample supply of image-level labels for effective model training. The establishment of query-gallery correspondences is widely regarded as crucial for developing a powerful image retrieval model. However, acquiring such labels is often a time-consuming and expensive task, especially in unconstrained real-world scenarios. Moreover, the query images often stem from complex and diverse data distributions, making the absence of labels even more critical. For deep learning models that thrive on large volumes of labeled training data, the scarcity of labels presents a significant obstacle.

Recent efforts have been directed towards learning from weakly labeled or even completely unlabeled data, with some encouraging progress in the domain of contrastive learning [23], [24]. However, to the best of our knowledge, most advancements are still confined to within-domain cases. Additionally, many existing unsupervised learning models are not fully unsupervised, as they may still require some form of extra labeled data for other tasks, such as fine-tuning [25], [26] pre-trained models on other labeled data [27], or they rely on external priors or constraints [10]. Consequently, learning-based approaches for fully unsupervised image retrieval remain largely unexplored.

To address these challenges, this work aims to formulate a more demanding, practical, and valuable task: *fully unsupervised domain-agnostic image retrieval*. Under this scenario, both domain-level and image-level labels are entirely unavailable, necessitating the development of innovative and comprehensive solutions to address the domain variation and missing label issues in a unified framework. Fully unsupervised domain-agnostic image retrieval holds significant importance and potential value, surpassing both cross-domain image retrieval and unsupervised image retrieval. However, the technical challenges posed by this problem are notably more intricate. This work sets out to explore an initial yet promising solution for this practical image retrieval task, where we aim to address the label-unknown (fully unsupervised) and domain-unknown (domain-agnostic) issues simultaneously.

Existing unsupervised learning models, such as the state-of-the-art contrastive learning [28], typically rely on self-generated supervision during training. For example, in methods like SimCLR [23], predefined image data augmentation is used to create positive pairs (augmented views from the same image) and negative pairs (augmented samples from different images). This self-generated supervision is crucial for learning a feature space that measures feature similarity between samples. However, it is essential to note that such self-supervision is not inherently domain-agnostic. The transformations used to obtain augmented views in previous works [23], [29]–[31] are not designed to learn discriminative domain-agnostic feature representations. Consequently, the pursuit of domain-agnostic self-generated supervision becomes a necessary yet highly challenging problem.

The key to obtaining domain-agnostic representations without relying on domain-level and image-level annotations lies in our approach, where we consider each image as belonging to an inherent "image domain". Instead of assigning each image a unique pseudo instance label, we assign it a unique pseudo domain label. This enables unpaired image-to-image translation between constructed pseudo domain pairs, effectively covering real domain variations.

To create pseudo domain pairs for unpaired domain translation, we first construct a special "one-shot" image domain that contains only a single image randomly sampled from the entire training dataset. The remaining training images are then regarded as belonging to a "source" domain. This setup allows us to perform One-Shot unpaired image-to-image Translation (OST [32]–[34]) between the constructed "one-shot" image domain and the "source" domain. Consequently, we can project all the training samples from the "source" domain to the "one-shot" domain, synthesizing images with the appearance of the "target" domain. Importantly, the original images and their corresponding synthesized counterparts serve as augmented views from the same instance. This enables us to formulate domain-agnostic self-supervision without any human annotations.

The proposed approach, One-Shot Synthesis-based Contrastive learning (OSSCo), tackles a novel and practical fully unsupervised domain-agnostic image retrieval task. The formulated task surpasses the widely studied cross-domain image retrieval in terms of applicability and generality.

We summarize the key contributions of this work as follows:

- Introduction of a novel fully unsupervised domain-agnostic image retrieval task, which addresses a more challenging and versatile problem compared to cross-domain image retrieval.
- Development of an innovative approach to solve the proposed domain-agnostic image retrieval task under a fully unsupervised setting, where no additional data or priors/constraints are required. The annotation-free training procedure of OSSCo sheds light on the extraction of efficient domain-agnostic representations without any annotations.
- Comprehensive experiments conducted on diverse datasets to demonstrate the effectiveness of the proposed approach. OSSCo exhibits competitive performance even when compared to state-of-the-art supervised image retrieval algorithms. Ablation studies have also been included to analyze the impact of the proposed synthesis-based transform derived from OST.

The rest of this paper is organized as below. Section II provides a concise overview of related works in the field. In Section III, the proposed approach, One-Shot Synthesis-based Contrastive learning (OSSCo), is thoroughly explained

and detailed, including its iterative training framework design. Section IV presents extensive experimental results conducted on various datasets to evaluate the effectiveness of the proposed approach. The limitations of the proposed approach and its practical scenarios are discussed in Section V. The paper concludes in Section VI, summarizing the key findings and contributions of this work.

II. Related Work

A. Cross-domain Image Retrieval

Cross-domain image retrieval methods [1], [17], [35] primarily focus on extracting features from images belonging to clearly classified domains, along with extensive query-gallery correspondences. These methods often involve handcrafted descriptors [36]–[38] or learnable feature representations [35], [39]. Handcrafted features, such as SIFT [36], BoW [38], and SURF [37], are commonly used for cross-domain image matching. On the other hand, deep learning methods [35], [39] have garnered significant attention for their more robust and accurate recognition performance. These learnable feature representations are typically extracted under supervised settings, relying on extensive labeled data [40], [41]. However, the collection and organization of such labeled data can be costly and time-consuming [42].

In contrast to existing cross-domain image retrieval methods, our work focuses on fully unsupervised domain-agnostic image retrieval under domain-unknown and label-unknown settings. The proposed domain-agnostic image retrieval task aligns with real-world scenarios, significantly reducing the burden of annotation efforts.

B. One-shot Unpaired Image Translation

One-Shot unpaired image-to-image Translation (OST) aims to learn the mapping function between two domains in which one domain only includes one sample. There are mainly two settings for this task: "one-to-many" and "many-to-one". For the former setting, Benaim *et al*. [32] aim to generate an analogous output from a single one-shot image in the source domain $\mathbb{X}$. The synthesized image is required to be similar to the redundant samples from the target domain $\mathbb{Y}$. In contrast, the latter "many-to-one" targets to convert the diverse source samples to the target domain with a single one-shot sample [33], [34], which is a more challenging task due to the insufficient target data. Besides, the data imbalance between the abundant source data and the one-shot target sample results in an overfitting problem during the learning process. To alleviate the overfitting problem, [33], [34] propose to mine the style appearance representation from the limited one-shot sample through multiple parallel threads [34] or an adversarial style mining module [33]. In this paper, we aim to utilize the intrinsic property of OST to simulate the appearance variance among all the training samples.

Different from these previous works [33], [34], which only perform OST and evaluate the translation performance with some objective evaluation metrics, the proposed method adopts OST as one effective data augmentation and mainly focuses on the downstream image retrieval task.

C. Contrastive Learning

In contrast to traditional supervised learning methods [43] that require data-label correspondences, contrastive learning aims to learn efficient global feature representations with self-generated supervision. This annotation-free training scheme makes contrastive learning methods easy to implement and alleviates the burden of heavy annotation costs. Notably, recent works [23], [29]–[31], [44], [45] have achieved remarkable success and introduced a novel paradigm for visual recognition.

These methods aim to extract effective feature representations through self-generated supervision [23], [29], [46] or pretext tasks [31], [44]. For instance, SimCLR [23] combines various traditional transformations for data augmentations, where the self-supervision minimizes the global representation distance among the transformed outputs and maximizes the distance between different instance samples. MoCo-v2 [31] incorporates a projection head and transformations from SimCLR [29] to boost recognition performance. SwAV [47] leverages intermediate prototypes to reduce computational costs.

However, current contrastive learning methods are limited to within-domain recognition tasks, as the domain gap between different domains poses a significant challenge for effectively extracting features without any annotation. To address this, advanced data augmentations [48], [49] such as MixUp [50] and AugMix [51] could be introduced to contrastive learning models. MixUp combines two visual images to augment the data, enhancing robustness to adversarial samples and generalization to unseen data. AugMix augments the same image multiple times using different methods and combines them to enhance model robustness. iMix [52] integrates MixUp with existing contrastive learning models and introduces a regularization strategy to improve contrastive representation learning.

In this work, we take a different approach by incorporating a one-shot image synthesis module to disentangle domain-specific and domain-agnostic representations. This allows our method to address the challenge of domain-agnostic image retrieval without relying on within-domain supervision.

III. Proposed Approach

A. Preliminary

To ensure clarity and avoid potential misunderstandings, we provide detailed explanations of the technical terms and the experimental setting used in our approach. An *image domain* refers to a domain where there is agreement and common semantic attributes expressed (*e.g.*, foggy, daytime, nighttime, sketch, photo, etc.). Images from the same image domain share similar feature representations. *Image labels* refer to semantic category labels and image correspondences among all the training images for the image retrieval task. The proposed fully unsupervised domain-agnostic image retrieval setting is both *domain-unknown* and *label-unknown*. Specifically, during the entire training process, domain labels and image-level annotations (such as category labels or image correspondences) are not available. It should be noted that while the domain is unknown, it does not imply that there are no definable

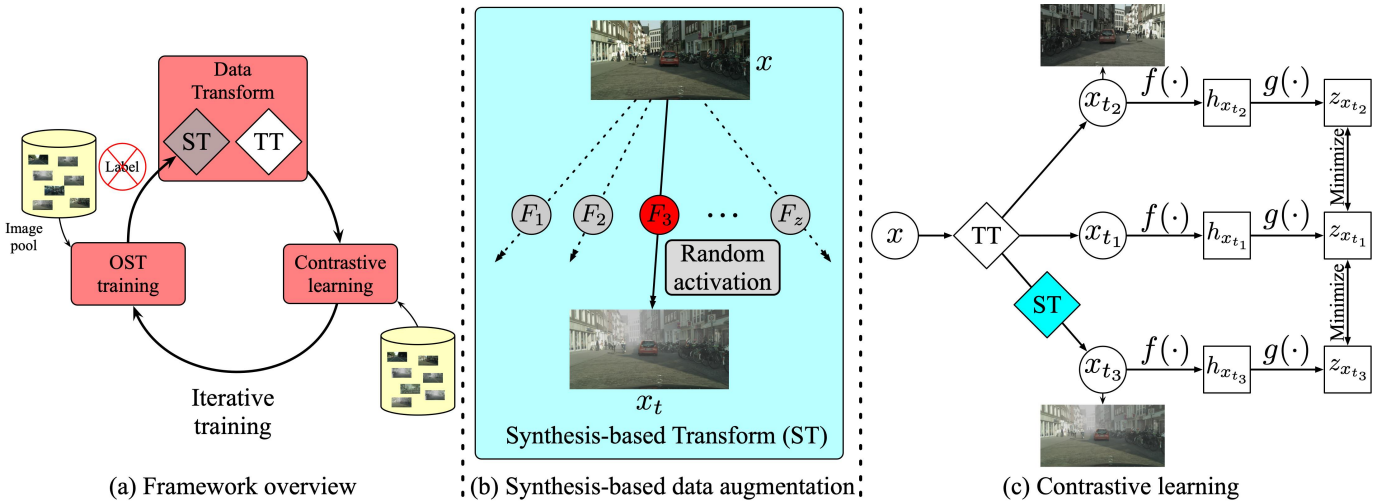


Fig. 2. The framework overview of our proposed OSSCo is depicted in (a). It is optimized through iterative training, which involves multiple iterations to approach the implicit real data variations. For more in-depth training details on the iterative process, please refer to Algorithm 1. (b) demonstrates the Synthesis-based Transform operation (ST). In this operation, we employ a random activation strategy to synthesize transformed images. This strategy enhances the model's robustness and flexibility in capturing domain-agnostic representations. (c) illustrates the training scheme of the contrastive learning used in our method. Contrastive learning is a key component of OSSCo, enabling the model to learn effective domain-agnostic representations. Through this process, the model minimizes the distances between positive pairs (images from the same instance) and maximizes the distances between negative pairs (images from different instances).

implicit domain labels. Ground-truth domain labels may exist, but we do not use them for training; they are only utilized for performance evaluation to ensure fairness when compared with other methods that require domain labels. Additionally, image labels are also not available for training under the unsupervised setting.

We provide detailed discussions regarding two concurrent lines of work: *domain adaptation* and *domain generalization*. *Domain adaptation* methods [53], [54] aim to bridge the shared knowledge between two visual domains, requiring both the source and target distributions to be available and fixed during the training procedure. In other words, the domain labels for training samples are available. The goal of domain adaptation is to extract domain-agnostic feature representations that achieve satisfactory recognition performance in the target domain. On the other hand, *domain generalization* refers to a setting where the testing domains did not appear or were not known in advance [55]. The training and testing samples are from different image domains with separate data distributions. Under both the domain adaptation and domain generalization settings, domain information and image-level label annotations are available during the training process. In contrast, our proposed method does not require any domain information or label information for training, making it more practical and generic.

The framework overview of our method is shown in Fig. 2 (a), which is optimized through the proposed iterative training strategy. First, the One-Shot unpaired image-to-image Translation (OST) training is conducted. For more training details, please refer to Section III-B. Then, OST is applied for the Synthesis-based Transform (ST) shown in Fig. 2 (b). We perform two categories of data transforms for the contrastive learning procedure in Fig. 2 (c): 1) the Traditional Transform (TT) operations (including horizontal flipping, resizing, cropping, color jittering, and grayscaling) and 2) the Synthesis-based Transform (ST).

B. Synthesis-based Transform

One-shot Unpaired Image-to-image Translation. We start by considering an *image pool* containing all the training images $\mathbb{X} : x \in \{x_i\}_1^M$ (total M samples) without any label annotations. Our goal is to generate *one-shot image sets* with z different instinctive images, as illustrated in Fig. 3. To achieve this, we split the image set $\mathbb{X}$ into $z+1$ subsets: $\{\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_z\}$ and $\mathbb{X}_s$, where $\mathbb{X}_s$ contains all the remaining training samples except the selected z images. Each subset $\mathbb{X}_i$ contains only a one-shot image. Due to the domain-unknown setting, the one-shot image may come from clearly definable domains (*e.g.*, foggy, rainy, or any other domain). The selected one-shot images carry separate style representations due to the random selection strategy. A large value of z ensures that images from each implicit domain can be sampled. By using these self-formulated subsets $\mathbb{X}_s$ and $\{\mathbb{X}_i\}_1^z$, we perform OST between the subset $\mathbb{X}_s$ and each $\mathbb{X}_i$.

To perform unpaired image translation between the source subset $\mathbb{X}_s$ and the target subsets $\{\mathbb{X}_i\}_1^z$, we design z tiny symmetric reverse generator pairs: $\{F_i, G_i\}_1^z$. Each generator pair (F_i, G_i) is responsible for the domain translation between $\mathbb{X}_s$ and $\mathbb{X}_i$. We adopt the vanilla CycleGAN architecture [56] for OST, as it is relatively easier to train and faster to converge than other architectures like StarGAN [57]. We use the widely adopted adversarial loss $\mathcal{L}_{adv}^i(F_i, D_i)$ from $\mathbb{X}_s$ to $\mathbb{X}_i$ to generate desired images:

$$\begin{aligned}\mathcal{L}_{adv}^i(F_i, D_i) = \mathbb{E}_{\mathbb{X}}[\log D_i(x_i^{ost})] \\ + \mathbb{E}_{\mathbb{X}}[\log(1 - D_i(F_i(x_s)))],\end{aligned} \tag{1}$$

where x_s is a randomly selected sample from $\mathbb{X}_s$ and x_i^{ost} is the one-shot image from $\mathbb{X}_i$. Each subset contains only one

one-shot image x_i^{ost}. D_i is the corresponding discriminator for the translator F_i. We focus on the forward "many-to-one" direction (from $\mathbb{X}_s$ to $\{\mathbb{X}_i\}_1^z$) under the OST setting and do not compute the reverse adversarial loss $\mathcal{L}_{adv}^i(G_i)$ for the reverse translator G_i. To preserve the content information of images from $\mathbb{X}_s$, we use the cycle-consistency loss $\mathcal{L}_{cyc}^i(F_i, G_i)$ [56] to link F_i and G_i:

$$\begin{aligned}\mathcal{L}_{cyc}^i(F_i, G_i) = \mathbb{E}_{\mathbb{X}}[\|G_i(F_i(x_s)) - x_s\|_1] \\ + \mathbb{E}_{\mathbb{X}}[\|F_i(G_i(x_i^{ost})) - x_i^{ost}\|_1],\end{aligned} \tag{2}$$

where the pixel-wise loss between the reconstruction and the original image helps preserve the content information after the one-shot unpaired image-to-image translation. It is worth mentioning that G_i targets to synthesize the counterpart reconstruction of x_s from the formulated source set $\mathbb{X}_s$ containing samples from several implicit, clearly definable domains. The specific implicit domain from which x_s comes does not matter. The final objective function for OST training is:

$$\mathcal{L}_{ost} = \sum_{i=1}^{z}(\mathcal{L}_{adv}^i(F_i, D_i) + \lambda \mathcal{L}_{cyc}^i(F_i, G_i)), \tag{3}$$

where λ is a hyper-parameter to balance the adversarial loss and the cycle-consistency loss, and we set $\lambda = 10$ following vanilla CycleGAN [56].

Synthesis-based Data Augmentation. After executing the OST training for m iterations, we suspend the OST training, and the parameters θ_F of $\{F_i\}_1^z$ are frozen. We utilize $\{F_i\}_1^z$ for the synthesis-based transform to generate images with different one-shot target appearances. Since $\{F_i\}_1^z$ can generate z different outputs simultaneously, to reduce computational costs and promote image diversity, we adopt an important *random activation* strategy, as shown in Fig. 2 (b). This strategy generates only one output x_t, which shares the same instance as x. The augmented output x_t is used for later contrastive learning to extract domain-agnostic feature representations. The random activation strategy, widely adopted in ensemble learning [58], effectively reduces variance. By using random activation, we can significantly reduce the inference time and nearly explore all domain differences. The synthesized sample diversity is promoted by a large z and random activation. Moreover, the random activation strategy can also mitigate the influence of noise and uncertainty caused by OST models, as there may be visible visual artifacts in the synthesized images. Additionally, the target goal of our synthesis-based data augmentation is to quickly learn the appearance representation of the self-formulated subsets $\{\mathbb{X}_i\}_1^z$ and achieve reasonable generation for further domain-agnostic feature extraction. Other image synthesis approaches (*e.g.*, StarGAN [57], neural-style [59], fast-neural-style [60]) could also be adopted, but in this paper, we focus on elaborating the use of vanilla CycleGAN for convenience and clarity of explanation.

C. Contrastive Learning

To obtain domain-agnostic feature representations for images from arbitrary domains, we employ both Traditional Transform operations (TT) and the Synthesis-based Transform (ST). The traditional transforms include random resizing, cropping, color jittering, flipping, and grayscaling. The TT operation generates two augmented views, x_{t_1} and x_{t_2}, while the combination of TT and ST generates another augmented view x_{t_3}, as illustrated in Fig. 2 (c). The combination of ST and TT could provide the model with a stronger ability to extract the domain-agnostic feature representations while preserving the original ability of contrastive learning models.

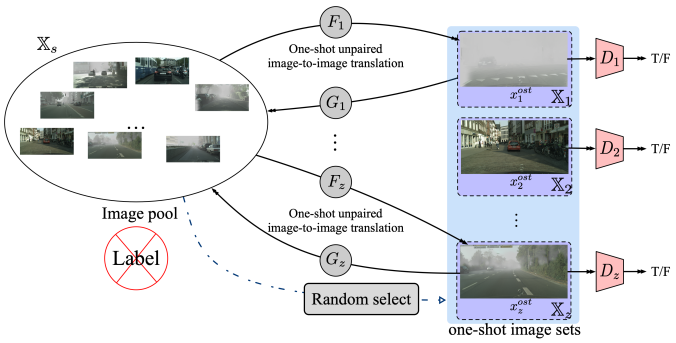


Fig. 3. Training details of the One-Shot unpaired image-to-image Translation (OST). We split all the training samples $\mathbb{X}$ into $z+1$ subsets: $\{\mathbb{X}_i\}_1^z$, each containing only one corresponding one-shot image x_i^{ost}, and $\mathbb{X}_s$ containing all the remaining training samples. During OST training, the CycleGAN architecture [56] is utilized to learn the mapping functions F_i and G_i for each one-shot image domain $\mathbb{X}_i$. The generator F_i aims to translate images from the source domain $\mathbb{X}_s$ to the one-shot image domain $\mathbb{X}_i$, while the generator G_i aims to translate images from $\mathbb{X}_i$ back to $\mathbb{X}_s$. The discriminators D_i are used to distinguish between real images from the one-shot domain $\mathbb{X}_i$ and generated images from $F_i(\mathbb{X}_s)$. The training process iterates over multiple epochs to optimize the generators and discriminators for each one-shot image domain. After the OST training is completed, the frozen generators $\{F_i\}_1^z$ are utilized for synthesis-based data augmentation and downstream contrastive learning for domain-agnostic feature extraction.

For contrastive learning, we utilize an encoder $f(\cdot)$ (*e.g.*, ResNet [43]) to extract global features $h_{x_{t_1}} = f(x_{t_1}) = ResNet(x_{t_1})$, where $h_{x_{t_1}} \in \mathbb{R}^d$ is the feature representation after the average pooling layer. A small neural network projection head $g(\cdot)$ is appended to map the global features to the output space where the loss is computed. The projection head $g(\cdot)$ consists of a single hidden layer, and it produces $z_{x_{t_1}} = g(h_{x_{t_1}}) = W^{(2)}\sigma(W^{(1)}h_{x_{t_1}})$, where σ is a ReLU non-linearity function.

During training, we sample a mini-batch of N examples and define the contrastive prediction task on pairs of augmented examples derived from the mini-batch, resulting in $3N$ data points. We do not explicitly sample negative examples. Instead, given a positive pair, we treat the other $3(N-1)$ augmented examples within the mini-batch as negative examples. The similarity between feature u and v is denoted as $sim(u, v) = u^T v/\|u\|\|v\|$. We compute two types of losses. First, the loss for the positive example pair (x_{t_1}, x_{t_2}) is given by:

$$\mathcal{L}(x_{t_1}, x_{t_2}) = -\log \frac{\exp(\frac{sim(z_{x_{t_1}}, z_{x_{t_2}})}{\tau})}{\sum_{x \in X, x \neq x_{t_1}} \exp(\frac{sim(z_{x_{t_1}}, z_x)}{\tau})}, \tag{4}$$

where X is the set of augmented images in the current mini-batch, and τ is a temperature parameter. Second, the loss for

examples (x_{t_1}, x_{t_3}) is given by:

$$\mathcal{L}(x_{t_1}, x_{t_3}) = -\log \frac{\exp(\frac{sim(z_{x_{t_1}}, z_{x_{t_3}})}{\tau})}{\sum_{x \in X, x \neq x_{t_1}} \exp(\frac{sim(z_{x_{t_1}}, z_x)}{\tau})}. \quad (5)$$

The final objective function for contrastive learning is defined as:

$$\mathcal{L}_{con} = \mathcal{L}(x_{t_1}, x_{t_2}) + \mathcal{L}(x_{t_1}, x_{t_3}), \quad (6)$$

which is used to optimize the contrastive learning model and extract domain-agnostic feature representations.

D. Training Pipeline

We propose an advanced iterative training approach, as shown in Algorithm 1, to combine the One-Shot unpaired image-to-image Translation (OST) training and contrastive learning. The training pipeline consists of three stages:

- **OST Training**: We start by executing the OST training for m iterations with a batch size of 1. During this stage, we split all the training samples $\mathbb{X}$ into $z+1$ subsets, where $\{\mathbb{X}_i\}_1^z$ each contains one corresponding one-shot image x_i^{ost}, and $\mathbb{X}_s$ contains all the remaining training samples. The OST training aims to learn the translation between the source subset $\mathbb{X}_s$ and each target subset $\mathbb{X}_i$, using the CycleGAN architecture [56]. After optimization, the OST model is suspended, and the parameters θ_F of the translators $\{F_i\}_1^z$ are frozen.
- **Contrastive Learning**: The frozen $\{F_i\}_1^z$ are then used for image synthesis in the contrastive learning stage, which is executed for n iterations with a batch size of b. During this stage, both Traditional Transform operations (TT) and the Synthesis-based Transform (ST) are applied to generate augmented views x_{t_1}, x_{t_2}, and x_{t_3} as illustrated in Fig. 2 (c). These augmented views are used for contrastive learning to extract domain-agnostic feature representations. The contrastive learning model is trained to optimize the objective function $\mathcal{L}_{con}$ defined in Eq. (6).
- **One-shot Image Set Refresh**: Considering that the one-shot image pool only contains z different images, which may not cover all the styles of the entire image pool, we design a one-shot image set refresh operation. After completing the contrastive learning training, we randomly pick another z images from the training set to replace the existing images in the one-shot image pool. This refresh operation helps to simulate different styles and enhance the diversity of the one-shot image pool.

We repeat these three training stages for k rounds, resuming the parameters of the entire network from the previous round. This iterative training strategy enables the model to gradually disentangle domain-specific and domain-agnostic representations and approach the implicit real data variations effectively.

Algorithm 1 Iterative Training for OSSCo

Require: k (total number of rounds), m (number of iterations for OST training in each round), n (number of iterations for contrastive learning in each round), b (batch size for contrastive learning), and z (number of samples in one-shot image set).
Ensure: Model parameters $\theta = \{\theta_F, \theta_G, \theta_D, \theta_f, \theta_g\}$.

```
for i = 1, 2, ..., k do
    Randomly initialize one-shot image pool x^ost ∈ {x_1^ost, x_2^ost, ..., x_z^ost} from 𝕏,
    for j = 1, 2, ..., m do
        Randomly sample x_s from 𝕏_s,
        Optimize the parameters θ_F, θ_G, and θ_D with L_ost,
    end for
    Freeze θ_F, θ_G, and θ_D,
    for j = 1, 2, ..., n do
        Randomly sample {x_1, x_2 ..., x_b} from 𝕏,
        Optimize θ_f and θ_g based on TT and ST with L_con,
    end for
    Unfreeze θ_F, θ_G, and θ_D,
    Perform one-shot image set refresh: randomly replace x^ost with z images from 𝕏 and obtain new 𝕏_s.
end for
```

IV. Experiments

A. Datasets

The experiments are conducted on four datasets: Foggy Cityscapes [19], CUFSF [20], Facades [56], and Weather Cityscapes [61].

Foggy Cityscapes [19] is a synthetic foggy dataset with three different levels of visibility: 150m, 300m, and 600m. For this dataset, we choose clear images from the Cityscapes dataset [62] and foggy images with 150m visibility following the original split. The training set consists of 2,975 clear images and their corresponding foggy versions, while the evaluation set contains 500 clear-foggy image pairs.

CUFSF (CUHK Face Sketch FERET Database [20]) is a face-sketch synthesis and recognition dataset, which includes 1,194 persons from the FERET database [65]. Each person is represented by a face photo and a corresponding sketch image with shape exaggeration drawn by an artist. We use the train/test split from DualGAN [66] and select 995 photo-sketch image pairs for training, leaving the rest for evaluation.

Facades dataset [56] consists of image pairs, each composed of a semantic label map and a photo of the same building. We follow the same train/test split as CycleGAN [56], using 400 pairs for training and 106 pairs for testing.

Weather Cityscapes [61] is a dataset generated from physics-based rendering. Each image provides 8 rain levels (ranging from drizzle to storm conditions) from the *Rainy Cityscapes* dataset, and 7 fog intensities (ranging from clear to dense fog) from the *Foggy Cityscapes* dataset. The rain levels are denoted by the amount of rain flow in millimeters (1mm, 5mm, 17mm, 25mm, 50mm, 75mm, 100mm, and 200mm), while the fog intensities are represented by visibility distances (30m, 40m, 50m, 75m, 150m, 375m, and 750m).

These datasets are carefully chosen to evaluate the effectiveness and generalizability of our proposed approach on different tasks and challenging scenarios.

TABLE I
DOMAIN-AGNOSTIC IMAGE RETRIEVAL PERFORMANCE ON FOGGY CITYSCAPES [19] DATASET. THE SCORES IN BOLD INDICATE THE BEST ACCURACY. THE METHODS WITH SUPERSCRIPT * INDICATE SUPERVISED METHODS, WHILE METHODS WITHOUT SUPERSCRIPT * ARE OPTIMIZED THROUGH UNSUPERVISED TRAINING.

Method	Clear → Foggy				Foggy → Clear				Clear ↔ Foggy			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Pretrained [43]	77.0	82.0	86.6	89.2	93.4	95.6	97.0	98.4	45.7	53.3	59.3	65.4
NPID [44]	22.8	29.4	37.2	46.4	21.6	28.4	35.6	43.6	5.9	8.3	11.2	14.1
NPID+Pretrained	75.0	84.4	90.8	94.6	77.4	85.2	90.8	94.2	40.1	49.2	59.6	70.0
SimCLR [23]	92.2	94.6	96.6	97.8	89.6	93.0	95.4	98.2	80.1	85.4	88.8	92.3
SimCLR+Pretrained	93.2	97.2	98.6	99.0	97.2	99.2	99.6	99.6	82.9	89.0	93.8	97.3
SimCLR+MixUp [50]	94.4	98.4	98.8	99.4	97.8	99.6	99.6	99.6	83.4	89.7	94.4	97.8
SimCLR+AugMix [51]	95.4	97.8	98.8	99.2	97.8	99.6	99.6	99.8	84.3	89.9	95.2	98.3
SimCLR+iMix [52]	95.0	97.6	98.6	99.4	97.4	99.6	99.6	99.6	83.6	89.9	94.6	98.1
OSSCo	**99.2**	**99.6**	**99.8**	**99.8**	**99.2**	**99.4**	**99.4**	**99.6**	**96.9**	**98.9**	**99.4**	**99.5**
SoftTriple* [63]	99.6	99.8	100	100	99.8	99.8	99.8	100	98.4	99.7	99.8	99.9
ProxyAnchor* [64]	99.6	100	100	100	99.6	99.8	99.8	100	98.8	99.6	99.6	99.8

B. Evaluation Metric and Experimental Setup

Evaluation Metric: The main evaluation metric used in our experiments is Recall@K [67]. At the inference stage, we compute the cosine similarities of feature vectors between the query image and other images in the test set. We then retrieve the top K images with the highest similarities. A higher Recall@K score indicates better retrieval performance.

Experimental Setup: We begin by using a pre-trained ResNet [43] model on ImageNet [27] to obtain feature vectors after the global average pooling layer. We use these feature vectors to compute the retrieval precision and refer to this setting as *Pretrained*. Next, we experiment with two contrastive learning methods, NPID [44] and SimCLR [23], under two settings: 1) random weight initialization and 2) initializing the weights with the pre-trained ImageNet model [43]. We denote these settings as *NPID+Pretrained* and *SimCLR+Pretrained*, respectively. Additionally, we combine MixUp [50], AugMix [51], and iMix [52] with SimCLR to explore the superiority of our proposed OSSCo over existing data augmentation algorithms.

For fully unsupervised learning, we train our OSSCo from scratch without any prior knowledge. We also include two currently supervised image retrieval methods, SoftTriple [63] and ProxyAnchor [64], for comparison. At the evaluation stage, we use domain information and label annotations to evaluate the domain-agnostic image retrieval performance. We design two settings for evaluation: 1) the query image is from domain $\mathbb{X}$ while the gallery images are from domain $\mathbb{Y}$ (denoted as $\mathbb{X} \rightarrow \mathbb{Y}$); 2) the query and gallery images are a mixture of images from domain $\mathbb{X}$ and domain $\mathbb{Y}$ (denoted as $\mathbb{X} \leftrightarrow \mathbb{Y}$).

C. Implementation Details

For the OST training, we adopt the architecture from CycleGAN [56]. Both $\{F_i\}_1^z$ and $\{G_i\}_1^z$ consist of three *Conv-InstanceNorm-ReLU* blocks to downsample the input images, with a kernel size of 3 and a stride of 2. The bottleneck consists of 9 residual blocks to preserve content information. Three *Deconv-InstanceNorm-ReLU* blocks are used to generate images of the same size as the input, and a Tanh activation function is applied to obtain the normalized outputs. The discriminator architecture follows the PatchGAN [68] design. To reduce computational cost and memory burden, the channel number of the first layer is set to 32 for $\{F_i\}_1^z$, $\{G_i\}_1^z$, and $\{D_i\}_1^z$.

At the second contrastive learning stage, we use the same architecture as SimCLR [23] as the backbone. The images are firstly resized to 256×256, and then applied with color jittering, after that, the grayscaling and horizontal flipping are used to augment the images. The projected output z_x is a 128-dimensional vector, and the temperature τ, inherited from [23], is set to 0.1. To preserve the content as much as possible, we change the random resize scale from (0.2, 1.0) used in SimCLR to (1.0, 1.12), as used in CycleGAN [56]. The number of negative samples in NPID [44] is set to 4,096, following the default setting. The momentum in NPID [44] is set to 0.5. All methods are optimized for the same number of iterations: 10,000. We use the Adam optimizer [69] with a learning rate of $1\mathrm{e}{-3}$ and weight decay of $1\mathrm{e}{-6}$ for the optimization. We set $k = 5$, $m = 4,000$, $n = 10,000$, and $b = 16$ in all our experiments through grid search. To ensure a fair comparison, the batch size of SimCLR is set to 32, while that of NPID and all supervised methods is 64. Our code is available on https://github.com/leftthomas/OSSCo, implemented with PyTorch [70] library.

D. Comparison with State-of-the-art

Retrieval-based Place Recognition. The Foggy Cityscapes dataset [19] contains images with different appearances due to dense fog, leading to a loss of content information in buildings and trees. This poses a challenge for unsupervised retrieval methods to extract efficient representations for robust image matching. The quantitative comparison on the Foggy Cityscapes dataset is illustrated in Table I. As shown, the proposed OSSCo outperforms other unsupervised methods by a large margin. Even when compared with the supervised method ProxyAnchor [64], the proposed OSSCo remains competitive. Notably, initializing the weights with a pre-trained model on the ImageNet dataset [27] significantly boosts the image retrieval performance of NPID [44], as

Fig. 4. The translated outputs of our OST model by choosing different one-shot images. The input images and one-shot images consist of original raw images from the Cityscapes [62] dataset and simulated foggy images from the Foggy Cityscapes [19] dataset. The images in the yellow boxes indicate the images generated by our OST model.

TABLE II
DOMAIN-AGNOSTIC IMAGE RETRIEVAL PERFORMANCE ON CUFSF [20] DATASET. THE SCORES IN BOLD INDICATE THE BEST ACCURACY. THE METHODS WITH SUPERSCRIPT * INDICATE SUPERVISED METHODS, WHILE METHODS WITHOUT SUPERSCRIPT * ARE OPTIMIZED THROUGH UNSUPERVISED TRAINING.

Method	Sketch → Photo				Photo → Sketch				Sketch ↔ Photo			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Pretrained [43]	9.0	13.1	18.1	25.6	16.6	24.1	30.7	38.2	0.3	0.3	1.3	3.0
NPID [44]	37.2	48.7	63.8	73.4	40.7	52.8	67.8	73.9	27.1	34.4	46.0	60.1
NPID+Pretrained	41.2	49.8	59.8	69.4	42.2	50.8	62.3	72.4	26.4	34.2	46.0	56.0
SimCLR [23]	24.1	39.2	56.3	72.4	32.7	45.2	56.3	68.8	15.1	21.9	33.7	49.0
SimCLR+Pretrained	35.2	42.7	52.8	64.8	34.2	45.2	55.6	62.3	21.4	31.7	37.9	46.2
SimCLR+MixUp [50]	37.0	45.2	55.6	67.4	35.0	46.2	57.0	64.3	23.0	33.5	39.1	49.1
SimCLR+AugMix [51]	39.2	48.7	59.2	68.6	37.1	49.2	59.4	67.1	23.6	35.2	39.2	51.2
SimCLR+iMix [52]	43.1	52.1	62.3	71.2	41.7	53.4	62.7	71.1	27.4	38.5	43.1	51.2
OSSCo	**82.4**	**93.5**	**97.5**	**99.5**	**88.4**	**98.0**	**99.5**	**99.5**	**55.0**	**70.4**	**87.2**	**94.5**
SoftTriple* [63]	86.4	92.5	95.5	99.0	89.4	93.5	97.5	99.5	79.6	85.9	92.7	96.2
ProxyAnchor* [64]	95.5	98.0	100	100	95.5	98.5	100	100	91.7	95.7	98.2	99.7

seen in "NPID+Pretrained" method. Additionally, we observe that when combined with MixUp [50], AugMix [51], or iMix [52], the SimCLR method shows some improvement in performance. However, there is still a significant gap between SimCLR and our proposed OSSCo method, indicating the superiority of our approach for domain-agnostic image retrieval.

Furthermore, we showcase the intermediate outputs of the OST model in Fig. 4. The OST model produces images with a similar style to the one-shot image, with the original domain appearance faded. Some local content information of the original inputs is lost after the image translation. However, despite these visual artifacts, our OSSCo method achieves accurate recognition. This is because contrastive learning models tend to focus on global content representation rather than local content information. In other words, the contrastive learning model does not require the generated images to have high image synthesis quality.

Sketch-photo Image Retrieval. Matching photo images with their corresponding person sketches holds great potential in facial recognition systems. However, due to the lack of colors, textures, and structural information, sketch images are highly abstract, making sketch-photo image retrieval a challenging task. In our experiment on the CUFSF dataset [20], where photos captured by cameras and sketch images drawn by artists are from different domains, we perform the proposed domain-agnostic image retrieval task. The quantitative comparison is shown in Table II.

The results demonstrate that the contrastive learning-based methods struggle to extract feature representations that are robust to color, texture, and appearance changes, leading to poor sketch-photo retrieval performance. In contrast, our method, which does not require any annotations, surpasses all the contrastive learning-based methods by a significant margin (our R@1 score is nearly twice as high as that of NPID [44] under all settings). Moreover, our OSSCo achieves comparable performance to the fully supervised method SoftTriple [63] (88.4% vs 89.4% for R@1 under the photo → sketch setting). These results highlight the effectiveness of our approach in addressing the challenges posed by sketch-photo image retrieval.

Facades-label Image Retrieval. In addition to sketch-photo image retrieval, we conduct experiments on the Facades dataset [56], where the semantic label images contain highly abstracted information from the original facades images. The quantitative comparison among different methods is presented in Table III. As shown, the pre-trained model on the ImageNet

TABLE III
DOMAIN-AGNOSTIC IMAGE RETRIEVAL PERFORMANCE ON FACADES [56] DATASET. THE SCORES IN BOLD INDICATE THE BEST ACCURACY. THE METHODS WITH SUPERSCRIPT * INDICATE SUPERVISED METHODS, WHILE METHODS WITHOUT SUPERSCRIPT * ARE OPTIMIZED THROUGH UNSUPERVISED TRAINING.

Method	Facades → Label				Label → Facades				Facades ↔ Label			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Pretrained [43]	5.7	6.6	8.5	17.0	3.8	4.7	7.5	12.3	0	0	0	0
NPID+Pretrained	3.8	4.7	8.5	11.3	4.7	4.7	9.4	15.1	3.3	3.8	4.7	7.5
SimCLR+Pretrained	36.8	45.3	53.8	65.1	37.7	48.1	59.4	67.9	25.5	33.5	45.3	50.5
SimCLR+MixUp [50]	43.2	51.1	58.2	69.3	42.1	50.6	64.1	68.7	28.3	39.1	49.1	55.1
SimCLR+AugMix [51]	45.7	53.1	65.2	68.3	43.4	51.7	65.7	69.3	29.1	40.5	50.3	57.5
SimCLR+iMix [52]	46.3	55.3	66.7	70.1	45.1	52.6	67.1	70.4	29.6	41.6	51.7	59.7
OSSCo	**77.4**	**83.0**	**86.8**	**91.5**	**81.1**	**85.8**	**89.6**	**92.5**	**61.8**	**70.8**	**78.3**	**84.0**
SoftTriple* [63]	98.1	98.1	100	100	96.2	98.1	99.1	100	95.3	96.2	98.6	99.1
ProxyAnchor* [64]	99.1	100	100	100	99.1	99.1	99.1	100	97.2	98.1	99.1	99.5

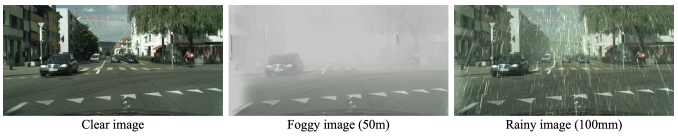


Fig. 5. An illustration of the utilized clear images, foggy images with 50m visibility, and rainy images with 100mm rainfall.

dataset performs poorly in this facades ↔ label image retrieval task, as the learned feature representations are not effectively suited for this specific domain. Similarly, SimCLR and NPID struggle to achieve effective retrieval results due to the difficulty in capturing distinct content representations. In contrast, our method excels in this task, demonstrating remarkable image retrieval performance without any image labels. This outcome further underscores the strength of our approach in addressing challenges in facades-label image retrieval.

Domain-agnostic Image Retrieval Among Multiple Domains. In this task, we perform domain-agnostic image retrieval among multiple domains, specifically three domains: foggy images with 50m visibility, rainy images with 100mm rainfall, and their corresponding clear images. The presence of heavy fog in the foggy images from the Weather Cityscapes dataset [61] makes it extremely challenging to identify whether two images are from the same location. Additionally, the rainy images and foggy images exhibit significant appearance variance, requiring the model to disentangle domain-specific and domain-agnostic representations. Fig. 5 shows some example images for intuitive comparison.

To evaluate the performance, we reorganize the Weather Cityscapes dataset for train/test split, selecting the images collected at "ulm", "weimar", and "zurich" for evaluation, and the remaining images for training. The dataset consists of a total of 8,148 images, with 2,716 images per domain for training and 777 images, with 259 images per domain, for testing.

For a fair comparison, all experiments are conducted under the same split. We report the results under the three most challenging settings: clear ↔ foggy, clear ↔ rainy, and foggy ↔ rainy in Table V. As shown, directly using feature representations from the pre-trained model [43] results in poor performance. Our OSSCo method achieves the highest scores among all unsupervised methods, outperforming others by a significant margin. Additionally, the retrieval performance under the clear ↔ rainy setting is much higher than under the clear ↔ foggy setting, likely due to the less severe occlusion present in the rainy images.

TABLE IV
QUANTITATIVE COMPARISON OF OUR OSSCO UNDER DIFFERENT COMBINATIONS OF m AND n ON THE CUFSF [20] DATASET FOR SKETCH ↔ PHOTO IMAGE RETRIEVAL. THE SCORES IN BOLD INDICATE THE BEST ACCURACY.

m	n	R@1	R@2	R@4	R@8
2,000	5,000	31.2	43.5	57.3	69.9
4,000	5,000	34.9	50.5	66.6	82.2
6,000	5,000	33.2	49.3	65.8	83.2
8,000	5,000	32.9	48.7	71.6	87.4
2,000	10,000	42.2	56.5	73.9	84.4
4,000	10,000	**55.0**	**70.4**	**87.2**	**94.5**
6,000	10,000	45.7	59.3	76.1	87.2
8,000	10,000	44.5	60.8	73.9	85.7

E. Ablation Studies

To demonstrate the effectiveness of the proposed components in our OSSCo, we perform ablation studies on the sketch-photo image retrieval task using the CUFSF dataset [20]. We focus on the sketch ↔ photo setting for this analysis.

Training Balance. To achieve a better balance between OST training and contrastive learning, we explore different combinations of m and n in Table IV. A small value of m may not allow the OST module to optimize effectively, leading to insufficient learning of domain variance through the synthesis-based transform module. According to our observations, a value of $m = 4,000$ is the best choice for the one-shot unpaired image-to-image translation. Similarly, a value of n is set to $10,000$ for the contrastive learning stage, ensuring the extraction of meaningful feature representations.

Selection of z. We also investigate the impact of different choices of z and present the quantitative comparison in Table VI. z represents the number of one-shot images used for the OST model. If z is too small, we may not be able to effectively simulate all domain appearances, leading

TABLE V
DOMAIN-AGNOSTIC IMAGE RETRIEVAL PERFORMANCE ON WEATHER CITYSCAPES [61] DATASET. THE SCORES IN BOLD INDICATE THE BEST ACCURACY. THE METHODS WITH SUPERSCRIPT * INDICATE SUPERVISED METHODS, WHILE METHODS WITHOUT SUPERSCRIPT * ARE OPTIMIZED THROUGH UNSUPERVISED TRAINING.

Method	Clear ↔ Foggy				Clear ↔ Rainy				Foggy ↔ Rainy			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Pretrained [43]	0.3	0.4	0.4	0.6	0.1	0.1	0.4	0.6	0	0	0	0.1
NPID+Pretrained	1.4	1.8	3.3	4.7	64.6	73.3	79.7	84.7	0.4	0.4	0.4	1.0
SimCLR+Pretrained	14.8	19.9	25.5	32.6	92.2	95.3	97.1	98.2	19.2	24.4	31.3	37.5
SimCLR+MixUp [50]	16.3	23.4	28.4	38.1	93.2	96.1	98.0	98.2	21.1	26.1	33.1	40.5
SimCLR+AugMix [51]	17.4	24.5	29.5	41.5	92.8	95.5	97.6	98.6	23.9	27.4	35.5	42.1
SimCLR+iMix [52]	18.1	26.7	30.7	42.7	93.3	95.9	97.9	99.1	24.1	30.3	36.8	44.1
OSSCo	**55.2**	**62.4**	**69.9**	**76.3**	**96.2**	**96.7**	**96.9**	**97.2**	**72.7**	**79.0**	**85.0**	**89.3**
SoftTriple* [63]	98.7	99.9	100	100	100	100	100	100	99.7	100	100	100
ProxyAnchor* [64]	95.3	97.8	98.3	98.5	98.3	98.6	98.9	99.0	98.2	99.3	99.6	99.9

TABLE VI
QUANTITATIVE COMPARISON OF OUR OSSCO UNDER DIFFERENT z ON THE CUFSF [20] DATASET FOR SKETCH ↔ PHOTO IMAGE RETRIEVAL. THE SCORES IN BOLD INDICATE THE BEST ACCURACY. WE ALSO REPORT THE PARAMETERS (M), MACS (G) AND OCCUPIED GPU MEMORIES (G).

z	Params	MACs	Mems	R@1	R@2	R@4	R@8
2	14.2	61.3	1.8	33.4	46.2	65.6	75.1
4	28.4	122.6	2.0	39.5	50.0	65.6	75.9
8	56.7	245.3	2.5	**55.0**	70.4	**87.2**	**94.5**
16	113.4	490.5	3.4	50.8	67.6	81.4	93.0
32	226.9	981.0	5.4	53.0	**71.1**	83.9	93.2

TABLE VII
QUANTITATIVE COMPARISON OF OUR OSSCO UNDER DIFFERENT TRANSFORMS ON THE CUFSF [20] DATASET FOR SKETCH ↔ PHOTO IMAGE RETRIEVAL. THE SCORES IN BOLD INDICATE THE BEST ACCURACY.

Transform	R@1	R@2	R@4	R@8
TT	21.4	31.7	37.9	46.2
ST	50.8	66.6	83.2	87.4
ST+TT	**55.0**	**70.4**	**87.2**	**94.5**

to a knowledge forgetting problem during iterative training. Conversely, a large z could lead to heavy memory burdens and computational costs. An appropriate value of z can effectively promote the feature extraction ability and avoid noise caused by the OST models. The choice of z is also influenced by the diversity of the training images.

Effectiveness of TT and ST. We also explore the effectiveness of the traditional rule-based transforms and the proposed synthesis-based transform. By adding these modules separately or combining them together, we assess the improvement of each module. The results are presented in Table VII, where it can be observed that the proposed synthesis-based transform contributes the most to the improvement.

Influence of Image Quality of Synthesized Images. We propose that our focus is not on generating high-quality synthesized images, but rather on utilizing OST to alleviate the influence of domain shift and extract domain-agnostic feature representations in a fully unsupervised manner. To demonstrate this, we use recent image synthesis frameworks such as CycleGAN [56], ForkGAN [71], and TSIT [72] to perform unpaired image translation with available domain labels. We then use the synthesized images for the synthesis-based transform equipped with contrastive learning. The FID score between the synthesized images and real images is computed to measure image quality. The experiments are conducted on the Foggy Cityscapes dataset [19], with results shown in Table VIII. Our OSSCo shows no performance drop even though other algorithms achieve better image synthesis performance. We have also conducted experiments under the domain-known setting, where the domain labels are available while constructing one-shot target domains and performing OST. The experimental results are also reported in Table VIII. We notice that there is only marginal performance improvement achieved, which indicates that the proposed method could simulate the domain shift among different image domains by randomly sampling one-shot target images.

TABLE VIII
QUANTITATIVE COMPARISON OF OUR OSSCO WITH OTHER UNPAIRED IMAGE TRANSLATION METHODS ON THE FOGGY CITYSCAPES [19] DATASET. † INDICATES THAT THE DOMAIN LABELS ARE AVAILABLE DURING CONSTRUCTING ONE-SHOT TARGET DOMAINS AND PERFORMING OST. THE SCORES IN BOLD INDICATE THE BEST ACCURACY.

Method	Task (FID ↓)		Foggy ↔ Clear			
	Clear → Foggy	Clear → Foggy	R@1	R@2	R@4	R@8
CycleGAN [56]	27.89	34.65	96.3	97.5	98.9	99.4
ForkGAN [71]	8.25	**11.73**	**97.3**	99.0	**99.6**	**99.8**
TSIT [72]	**7.72**	14.21	97.2	**99.1**	99.5	99.6
OSSCo	37.19	56.24	96.9	98.9	99.4	99.5
OSSCo†	30.56	37.82	96.4	97.5	99.0	99.3

Distribution Visualization. To explore why contrastive learning-based methods fail to perform well in our challenging fully unsupervised domain-agnostic image retrieval, we visualize the test sample distribution based on the learned global feature representations using T-SNE. The sample distribution of different methods is illustrated in Fig. 6. The contrastive learning-based methods struggle to achieve content disentanglement, while our OSSCo better extracts domain-agnostic feature representations, resulting in a more reliable feature space for content similarity measurement.

More OST Results. We provide additional visual results generated by our one-shot unpaired image-to-image translation model in Figs. 7 and 8. The OST model demonstrates a strong ability to synthesize images with the same appearance as the

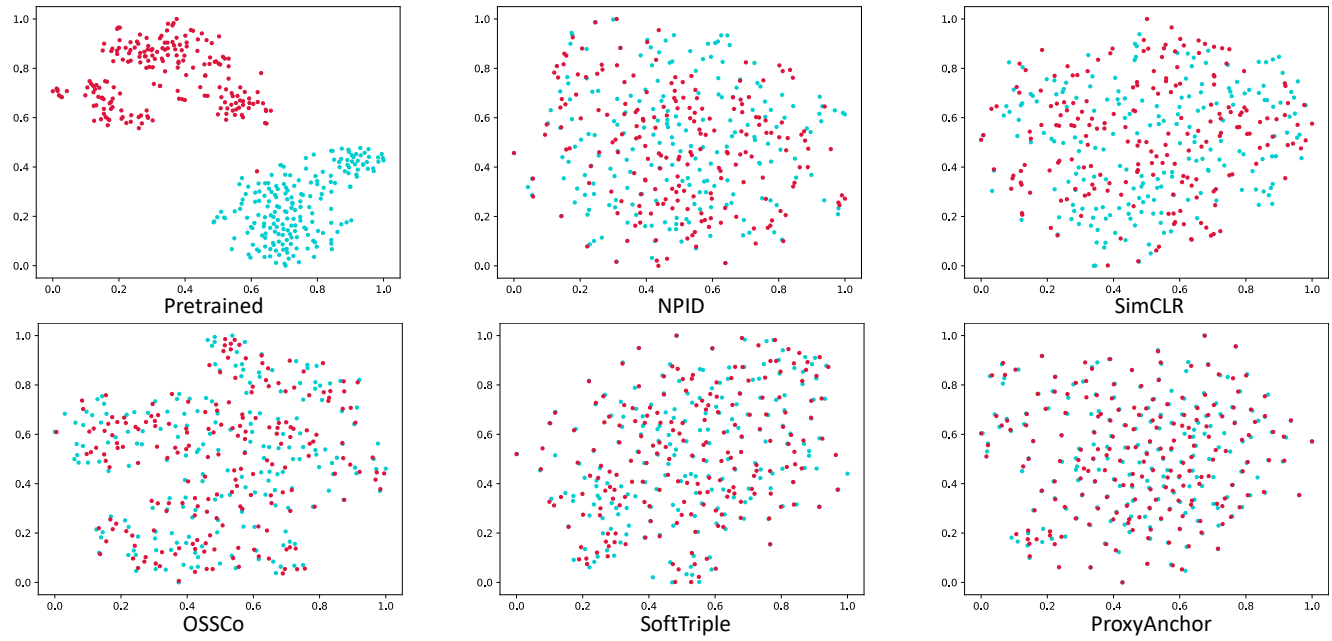


Fig. 6. The test sample distribution of different methods based on the learned global representation, conducted on the CUFSF [20] dataset. The cyan-blue and red points indicate the photo and sketch samples, respectively.

Fig. 7. The translated outputs of our OST model by choosing different one-shot images. The input images and one-shot images consist of original raw images from the Cityscapes [62] dataset and simulated foggy images from the Foggy Cityscapes [19] dataset. The images in the yellow boxes indicate the images generated by our OST model.

selected one-shot image for both Foggy Cityscapes [19] and CUFSF [20] datasets.

V. Discussion

Limitations. While our proposed method shows promising results in various domain-agnostic image retrieval tasks, it does have some limitations. One key limitation is that the method may not lead to significant performance improvements in all vision tasks. In certain scenarios, such as complex and diverse datasets like ImageNet [27] and COCO [73], where objects and backgrounds vary greatly, the limited one-shot samples may not be sufficient for the generative model to learn effective representations. Additionally, our method is not applicable in cases where there is no clear definition of a "domain" within the image data, making it impractical for datasets like ImageNet and COCO. We provide the failure cases of OST in Fig. 9. As demonstrated, when the images are too complicated or contain too many conceptions, our OST model cannot model the required implicit mapping function between two visual domains. Furthermore, if there is no explicit definition of the clear image domains, our method will also fail to capture the domain gap.

Practical Scenarios. Our method excels in scenarios where there is content overlap between samples from different distributions, such as sketch $\leftrightarrow$ photo, facades $\leftrightarrow$ label, and various illumination and weather translations. It is particularly effective in handling low-level vision tasks involving style, texture, and global transformations. However, it may struggle in cases with large pose, viewpoint, and background diversity among training data. The current OST model is limited to low-level tasks and lacks the ability to abstract high-level semantic representations.

Future Work. Future work could focus on addressing the overfitting problem in OST training, where the model tends to generate extremely similar samples as the one-shot image, regardless of the input image [33], [34]. Exploring techniques to avoid noise and uncertainty caused by image translation is also essential. Additionally, extending the application of

Fig. 8. The translated outputs of our OST model by choosing different one-shot images. The input images and one-shot images consist of original raw images from the CUFSF [20] dataset. The images in the yellow boxes indicate the images generated by our OST model.

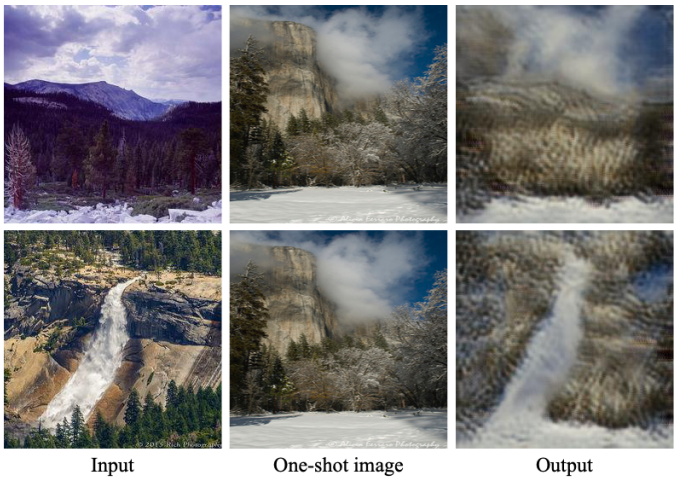


Fig. 9. The failure cases of OST. We cannot generate reasonable synthesized images when the one-shot image is complicated and there is no explicit domain shift between the source domain and the formulated one-shot target domain.

the proposed synthesis-based transform to other unsupervised learning tasks and exploring its effectiveness in different vision domains could be interesting areas of research.

VI. Conclusion

In this paper, we introduce a novel domain-agnostic image retrieval task, where domain information and label annotations are unavailable. To address this challenging problem, we propose the OSSCo method, which achieves competitive image retrieval performance even when compared with supervised methods that rely on extensive label annotations. The key contribution of our work lies in the effective combination of OST and contrastive learning through the one-shot synthesis-based transform module, which can simulate domain disparities. The impressive image retrieval results and the annotation-free training scheme demonstrate the high potential of our method for real-world applications. Our work pioneers a new direction by integrating image synthesis as an effective data transformation to enhance domain-agnostic feature extraction ability.

References

[1] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 799–807. 1, 3

[2] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017. 1

[3] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2021. 1

[4] C. Bai, H. Li, J. Zhang, L. Huang, and L. Zhang, "Unsupervised adversarial instance-level image retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 2199–2207, 2021. 1

[5] F. Liu, C. Gao, Y. Sun, Y. Zhao, F. Yang, A. Qin, and D. Meng, "Infrared and visible cross-modal image retrieval through shared features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4485–4496, 2021. 1

[6] S. Li, Y. Guo, H. Ren, Z. Wang, K. Ren, C. Liu, H. Lin, and J. Shi, "Fcnet: A feature context network based on ensemble framework for image retrieval," *IET Computer Vision*, vol. 16, no. 4, pp. 295–306, 2022. 1

[7] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *arXiv preprint arXiv:1706.06064*, 2017. 1

[8] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000. 1

[9] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)," *Neurocomputing*, vol. 452, pp. 675–689, 2021. 1

[10] H. Ren, Z. Zheng, Y. Wu, and H. Lu, "Daco: domain-agnostic contrastive learning for visual place recognition," *Applied Intelligence*, pp. 1–14, 2023. 1, 2

[11] L. Wang, X. Qian, X. Zhang, and X. Hou, "Sketch-based image retrieval with multi-clustering re-ranking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4929–4943, 2019. 1

[12] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3226–3237, 2019. 1

[13] F. Yang, Y. Wu, Z. Wang, X. Li, S. Sakti, and S. Nakamura, "Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 2347–2360, 2020. 1

[14] H. Ren, Z. Zheng, and H. Lu, "Energy-guided feature fusion for zero-shot sketch-based image retrieval," *Neural Processing Letters*, vol. 54, no. 6, pp. 5711–5720, 2022. 1

[15] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *ACM International Conference on Multimedia*, 2018, pp. 609–617. 1

[16] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *International Conference on Computer Vision*, 2017, pp. 3456–3465. 1

[17] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 5958–5964. 1, 3

[18] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 1643–1649. 1

[19] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *European Conference on Computer Vision*, 2018, pp. 687–704. 1, 6, 7, 8, 10, 11

[20] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2008. 1, 6, 8, 9, 10, 11, 12

[21] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2179–2188. 1

[22] H. Ren, Z. Zheng, Y. Wu, H. Lu, Y. Yang, Y. Shan, and S.-K. Yeung, "Acnet: Approaching-and-centralizing network for zero-shot sketch-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5022–5035, 2023. 1

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607. 2, 3, 7, 8

[24] X. Dong, L. Liu, L. Zhu, Z. Cheng, and H. Zhang, "Unsupervised deep k-means hashing for efficient image retrieval and clustering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3266–3277, 2020. 2

[25] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *European Conference on Computer Vision*, 2016, pp. 3–20. 2

[26] C. Chang, G. Yu, C. Liu, and M. Volkovs, "Explore-exploit graph traversal for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9423–9431. 2

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. 2, 7, 11

[28] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020. 2

[29] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Annual Conference on Neural Information Processing Systems*, vol. 33, 2020, pp. 22 243–22 255. 2, 3

[30] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," in *Annual Conference on Neural Information Processing Systems*, vol. 33, 2020, pp. 21 271–21 284. 2, 3

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[32] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *Annual Conference on Neural Information Processing Systems*, vol. 31, 2018. 2, 3

[33] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Adversarial style mining for one-shot unsupervised domain adaptation," in *Annual Conference on Neural Information Processing Systems*, vol. 33, 2020, pp. 20 612–20 623. 2, 3, 11

[34] Z. Zheng, Z. Yu, H. Zheng, Y. Yang, and H. T. Shen, "One-shot image-to-image translation via part-global learning with a multi-adversarial framework," *IEEE Transactions on Multimedia*, vol. 24, pp. 480–491, 2021. 2, 3, 11

[35] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307. 3

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. 3

[37] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417. 3

[38] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8. 3

[39] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *European Conference on Computer Vision*, 2020, pp. 369–386. 3

[40] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209. 3

[41] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger, "Visual localization by learning objects-of-interest dense match regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5634–5643. 3

[42] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610. 3

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 3, 5, 7, 8, 9, 10

[44] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742. 3, 7, 8

[45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 3

[46] V. Verma, T. Luong, K. Kawaguchi, H. Pham, and Q. Le, "Towards domain-agnostic contrastive learning," in *International Conference on Machine Learning*, 2021, pp. 10 530–10 541. 3

[47] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Annual Conference on Neural Information Processing Systems*, vol. 33, 2020, pp. 9912–9924. 3

[48] E. S. Lee, J. Kim, S. Park, and Y. M. Kim, "Moda: Map style transfer for self-supervised domain adaptation of embodied agents," in *European Conference on Computer Vision*, 2022, pp. 338–354. 3

[49] R. Brüel-Gabrielsson, T. Wang, M. Baradad, and J. Solomon, "Deep augmentation: Enhancing self-supervised learning through transformations in higher activation space," *arXiv preprint arXiv:2303.14537*, 2023. 3

[50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. 3, 7, 8, 9, 10

[51] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *International Conference on Learning Representations*, 2020. 3, 7, 8, 9, 10

[52] K. Lee, Y. Zhu, K. Sohn, C.-L. Li, J. Shin, and H. Lee, "I-mix: A domain-agnostic strategy for contrastive representation learning," in *International Conference on Learning Representations*, 2021. 3, 7, 8, 9, 10

[53] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018. 4

[54] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *European Conference on Computer Vision*, 2018, pp. 289–305. 4

[55] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain

generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022. 4

[56] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, 2017, pp. 2223–2232. 4, 5, 6, 7, 8, 9, 10

[57] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797. 4, 5

[58] Z.-H. Zhou and Z.-H. Zhou, *Ensemble learning*. Springer, 2021. 5

[59] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015. 5

[60] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711. 5

[61] S. S. Halder, J.-F. Lalonde, and R. d. Charette, "Physics-based rendering for improving robustness to rain," in *International Conference on Computer Vision*, 2019, pp. 10 203–10 212. 6, 9, 10

[62] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223. 6, 8, 11

[63] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *International Conference on Computer Vision*, 2019, pp. 6450–6458. 7, 8, 9, 10

[64] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3238–3247. 7, 8, 9, 10

[65] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000. 6

[66] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *International Conference on Computer Vision*, 2017, pp. 2849–2857. 6

[67] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012. 7

[68] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807. 7

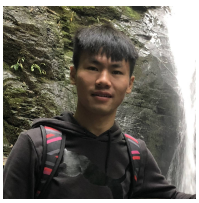
[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015. 7

[70] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Annual Conference on Neural Information Processing Systems*, 2019, pp. 8026–8037. 7

[71] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *European Conference on Computer Vision*, 2020, pp. 155–170. 10

[72] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "Tsit: A simple and versatile framework for image-to-image translation," in *European Conference on Computer Vision*, 2020, pp. 206–222. 10

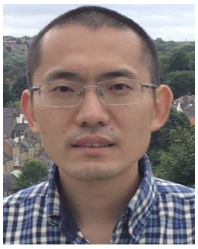
[73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755. 11



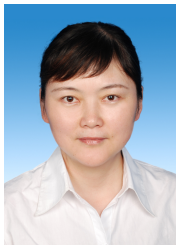
Ziqiang Zheng received his B.Eng. degree in communication engineering from the Ocean University of China in 2019. He is currently with the Center for Future Media, the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia content analysis and computer vision.

Hao Ren (Member, IEEE) received his B.Eng. degree in software engineering from the Zhejiang University of Technology in 2017. He is currently with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China. His research interests include pattern recognition and computer vision.

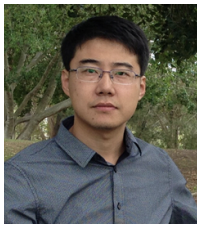
Yang Wu (Member, IEEE) received a BS degree and a Ph.D. degree from Xi'an Jiaotong University in 2004 and 2010, respectively. He is currently a principal researcher with Tencent AI Lab. From Jul. 2019 to May 2021, he was a program-specific senior lecturer with the Department of Intelligence Science and Technology, Kyoto University. He was an assistant professor of the NAIST International Collaborative Laboratory for Robotics Vision, Nara Institute of Science and Technology (NAIST), from Dec. 2014 to Jun. 2019. From 2011 to 2014, he was a program-specific researcher with the Academic Center for Computing and Media Studies, Kyoto University. His research is in the fields of computer vision, pattern recognition, as well as multimedia content analysis, enhancement and generation.



Weichuan Zhang (Member, IEEE) received the MS degree in signal and information processing from Southwest Jiaotong University, China, and the PhD degree in signal and information processing in National Lab of Radar Signal Processing, Xidian University, China. He is a research fellow with Griffith University, QLD, Australia. His research interests include computer vision, image analysis, and pattern recognition.



Hong Lu (Member, IEEE) received the B.Eng. and M.Eng. degrees in computer science and technology from Xidian University, Xi'an, China, in 1993 and 1998, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2005. From 1993 to 2000, she was a Lecturer and a Researcher with the School of Computer Science and Technology, Xidian University. From 2000 to 2003, she was a Research Student with the School of Electrical and Electronic Engineering, Nanyang Technological University. Since 2004, she has been with the School of Computer Science, Fudan University, Shanghai, China, where she is currently a Professor. Her current research interests include computer vision, machine learning, pattern recognition, and robotic tasks.



Yang Yang (Senior Member, IEEE) received the bachelor's degree from Jilin University, Changchun, China, in 2006, the master's degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from The University of Queensland, Brisbane, Australia, in 2012, all in computer science. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively. He is currently a Professor of the National Thousand Talents Plan and the Dean of the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, where he is also the Director of the Center for Future Media. He is also an Honorary Professor with The University of Queensland, Brisbane, QLD, Australia. His research interests include multimedia search, computer vision, artificial intelligence, and big data management. Dr. Shen is an ACM Distinguished Member and an Optical Society of America (OSA) Fellow.