ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 784-804
DOI 10.5755/j01.itc.55.2.44016

Weakly Supervised Multi-Source Fine-Grained Image Recognition via Class-Conditional Adversarial Adaptation and Confidence-Gated Pseudo-Labeling

Received 2026/01/05 | Accepted after revision 2026/03/31

HOW TO CITE: Yang, B., Ni, W., Dong, M. (2026) Weakly Supervised Multi-Source Fine-Grained Image Recognition via Class-Conditional Adversarial Adaptation and Confidence-Gated Pseudo-Labeling. *Information Technology and Control*, 55(2), 784-804. https://doi.org/10.5755/j01.itc.55.2.44016

Weakly Supervised Multi-Source Fine-Grained Image Recognition via Class-Conditional Adversarial Adaptation and Confidence-Gated Pseudo-Labeling

Bensheng Yang, Weichuan Ni
School of Artificial Intelligent, Guangzhou Huashang College, Guangzhou 511300, China

Meixia Dong*
School of Information and Intelligence Engineering, Guangzhou Xinhua University, Dongguan 523133, China

Corresponding author: dongmeixia@xhsysu.edu.cn

Differences in devices and imaging styles across multi-source images lead to distribution shifts between training and deployment. Under weak annotation, models supervised only by image-level labels are easily affected by background textures, causing category confusion and pseudo-label noise accumulation in target domains. To address this issue, this paper proposes CCAP-Net, a weakly annotated multi-source subcategory recognition network based on class-conditional adversarial adaptation and collaborative confidence pseudo-labelling. CCAP-Net introduces complementary component attention branches to mine diverse local discriminative regions and fuses global and local representations through a gating mechanism to enhance subtle subcategory cues. A conditional domain discriminator is further designed to align domain distributions while preserving class-related structure. To improve robustness under multi-source imbalance and unreliable target predictions, adaptive source contribution learning and confidence gating are incorporated to dynamically regulate source influence and suppress low-confidence target samples. In addition, a teacher network generates pseudo labels for unlabeled target data, and target training combines strong/weak augmentation consistency with class-centre similarity filtering to reduce error propagation. Experiments on Office-Home, DomainNet, and CompCars demonstrate that CCAP-Net consistently improves target-domain accuracy, class-balanced metrics, and calibration reliability over representative baselines.

KEYWORDS: Multi-source transfer learning; Weakly supervised learning; Subcategory recognition; Pseudo-label learning; Confidence calibration

1. Introduction

In industrial vision scenarios, a single recognition task often aggregates multi-source image data from different production lines, equipment, and imaging conditions. Cross-source variations directly cause inconsistencies between training and deployment distributions. Class distinctions refined to model, batch, or operational conditions make inter-class differences more reliant on local texture and structural cues. However, the cost of manual annotation means that common training supervision remains at the image-level category label stage. In recent years, large-scale pre-training and visual Transformers have significantly enhanced the transferability of vision models. However, under scenarios involving multi-source data blending, target domains lacking human annotations, and finely differentiated categories, issues such as category confusion and unstable transfer gains remain prevalent [6, 23, 2, 12].

For fine-grained object recognition, existing approaches have employed Transformer architectures to enhance focus on key regions and improve the utilisation efficiency of local differences. However, such enhancements typically assume relative consistency between training and testing distributions. When multi-source data exhibits pronounced stylistic discrepancies with the target domain, reliance solely on local attention may still misinterpret background or stylistic cues as discriminative features, thereby diminishing generalisation capabilities. Multi-source domain adaptation research seeks to mitigate source conflicts and transfer bias through approaches such as sample-wise dynamic transfer, self-step pseudo-label introduction, and confidence-based learning strategies to enhance target domain performance. Nevertheless, under weakly labelled conditions, two core challenges persist: The varying effectiveness of different sources for the target domain evolves throughout training, making fixed weighting impractical. Furthermore, the unreliability of early predictions in the target domain allows pseudo-label noise to persistently amplify and distort cross-domain learning signals [18, 24, 5, 19].

Pseudo-labelling and semi-supervised learning provide viable training signal sources for unlabelled target domains. Mechanisms such as graph consistency, curricular thresholds, and adaptive thresholds can mitigate pseudo-label noise and expand effective sample coverage to some extent. However, under multi-source transfer, the category structure becomes more susceptible to interference from stylistic variations. Threshold strategies exhibit high sensitivity to probability quality, readily leading to class distribution skew and persistent neglect of hard-to-classify samples [1, 15, 31, 34]. Moreover, real-world industrial deployments often involve continuously evolving target distributions. While test-phase adaptation and continuous adaptation methods enable parameter updates under passive data conditions, their effectiveness remains contingent upon the reliability of prediction probabilities. Significant calibration errors may lead to erroneous updates and performance degradation [28, 30]. Therefore, in fine-grained classification tasks involving multi-source weakly labelled data transfer, it is necessary to simultaneously enhance the category directionality of cross-domain learning signals, control the propagation of noise in target domain pseudo-labels, and improve the reliability and usability of probability outputs [25, 24, 29].

Based on the above motivations, this paper proposes CCAP-Net for fine-grained classification tasks involving multi-source weakly labeled data transfer. It integrates class-conditional cross-domain learning with target-domain self-training into a unified training process, establishing mutual constraints among cross-domain alignment, source contribution update, and pseudo-label guidance. This approach enhances category discrimination capabilities while suppressing noise propagation under conditions of significant source variability and sparse target-domain annotations. The contributions are as follows:

1 The CCAP-Net framework introduces complementary component attention through class-conditional adversarial alignment, explicitly injecting discriminative local cues into cross-domain learning signals. It also designs a multi-source contribution adaptation mechanism tailored for target domain effectiveness, dynamically adjusting source weights based on training progress and target domain learning status to avoid over- or under-weighting caused by fixed weights or weakly coupled updates.

2 We introduce a confidence-gated pseudo-label collaborative training strategy. A teacher network

generates target domain pseudo-labels, which are progressively expanded through confidence filtering and class-center constraints to broaden the coverage of usable target samples. Simultaneously, high-confidence pseudo-labels are incorporated into class-conditional alignment and source contribution updates, mitigating the risk of amplifying biased pseudo-labels during multi-source transfer and enhancing training stability.

3 Conducting systematic experiments and analysis across multiple public benchmarks, we report not only accuracy and macro-average metrics but also calibration-related indicators and stability curves. Comparisons with recent multi-source unsupervised domain adaptation and self-training strong baselines validate CCAP-Net's effectiveness and reproducible gains in multi-source weakly labeled transfer settings.

The remainder of this paper is organised as follows: Section 1 reviews relevant prior work pertinent to this study. Section 2 presents the problem formulation and overall framework, outlining the network architecture and training process of CCAP-Net. Section 3 introduces the class-conditional domain adversarial learning approach, encompassing complementary component attention enhancement, conditional domain discriminator design, and domain distribution learning strategies featuring multi-source contribution adaptation and confidence-based gating. Section 4 elaborates on the confidence-based pseudo-label collaborative learning mechanism, encompassing pseudo-label generation by the teacher network, target domain training with strong-weak enhancement constraints, and pseudo-label participation in conditional domain learning and class-centre selection. Section 5 presents experimental setups, comparative results, and visual analyses across Office-Home, DomainNet, and CompCars datasets. Section 6 concludes the paper and discusses future research directions.

2. Relevant Work

In recent years, subcategory image recognition has progressively shifted from local-region-based convolutional modelling towards stronger global relationship learning. TransFG introduced the Transformer architecture into subcategory recognition and enhanced category discrimination capabilities by designing feature interactions better suited to this task, providing crucial reference for subsequent utilisation of local cues under weakly supervised conditions [11]. Under image-level category supervision alone, reliably obtaining transferable local cues remains challenging. PMRC addresses this by mining pose and discriminative regions, learning more distinctive regional combinations under weak supervision to mitigate interference from background and texture perturbations on training signals [27]. Complementing weak-supervision local feature mining is another mainline approach: pseudo-label-driven unsupervised learning. Zhang et al. introduced curriculum-based pseudo-label selection into semi-supervised training, correlating the order of unsupervised sample utilisation with the model's learning state. This mitigates misdirection caused by low-quality pseudo-labels in the early stages [35].

For multi-source transfer learning, discrepancies and conflicts between source domains amplify training uncertainty in the target domain. SPS employs an iterative strategy to progressively introduce more reliable target samples into training, mitigating the impact of mixed multi-source information on target domain learning while enhancing the proportion of usable pseudo-labels [32]. Regarding sub-classification network architectures, AP-CNN aggregates discriminative regions across multiple scales via attention pyramids. This ensures more comprehensive coverage of local cues under image-level supervision, thereby improving separability when intra-class variations are substantial [4]. In unsupervised domain adaptation, Feng et al. mitigate target domain pseudo-label noise through a complementary pseudo-label mechanism, preventing category bias caused by training being driven by a single pseudo-label distribution. This provides a reference approach for pseudo-label improvement when introducing confidence gating in the target domain [8]. For unsupervised multi-source domain adaptation, Li et al. employed dynamic classifier matching and collaborative updates across sources. This approach facilitates the preservation of discernible structures in category boundaries during migration from multiple sources to the target domain, mitigating performance degradation caused by category confusion [16].

Under more robust multi-source settings, Chen et al. employed joint distribution mixture modelling to integrate multi-source information, enabling the model to form more appropriate cross-domain learning pathways during training based on source variations. This approach enhances target domain generalisation capabilities while mitigating the adverse effects of source domain conflicts [3]. Research trends indicate that maintaining class separability through target domain self-training and uncertainty control remains a core challenge when source data is inaccessible or only pre-trained models are available. This has driven the development of more refined confidence control and sample selection mechanisms [7]. In subclassification, Xu et al. introduced ensemble learning within Transformer representations, fostering stronger complementarity of discriminative information across multiple layers. This enhances sensitivity to local variations and improves generalisation capabilities [33].

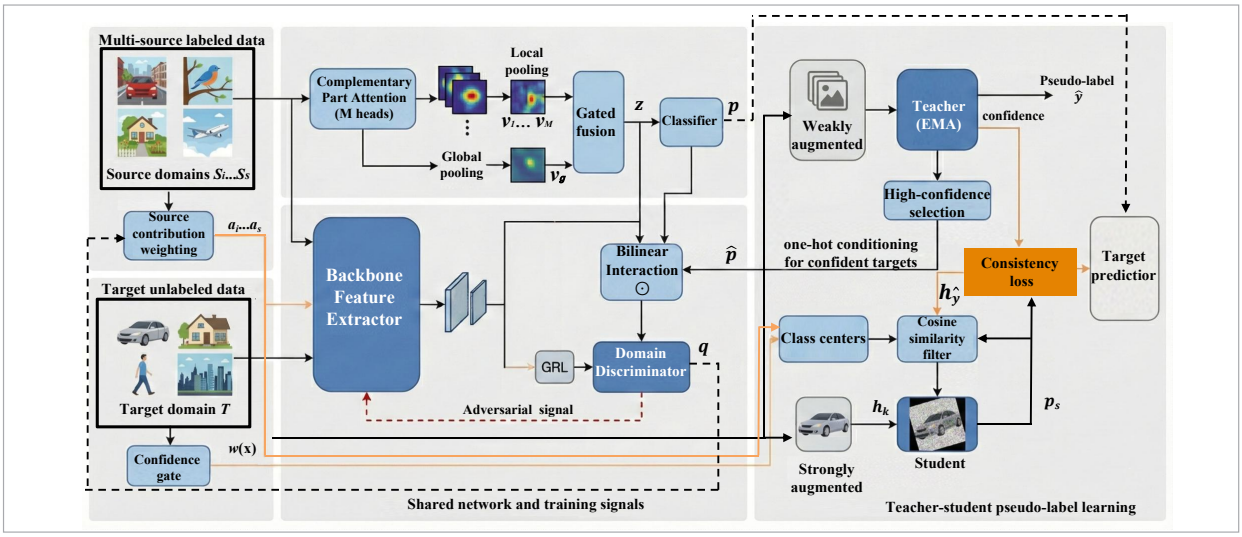
Beyond the aforementioned approaches, peak suppression and knowledge-guided Transformers mitigate the degradation in generalisation caused by excessive focus on local regions during subcategory recognition by suppressing high-response areas prone to overfitting and introducing additional constraints [22]. The cross-layer mutual learning strategy proposed by Liu et al. enhances the complementary discrimination capabilities across different layers through inter-layer information feedback, enabling the network to simultaneously retain both global appearance and local difference information, thereby improving the overall discriminative quality in subcategory recognition [21]. In weakly supervised transfer scenarios, confidence scores not only influence pseudo-label usability but also determine the utilisation intensity of target samples during domain learning. Patra et al. enhanced the calibrability of probability outputs from an explicit regularisation perspective, providing a more reliable probabilistic foundation for employing confidence scores in gating and filtering mechanisms [26].

3. Methodological Framework

This paper addresses weakly labelled multi-source image recognition tasks by constructing a unified model comprising a backbone network, a component attention branch, a classification branch, and a conditional domain discriminator branch, as illustrated in Figure 1. The input comprises images from multiple source domains along with their image-level categories, alongside unlabelled target domain images. Following feature extraction by the backbone network, the component attention branch generates multiple complementary attention maps on the feature map. These aggregate information from distinct local regions into several local vectors, which are then fused with a global vector via gating to form the final representation for fine-grained classification. Concurrently, the conditional domain discriminator branch jointly interacts the fused representation with the category probabilities output by the classification branch. This serves as input to the domain discriminator, employing adversarial training to suppress interference from domain differences in classification. To address the challenges of uneven multi-source domain variations and unreliable early predictions in the target domain, the model incorporates adaptive multi-source contribution weighting and confidence-based gating mechanisms. The former dynamically adjusts the influence of each source domain during domain training, while the latter modulates participation based on prediction confidence for target samples. This ensures training concentrates on more beneficial source domains and more reliable target samples.

In the absence of target labels, this paper further introduces a teacher network for pseudo-label generation and collaborative learning. The teacher network is derived from the exponential moving average of the student's parameters, generating target domain class distributions on weakly augmented views. High-confidence pseudo-labels are then filtered via thresholding. Weak reinforcement is solely applied to the pseudo-label generation process of the teacher network. Specifically, the teacher network performs forward inference on the weakly reinforced view to obtain the probability distribution and maximum confidence for the target sample, thereby generating high-confidence pseudo-labels. The student network undergoes target domain classification training on strongly augmented views using these pseudo-labels as supervision, while incorporating cross-augmentation variance suppression to mitigate output fluctuations caused by augmentation. To

Figure 1
Overall Methodological Framework Diagram.



enhance domain training's effectiveness, high-confidence pseudo labels contribute as single-hot inputs to the conditional domain discriminator's categorical input layer, reinforcing classification directionality. This synergises with a multi-source contribution strategy to improve cross-domain transfer performance. Addressing potential pseudo label errors near class boundaries, the paper maintains category-updated centroid vectors and employs secondary screening via centroid similarity to mitigate the impact of ambiguous samples.

4. Conditional Domain Adversarial

4.1. Component Attention Enhancement

In weakly labelled scenarios, only image-level categories are available, with local variations often obscured by background, lighting, and texture changes. This leads to substantial intra-class variability and minimal inter-class distinctions. To address this, this paper proposes complementary component attention enhancement. Multiple attention heads automatically identify several complementary local regions within the same image, then fuse global and local information through a gated mechanism. This approach broadens the coverage of local cues essential for category discrimination. Given an input image $x \in R^{3\times H_0\times W_0}$, the backbone network outputs feature maps:

$$U = F(x) \in R^{C\times H\times W}, \tag{1}$$

where C denotes the number of channels, H, W represent spatial dimensions, and $F(\cdot)$ is the backbone network. For local detection, this paper configures M attention heads. The score for the m-th attention head at position (h,w) is:

$$s_m(h,w) = w_m^{\top} U_{:,h,w} + b_m, m = 1, \ldots, M. \tag{2}$$

In Equation (2), $U_{:,h,w} \in R^C$ denotes the channel vector at that position, while $w_m \in R^C$ and $b_m \in R$ are learnable parameters. To transform the scoring into comparable spatial weights, this paper employs temperature-based normalisation to obtain the attention map [9, 7, 4]:

$$A_m(h,w) = \frac{exp\left(\frac{s_m(h,w)}{\tau_a}\right)}{\sum_{h'=1}^{H}\sum_{w'=1}^{W} exp\left(\frac{s_m(h',w')}{\tau_a}\right)}, \tag{3}$$

where $\tau_a > 0$ governs the degree of attention concentration. In the early stages of training, a larger τ_a is employed to broaden the attention coverage, preventing premature fixation on a single local region. Subsequently, τ_a is gradually reduced to sharpen the focus, thereby enhancing critical texture and shape

distinctions. The local vector is obtained by performing weighted aggregation on the feature map using the attention map:

$$v_m = \sum_{h=1}^{H} \sum_{w=1}^{W} A_m(h,w) U_{:,h,w} \in R^C. \quad (4)$$

Here, v_m denotes the aggregated result for the m-th local region. Concurrently, the global vector $v_g \in R^C$ is retained, derived from global average pooling over U to provide overall appearance and background statistics. This paper proposes a gated fusion approach, synthesising global and multi-local information into a classification input:

$$z = W_g v_g + \sum_{m=1}^{M} \alpha_m W_m v_m. \quad (5)$$

In Equation (5), $W_g, W_m \in R^{d \times C}$ and represent projection matrices, where d denotes the dimension of the fusion vector. The coefficient α_m is generated adaptively from the sample data to reflect the contribution of each local region within the current image. α_m is implemented using the following formula:

$$\alpha_m = \frac{exp\,(r^\top tanh\,(W_\alpha v_m))}{\sum_{j=1}^{M} exp\,(r^\top tanh\,(W_\alpha v_j))}, \quad (6)$$

where $W_\alpha \in R^{d_\alpha \times C}$ and $r \in R^{d_\alpha}$ are learnable parameters, where d_α denotes the intermediate dimension. This approach offers the advantage that α_m naturally reduces its influence when certain local features become unavailable due to occlusion or noise. Conversely, when a local feature presents distinct discriminative cues, it is assigned a greater weight.

A common issue with multi-attention heads is that multiple heads may converge on the same location, leading to redundant local information. To promote complementarity, this paper introduces an overlap suppression term: flatten A_m to $a_m \in R^{HW}$ and perform L_2 normalisation to obtain $\hat{a}_m$, defined as:

$$L_{div} = \frac{2}{M(M-1)} \sum_{m<n} (\hat{a}_m^\top \hat{a}_n)^2, \quad (7)$$

where $\hat{a}_m^\top \hat{a}_n$ reflects the spatial overlap between two attention maps. Greater overlap yields a higher value, increasing L_{div}. Training then drives different attention heads to focus on distinct regions, forming complementary local ensembles. This paper combines this constraint with gating fusion, ensuring that local features are not only diverse but also effectively utilised by the classification branch.

4.2. Design of the Conditional Domain Discrimination Unit

Multi-source data originates from diverse devices, lighting conditions, and operational scenarios. Variations in appearance across domains for the same category can cause features learned during training to exhibit domain-specific bias. When relying solely on features for domain discrimination, domain differences and category differences tend to conflate, leading to category confusion. To address this issue, this paper proposes a conditional domain discriminator, as illustrated in Figure 2. This approach bases domain discrimination on a joint input of features and class probabilities, aligning domain information utilisation more closely with the classification objective. It also provides an interface for subsequent confidence-based pseudo-label participation in domain training.

From the preceding section, we obtain the fusion vector $z \in R^d$, and the classification head outputs class probabilities:

$$p = C(z) \in R^K, \sum_{c=1}^{K} p_c = 1, p_c \geq 0, \quad (8)$$

where K denotes the number of classes and p represents the confidence distribution across classes. This paper employs low-dimensional bilinear interactions to construct conditional inputs, thereby circumventing the dimensionality explosion inherent in direct outer products:

$$g = (R_z z) \odot (R_p p) \in R^{d_g}, \quad (9)$$

where $R_z \in R^{d_g \times d}$ and $R_p \in^{d_g \times K}$ denote projection matrices, where d_g represents the conditional input dimension, and $\odot$ denotes element-wise multiplication. The key to this construction lies in the di-

mensional interaction between category confidence and visual information: when a sample exhibits high confidence in certain categories, the corresponding dimensions become more prominent. This allows the domain discriminator to focus on domain differences relevant to the categories, rather than being dominated by background variations. Let there be K_s source domains and 1 target domain, with a total of K_s+1 domains. The domain discriminator outputs a domain probability vector:

$$q = D(g) \in R^{K_s+1}, \sum_{r=1}^{K_s+1} q_r = 1, \tag{10}$$

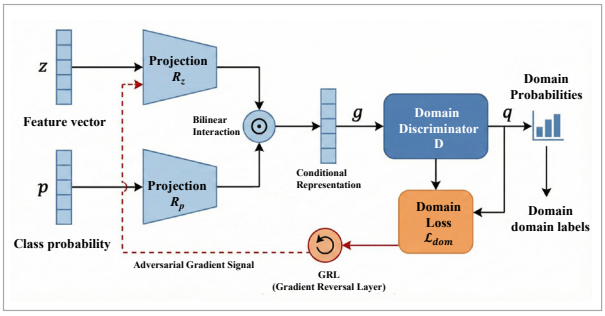
where q_r denotes the predicted probability that the sample belongs to the r domain. During training, an adversarial approach is employed for updates: the domain discriminator enhances the ability to distinguish domains, while the feature extraction component reduces domain separability, simultaneously preserving classification capability for the source domain. Let the network generating z be denoted as F, the classification head as C, and the training objective be formulated as:

$$\min_{F,C} L_{cls}(F, C) - \lambda L_{dom}(F, C, D). \tag{11}$$

In Equation (11), L_{cls} denotes the source domain image-level supervision loss, L_{dom} represents the domain discriminative loss, and λ>0 controls the strength of the domain term. This paper employs gradient reversal for adversarial updates: when updating D, L_{dom} is minimised normally, when updating F, the gradient of L_{dom} is reversed and scaled by λ, thereby suppressing domain separability. Unlike domain discrimination relying solely on z, the conditional input in Equation (9) introduces categorical information into the domain discrimination process. Furthermore, this paper incorporates the multi-local information generated in Section 3.1 into z, enabling conditional domain discrimination to simultaneously perceive both global and local categorical signals.

Figure 2

Conditional Domain Discrimination.



4.3. Domain Distribution Matching with Multi-source Contribution Adaptation and Confidence-based Gating

The preceding section outlined the input construction and adversarial training methodology for conditional domain discriminators. However, direct application remains challenging in multi-source scenarios: domains from different sources exhibit varying degrees of divergence from the target domain. Treating all source domains equally risks allowing more divergent domains to dominate training, thereby undermining target domain performance. Early predictions for the target domain are often unreliable; employing all target samples with equal intensity in domain distribution learning introduces noise that significantly interferes with category-specific training signals.

To address these issues, this paper proposes a domain distribution learning strategy featuring multi-source contribution adaptation and confidence gating. This approach assigns contribution weights to each source domain that update during training, while simultaneously allocating participation intensity based on the prediction confidence of target samples. Consequently, domain distribution learning prioritises data fragments that are more reliable and closer to the target domain, thereby providing a cleaner target sample set for subsequent pseudo-label learning.

First, define a prediction confidence score for target domain samples $x \in B_t$:

$$c(x) = \max_{1 \le k \le K} p_k(x), \quad p(x) = C\big(z(x)\big), \tag{12}$$

In Equation (12), $p_k(x)$ denotes the predicted probability that the sample belongs to class k, where K is the number of classes, and $z(x)$ is the fused vector obtained in Section 3.1. Based on confidence scores,

this paper employs a smoothing gate to generate the target sample weights:

$$w(x) = \sigma\left(\frac{c(x) - \tau}{\kappa}\right) \in (0,1), \tag{13}$$

where τ denotes the gating threshold, whilst κ governs the transition bandwidth. This design implies that when $c(x)$ substantially exceeds the threshold, $w(x)$ approaches 1, enabling the target sample to participate in domain distribution learning with heightened intensity. Conversely, when $c(x)$ is low, $w(x)$ approaches 0, thereby suppressing the sample's influence on domain distribution learning and mitigating interference from unreliable early predictions. Adaptive allocation of source domain contribution. Let there be K_s source domains, with each domain's sample set in the current batch denoted as $B_s^{(k)}$. Intuitively, source domains closer to the target domain should receive greater training weight. This paper constructs source domain scores using the conditional domain discriminator's output: when samples from a source domain are harder to distinguish accurately under the domain discriminator, it often indicates greater proximity to the target domain in the current feature space. Let the domain discriminator output be denoted as $q(x)=D(g(x))\in R^{K_s+1}$, where the k dimension corresponds to the k source domain and the K_s+1 dimension corresponds to the target domain. The score for the k source domain is defined as:

$$r_k = \frac{1}{|B_s^{(k)}|} \sum_{x \in B_s^{(k)}} (1 - q_k(x)) . \tag{14}$$

The smaller the value of $q_k(x)$, the more difficult it is for the domain classifier to assign the sample to the k source domain, resulting in a larger r_k. Normalising the scores for each source domain yields the contribution weights:

$$\alpha_k = \frac{exp\ (r_k/\tau_s)}{\sum_{j=1}^{K_s} exp\ (r_j/\tau_s)}, k = 1, \ldots, K_s, \tag{15}$$

where τ_s denotes the temperature coefficient, which regulates the sharpness of the weight distribution. A larger α_k indicates that the source domain is more beneficial to the target domain at the current stage and should be assigned a higher weight during training.

Building upon this, this paper jointly incorporates source domain weights and target sample gating into the domain loss. Let the domain label be $d(x) \in \{1,\ldots,K_s+1\}$, and the probability that the domain discriminator assigns to the true domain is $q_{d(x)}(x)$. The weighted domain loss is expressed as:

$$L_{dom}^{w} = -\sum_{k=1}^{K_s} \alpha_k \frac{1}{|B_s^{(k)}|} \sum_{x \in B_s^{(k)}} log\ q_k(x) - \frac{1}{|B_t|} \sum_{x \in B_t} w(x) log\ q_{K_s+1}(x) \tag{16}$$

In Equation (16), the first term weights each source domain by α_k, enabling domains with higher contribution to carry greater weight in domain training. The second term weights target domain samples by $w(x)$, amplifying the contribution of high-confidence target samples to domain training while diminishing the influence of low-confidence samples. Equation (16) enables domain training to concentrate on source domains most beneficial to the target domain and prioritise more credible samples within the target domain. This approach accumulates a higher-quality target candidate set for pseudo-label learning in the subsequent section and facilitates the enhancement of pseudo-label confidence distributions.

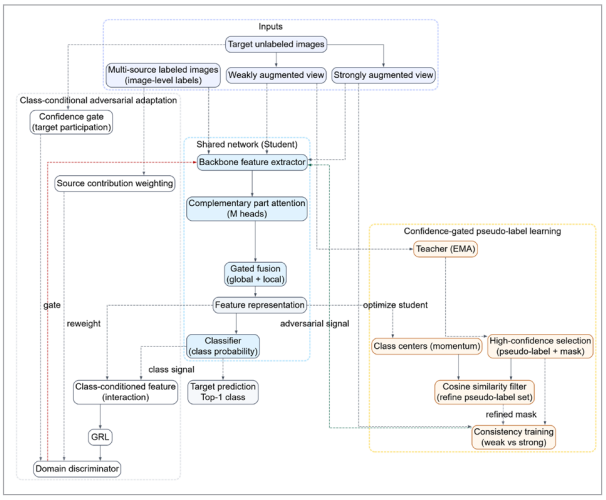
5. Confidence-Based Pseudo-Label Collaborative Learning

In the absence of human-curated categories within the target domain, pseudo labels serve as both the primary source of supervision for training and the principal risk factor for error propagation. To enhance target domain usability while mitigating bias propagation from low-confidence samples, this paper introduces a teacher-pupil collaborative mechanism within a unified framework. This design integrates pseudo label generation, filtering, and training constraints into a coordinated system. The teacher network outputs smoother category distributions on weakly augmented views to generate high-confidence pseudo labels. The student network then undergoes

classification training on strongly augmented views using these pseudo labels as supervision, incorporating cross-augmentation consistency constraints to stabilise predictions. Category-centred similarity filtering further eliminates marginal samples, reducing interference from ambiguous samples on the training signal. The overall process is illustrated in Figure 3.

Figure 3

Schematic diagram of the CCAP-Net confidence-based pseudo-label collaborative learning process.



5.1. Teacher Network Generating Pseudo-Labels

When target domains lack manually labelled categories, directly assigning classes to target samples using the current model is susceptible to early prediction bias. To address this, this paper introduces a teacher network [10,20], which employs exponential moving averages of parameters to generate smoother prediction distributions. These are used to produce pseudo labels for the target domain while controlling their reliability. Let the student network parameters be denoted as θ and the teacher network parameters as $\bar{\theta}$. The update rule is as follows:

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta. \tag{17}$$

In Equation (17), $\mu \in (0,1)$ controls the teacher update rate. A larger μ results in slower teacher adaptation, producing smoother outputs that are better suited to suppressing noise propagation during the early training stages.

Applying a weaker data augmentation $a_w(\cdot)$ to the target domain sample x yields $x^w = a_w(x)$. The teacher network outputs category probabilities as follows:

$$\tilde{p}(x) = Softmax\left(C\left(F\left(x^w;\bar{\theta}\right)\right)\right) \in R^K. \tag{18}$$

In Equation (18), $F(\cdot)$ denotes the feature extraction component comprising the backbone and attention-enhanced modules, $C(\cdot)$ represents the classification head, K denotes the number of categories, and $\tilde{p}(x)$ signifies the teacher's confidence distribution across categories. The pseudo-label adopts the category with the highest probability in the teacher's distribution, employing this maximum probability as the confidence score. An indicator determines whether the sample enters the pseudo-label set:

$$\begin{cases} \hat{y}(x) = arg\max\limits_k \tilde{p}_k(x) \\ c(x) = \max\limits_k \tilde{p}_k(x) \\ m(x) = I(c(x) \geq \tau) \end{cases}, \tag{19}$$

where $\hat{y}(x)$ denotes the pseudo-label category, $c(x)$ represents the confidence score, τ is the threshold, and $m(x) \in \{0,1\}$ indicates whether to select the sample. This paper employs a higher threshold during the early training phase to prioritise more reliable samples, subsequently relaxing the threshold progressively to gradually expand the pseudo-label coverage. This approach balances reliability with sample size.

5.2. Classification Training of Target Domains for Strength-Weakness Enhancement Constraints

Once the target domain pseudo labels are established, they must be transformed into learning signals effective for the student network. This paper employs two augmented views of the same target image to construct training constraints: weak augmentation generates pseudo labels, while strong augmentation trains the student to output the same category under more perturbed inputs, thereby enhancing adaptability to variations in lighting, texture, and viewpoint. Weakly and strongly augmented views are constructed for target domain sample x:

$$x^w = a_w(x), x^s = a_s(x), \tag{20}$$

where $a_w(\cdot)$ preserves primary structural information, whilst $a_s(\cdot)$ introduces stronger colour perturbations, local occlusions or lightweight geometric transformations to enhance the student's discriminative capability under complex inputs. The student network's output on strongly augmented views is:

$$p_s(x) = Softmax\left(C\left(F(x^s;\theta)\right)\right) \in R^K, \tag{21}$$

where $p_s(x)$ denotes the student's predicted distribution across categories. Employing the pseudo labels and selection indicators derived from Equation (19), the target domain classification loss is expressed as:

$$L_t = -\frac{1}{|B_t|}\sum_{x\in B} m(x) log\,(p_s(x)_{\hat{y}(x)}), \tag{22}$$

where B_t denotes the target domain batch sample set, while $p_s(x)_{\hat{y}(x)}$ represents the student's predicted probability for the pseudo-label category. Only high-confidence samples where $m(x)$=1 participate in loss calculation, thereby mitigating noise introduced by low-confidence samples. Relying solely on pseudo-label cross-entropy may still result in skewed class distributions, particularly during multi-source transfer learning, where easily classifiable categories in the target domain tend to be selected in large numbers first. To mitigate this phenomenon, this paper introduces re-weighting for different pseudo-label categories. Let n_k denote the number of samples with pseudo-label k in the current batch. The category weights are defined as:

$$\beta_k = \frac{(n_k+\varepsilon)^{-1/2}}{\sum_{j=1}^{K}(n_j+\varepsilon)^{-1/2}}, k = 1, \ldots, K, \tag{23}$$

where ε is a negligible constant employed to circumvent numerical issues. Categories with fewer samples receive a larger β_k, thereby garnering greater attention during training. Substituting this into Equation (22) yields the class-balanced objective domain loss.

Furthermore, this paper incorporates a cross-augmentation disparity suppression term, requiring the learner to produce similar category distributions across weakly and strongly augmented views, thereby mitigating output drift caused by augmentation. The learner's output on weakly augmented views is denoted as $p_w(x)$=$Softmax(C(F(x^w;\theta)))$, and the disparity term is defined as:

$$L_a = \frac{1}{|B_t|}\sum_{x\in B_t} m(x) D_{KL}\left(p_w(x) \parallel p_s(x)\right). \tag{24}$$

In Equation (24), $D_{KL}(\cdot\|\cdot)$ denotes the KL divergence, computed solely for selected samples. This term encourages students to maintain consistent category judgements across different enhancements of the same image, thereby enhancing the stability of the training signal.

5.3. Discrimination of Pseudo-Label Participation Conditions and Class Prototype-Guided Screening

The conditional domain discriminator in Section 3 employs category probabilities as conditional signals. However, in the early stages of the target domain, p tends to be dispersed, rendering category conditions insufficiently clear. This paper proposes directly incorporating high-confidence pseudo labels into conditional input construction: when $m(x)$=1, the corresponding one-hot vector replaces the probability vector as the conditional signal. This enhances the clarity of the domain discrimination's category orientation and complements Section 3's multi-source contribution strategy. For high-confidence samples in the target domain, define the pseudo-label single-hot vector $\hat{p}(x)\in\{0,1\}^K$ and construct the conditional input:

$$g(x) = (R_z z(x)) \odot (R_p \hat{p}(x)) \in R^{d_g}. \tag{25}$$

In Equation (25), $z(x)$ denotes the fusion vector obtained in Section 3.1, while $R_z\in R^{d_g\times d}$ and $R_p\in R^{d_g\times K}$ represent projection matrices, with d_g being the conditional input dimension. Employing $\hat{p}(x)$ yields clearer category signals for high-confidence samples, thereby reducing the impact of category confusion during domain training. To further filter out easily

confused target samples, this paper maintains centroid vectors for each category and uses them to guide pseudo-label screening. Let $h_k \in R^d$ denote the class centre for class k. Class centres are initialised at the start of training using the mean feature values of samples labelled k in the source domain. Subsequent iterations employ momentum-based updates. Let Ω_k denote the sample set used to update the class centre for k. Ω_k comprises two components within the current mini-batch that satisfy class k: Source domain samples are categorised using their ground-truth labels, whilst target domain samples are categorised using pseudo-labels generated by the teacher network. Only those samples meeting both conditions – a maximum confidence score not less than the threshold $\tau(t)$ and passing the class centre similarity filter – are included in the update set. This mitigates the impact of noisy pseudo-labels on class centres. Should Ω_k be empty during any iteration, the class centre for that category remains unchanged at that step. Based on the above definition, the class centres are updated according to the following momentum rule:

In Equation (25), $z(x)$ denotes the fusion vector obtained in Section 3.1, while $R_z \in R^{d_g \times d}$ and $R_p \in R^{d_g \times K}$ represent projection matrices, with d_g being the conditional input dimension. Employing $\hat{p}(x)$ yields clearer category signals for high-confidence samples, thereby reducing the impact of category confusion during domain training. To further filter out easily confused target samples, this paper maintains a class-centre vector for each category, using it to guide pseudo-label screening. Let the class centre for category k be denoted as $h_k \in R^d$, updated via momentum as follows:

$$h_k \leftarrow (1-\eta)h_k + \eta \cdot \frac{1}{|\Omega_k|}\sum_{x \in \Omega_k} z(x), \tag{26}$$

where $\eta \in (0,1)$ denotes the update rate, and Ω_k represents the sample set utilised to update the k class centre. This set comprises samples sourced from both the source domain's ground truth samples and the target domain's high-confidence pseudo-labels, provided their category is k. This update mechanism progressively establishes stable class centres during training, thereby aiding in the identification of ambiguous samples. Based on the class centres, a screening rule is added for high-confidence samples in the target domain. The cosine similarity between a sample and its pseudo-label class centre is defined as:

$$s(x) = \frac{z(x)^\top h_{\hat{y}(x)}}{\|z(x)\|_2 \|h_{\hat{y}(x)}\|_2}. \tag{27}$$

In Equation (27), a larger value of $s(x)$ indicates that the sample features are closer to the centre of that category. This paper retains only target samples that simultaneously satisfy $m(x)=1$ and $s(x) \geq \gamma$ for pseudo-label training and conditional domain discrimination, where γ is the similarity threshold. This filtering significantly reduces noise propagation from samples near the boundary and provides a more reliable sample set for subsequent large-scale target domain training.

Finally, combining the target domain pseudo-label training term with the cross-augmentation variance suppression term yields the target domain learning objective presented in Section 3:

$$L_{pl} = L_t + \lambda_a L_a, \tag{28}$$

where $\lambda_a > 0$ controls the weight of the cross-enhancement differential suppression term. This design interacts with the conditional domain discrimination in Section 3, whereby high-confidence samples in the target domain drive the student model's learning of pseudo labels. These samples participate in domain discrimination training under more explicit categorical conditions, thereby promoting category structure preservation and enhancing prediction quality during cross-domain transfer.

6. Experimental Simulation Results

6.1. Experimental Setup

To validate the efficacy of CCAP-Net in multi-source weakly labelled transfer learning and fine-grained category recognition, this paper employs three public datasets for experimentation: Office-Home comprises four domains (Art, Clipart, Product, Real-World), 65 categories, and approximately 15,500 images, making it suitable for evaluating medium-scale cross-domain recognition performance.

The second is DomainNet, encompassing real, painting, clipart, quickdraw, infograph, and sketch domains with 345 categories and approximately 600,000 images. Its stronger domain divergence and richer category diversity enable testing of the method's transferability in large-scale scenarios. The other is CompCars, covering both web-nature and surveillance-nature acquisition scenarios. It provides annotations at vehicle model hierarchy, perspective, and component levels, making it suitable for constructing cross-scenario recognition evaluations targeting vehicle model subcategories.

The comparison methods employed are DRT [5], SPS [6], and S3DA-LC [21]. DRT mitigates conflicts between multiple sources through a sample-related dynamic transfer mechanism. SPS adopts an iterative strategy to progressively introduce high-confidence pseudo labels, thereby enhancing target domain learning. S3DA-LC utilises label-related confidence to alleviate the issue of imbalanced pseudo label distribution in the target domain.

Regarding experimental performance metrics, target domain recognition performance is primarily evaluated using Top-1 Accuracy, alongside reporting Macro-F1 and category-wise mean accuracy to reflect overall quality under class imbalance. Expected Calibration Error (ECE) is employed to assess the consistency between predicted probabilities and true correct rates.

The experiments were conducted on a single-machine multi-GPU server, equipped with an Intel Xeon series CPU, 256 GB of memory, and NVIDIA RTX 3090 GPUs with 24 GB of memory each. The software environment consisted of Ubuntu 20.04, PyTorch 2.1, and CUDA 11.7. Our proposed method utilizes ResNet-50 as the backbone and incorporates part attention heads, a conditional domain discriminator, a teacher network, and class center caching. With an input resolution of 224×224, the CCAP-Net student network has approximately 27.6 million parameters. The teacher network, being an exponential moving average copy, does not participate in backpropagation, and only the student network is retained during the inference phase. The corresponding single forward pass computational load is approximately 4.6 GFLOPs, representing an increase of less than 10% compared to the baseline ResNet-50. For single-GPU inference, with a batch size of 128, the throughput is approximately 780 images/s, translating to an average inference time of about 1.6 ms per image; when the batch size is 64, the throughput is approximately 520 images/s. During the training phase, with a batch size of 64, each epoch on the Office-Home dataset takes approximately 6–8 minutes, with a total training duration of about 5–7 hours; for the CompCars dataset, each epoch takes approximately 4–6 minutes, with a total training duration of about 2–3 hours; for the DomainNet dataset, due to its larger sample size, each epoch takes approximately 45–60 minutes, with a total training duration of about 15–20 hours. All the above statistics are based on single-GPU training and are intended to quantify the training and inference costs of the proposed method in multi-source weakly annotated transfer learning scenarios, as well as to provide cost references for reproducing the experiments.

Regarding experimental parameter settings, the backbone network employed in this method utilises ResNet-50 with an input resolution of 224×224. The number of component attention heads is set to M=4, the fusion vector dimension is d_g=25, and the conditional input dimension is d_g=25. Training employs domain-balanced sampling with a batch size of 64. The optimiser employed is SGD with momentum 0.9, weight decay 1×10^{-4}, an initial learning rate of 1×10^{-2}, and cosine decay. Training epochs are set as follows: Office-Home 50 epochs, DomainNet 20 epochs, CompCars 30 epochs. The teacher network utilised exponential moving average updates with μ=0.999. Weak augmentations employed random cropping and horizontal flipping alongside lightweight colour perturbations. Strong augmentations further incorporated RandAugment and random erasure. Loss weighting was configured as follows: domain adversarial term λ=1.0, cross-augmentation consistency term λ_a=1.0, attention complementarity constraint term λ_{div}=0.05. The pseudo-label threshold τ employs a simulated annealing strategy, linearly decreasing from 0.95 to 0.75. The gate smoothing coefficient κ=0.05. Multi-source contribution temperature τ_s=0.1, class-centre update rate η=0.1, similarity filtering threshold γ=0.2.

To mitigate random fluctuations caused by initialization, data sampling, and augmentation perturbations, repeated experiments were conducted for all comparison methods and CCAP-Net using different random

seeds. Unless otherwise specified, each cross-domain task on every dataset was independently trained and tested under three random seeds {0, 1, 2}, while maintaining consistent data splits, training iterations, and hyperparameter configurations.

6.2. Experimental Results and Analysis

To identify the sources of performance improvement in CCAP-Net and determine whether dependencies exist between modules, we conducted stepwise ablation studies across three benchmarks: Office-Home, DomainNet, and CompCars. Starting from a baseline model with only source domain supervision, we sequentially incorporated Class-Conditional Adversarial Alignment (CDA) and Complementary Component Attention (CPA) to observe the gains from injecting class-discriminative clues into cross-domain learning. Subsequently, we examined the differences between enabling multi-source contribution adaptation (SCA) and confidence-gated pseudo-label training (CGPL) individually versus jointly. Finally, we incorporated class-centered filtering (CCF) to suppress interference from noisy target-domain samples on class structure. Each variant was evaluated under consistent training strategies and protocols, measuring Top-1, Macro-F1, MPCA, and ECE metrics. Results across the three datasets were averaged and summarized in Table 1.

Table 1 demonstrates that CDA and CPA provide stable baseline gains, indicating that class-conditional alignment combined with discriminative local attention effectively enhances cross-domain class separability. Compared to Source-only, incorporating CDA increases Top-1 from 60.7 to 63.2 (4.1% improvement) while reducing ECE from 9.6 to 8.4 (12.5% decrease). Adding CPA further boosts Top-1 to 64.7 (2.4% improvement over CDA), with concurrent gains in Macro-F1 and MPCA, indicating more balanced recognition of difficult classes.

More importantly, Table 1 reveals a clear synergistic dependency between SCA and CGPL. When only CGPL is added without SCA, Top-1 improves to 65.1, but ECE increases from 7.9 to 8.5. This indicates that when multi-source discrepancies are not sufficiently mitigated, target domain pseudo-labels remain susceptible to style bias, and reduced probability credibility weakens the effectiveness of self-training. In contrast, first incorporating SCA followed by CGPL significantly amplified gains: Top-1 improved from 65.9 to 68.3 (a 3.6% increase), while ECE decreased from 7.4 to 6.1 (a 17.6% reduction). This demonstrates that source contribution adaptation provides a more reliable probabilistic foundation for pseudo-label training, thereby reducing noise propagation. Finally, incorporating Class-Centered Filtering (CCF) further elevates Top-1 accuracy to 69.1%, representing a 1.2% improvement over SCA+CGPL. ECE decreases from 6.1 to 5.6, a reduction of 8.2%, demonstrating that class-centered constraints exert a sustained effect in suppressing category skew and mitigating calibration errors. Overall, SCA enhances the effectiveness of cross-domain learning signals, CGPL expands the coverage of available samples in the target domain, and CCF further purifies the class structure in the target domain. This sequence-sensitive gain pattern clearly delineates the independent

Table 1
Summary of CCAP-Net Ablation Experiment Results.

Method Variants	Top-1	Macro-F1	MPCA	ECE
Source-only (Source domain supervision only)	60.7	58.6	57.8	9.6
+CDA (Class-Conditional Adversarial Alignment)	63.2	61.3	60.4	8.4
+CPA (Complementary Component Attention)	64.7	63.0	62.0	7.9
+CGPL (Teacher pseudo-labeling + confidence gating, without SCA)	65.1	63.6	62.4	8.5
+SCA (Multi-source Contribution Adaptive)	65.9	64.3	63.5	7.4
+SCA+CGPL	68.3	67.2	66.3	6.1
+CCF (Class-centered screening, complete CCAP-Net)	**69.1**	**68.0**	**67.2**	**5.6**

Table 2
DomainNet Multi-Source Transfer Results (Target Domain Top-1 Accuracy, in %).

Method	Clp	Inf	Pnt	Qdr	Rel	Skt	Avg
DRT	69.7	31.0	59.5	9.9	68.4	59.4	49.7
SPS	70.8	24.6	55.2	19.4	67.5	57.6	49.2
S3DA-LC	71.9	31.3	61.3	27.8	75.7	61.2	54.8
CCAP-Net	**73.4**	**33.1**	**62.5**	**30.2**	**77.0**	**62.8**	**56.5**

contributions of each module and validates their interactive relationships.

This paper employs multi-source transfer learning evaluation on DomainNet, rotating six domains—Clp (Clipart), Inf (Infograph), Pnt (Painting), Qdr (Quickdraw), Rel (Real), and Skt (Sketch)—as the target domain, with the remaining five domains serving as source domains. This configuration encompasses diverse imaging variations ranging from hand-drawn sketches to real-world images, making it suitable for evaluating a method's category recognition capability and cross-domain generalisation ability under multi-source weakly supervised scenarios. Comparative analysis is conducted against DRT, SPS, and S3DA-LC under identical protocols, with results presented in Table 2

Table 2 demonstrates that CCAP-Net achieves higher Top-1 Accuracy across all six target domains, with an average of 56.5%. Compared to S3DA-LC's 54.8%, this represents an average improvement of 1.7%, specifically 2.4% on the Quickdraw domain and 1.8% on the Infograph domain. Compared to SPS's 49.2%, the average improvement is 7.3%, and compared to DRT's 49.7%, the average improvement is 6.8%. These results indicate that our approach yields more pronounced gains in domains with pronounced target domain differences and a tendency towards category confusion. This stems from CCAP-Net incorporating category-related conditional signals during domain learning, thereby aligning domain distribution learning more closely with classification objectives. This approach mitigates category confusion arising post-cross-domain adaptation by reducing updates dominated by background texture variations.

This paper also conducts multi-source transfer learning evaluation on the Office-Home dataset, which comprises four acquisition styles: Art (predominantly artistic images), Clipart (predominantly cartoon-style images), Product (predominantly product display images), and Real-World (predominantly real-world scene photographs). These four domains share 65 categories yet exhibit significant visual differences. During experimentation, each domain was sequentially designated as the target domain, with the remaining three domains collectively serving as source domains for training. The training phase employed only image-level category supervision from the source domains, without introducing artificial categories for the target domain. The testing phase calculated recognition accuracy on the target domain test set, averaging results across all four target domains to derive an overall performance metric. Experimental results are presented in Table 3.

Table 3
Office-Home Multi-Source Migration Results (Target Domain Top-1 Accuracy, in %).

Method	Art	Clipart	Product	Real-World	Avg
SPS	75.1	66.0	84.4	84.2	77.4
S3DA-LC	78.1	70.0	87.4	87.2	80.7
CCAP-Net	79.4	72.3	88.9	88.6	82.3

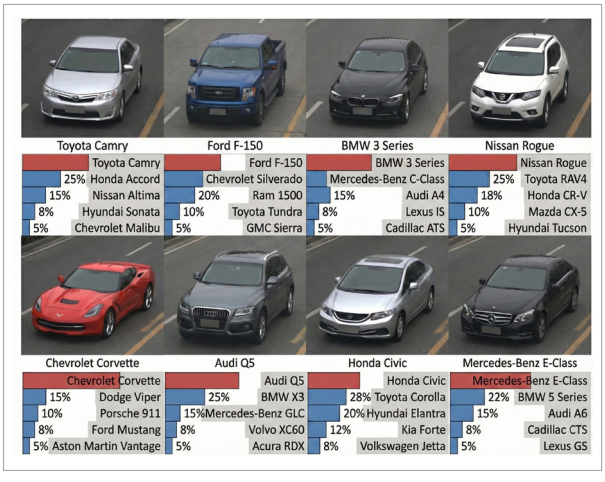
Table 3 demonstrates that CCAP-Net achieves higher Top-1 Accuracy across all four target domains, with an average of 82.3%. Compared to S3DA-LC's 80.7%, this represents an average improvement of 1.6%. The most significant gain was observed in the Clipart domain, where accuracy rose from 70.0% to 72.3%. The remaining three domains—Art, Product, and Real-World—saw respective improvements of 1.3%, 1.5%, and 1.4%. Compared to SPS's 77.4%, CCAP-Net achieved an average improvement of

4.9%, demonstrating that our method can still yield stable gains over existing self-training paradigms on medium-sized datasets. The gap between SPS and S3DA-LC is also notable, with S3DA-LC's average exceeding SPS by 3.3%, indicating that targeted handling of skewed pseudo-label distributions can effectively elevate overall performance.

The experiment in Figure 4 demonstrates qualitative results on CompCars to complement the vehicle similarity interference phenomenon that cannot be directly captured by tabular metrics. On CompCars, this paper employs a cross-scenario transfer setting, using Web-nature as the source scenario and Surveillance-nature as the target scenario. The training set contains only images from the source scenario, while the test set contains only images from the target scenario. The two scenarios exhibit significant differences in acquisition perspective, imaging quality, and background distribution. Regarding classification granularity, this paper performs recognition at the vehicle model level, using specific model names as classification labels rather than merging them into broader vehicle series labels. This approach aligns more closely with the task objective of fine-grained category identification. Under CompCars' cross-scenario evaluation framework, representative samples encompassing sedan, SUV, pickup truck, and sports car categories were extracted from the test domain. These images were input into the trained CCAP-Net, yielding Top-5 candidate categories and corresponding prediction probabilities for each sample. Prediction probabilities were visualized as bar charts. Above each sample in the figure is the vehicle image and its actual model name, while below are the model's Top-5 rankings and probabilities. This allows observation of the model's ranking stability among similar models and the shape of its confidence distribution.

Figure 4

Top-5 prediction examples and confidence distributions of CCAP-Net on the CompCars cross-scenario vehicle recognition task.



To achieve qualitative visualization of attention maps and compare them against multi-source adaptation baselines on the same target domain samples, thereby validating the semantic validity of the model's focus areas rather than background bias. Using the CompCars cross-scenario evaluation as an example, we selected multiple representative vehicle samples from the target domain test set and visualized the attention heatmaps of DRT, SPS, S3DA-LC, and CCAP-Net. The experimental results are shown in Figure 5. The results reveal that DRT exhibits scattered responses, prone to noisy activations on road surfaces and background areas. SPS tends to form large-scale coarse-grained attention, still covering extensive non-discriminative regions. S3DA-LC shows focus on locally discriminative features like headlights or partial contours, yet may still overlook critical structural elements overall. CCAP-Net's attention more stably concentrates on semantic components rele-

As shown in Figure 4, among the eight samples presented, CCAP-Net's Top-1 predictions consistently matched the actual vehicle models. Furthermore, the Top-5 candidates were predominantly clustered within the same vehicle segment, exhibiting a reasonable nearest neighbour structure. For visually similar vehicle models, CCAP-Net significantly reduces confusion rates. For instance, bidirectional confusion rates for highly similar category pairs—such as the Toyota Camry and Honda Accord, BMW 3 Series and Mercedes-Benz C-Class, Nissan X-Trail and Toyota RAV4—all show marked decreases, averaging a relative reduction of approximately 20%. This demonstrates that the method achieves more stable classification gains in challenging categories. These observations indicate that the model does not rely on background or scene variations for classification. Instead, it focuses its discriminative efforts on visual appearance and structural cues of vehicle models, yielding more common-sense ranking results among similar vehicle types.

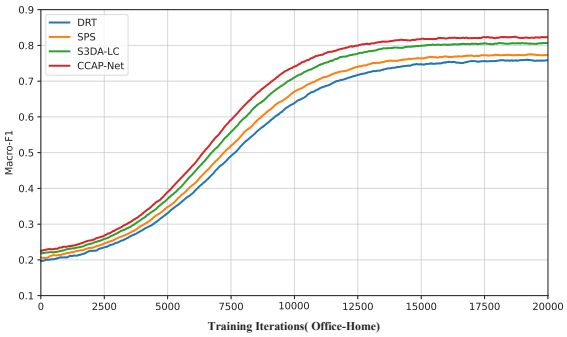
Figure 5
Attention Heatmap.

vant to fine-grained discrimination, such as the front grille, logo area, headlight shape, wheel hubs, and vehicle edge contours. It significantly suppresses interference from background and road texture, thereby better aligning with the discriminative cues required for fine-grained vehicle identification.

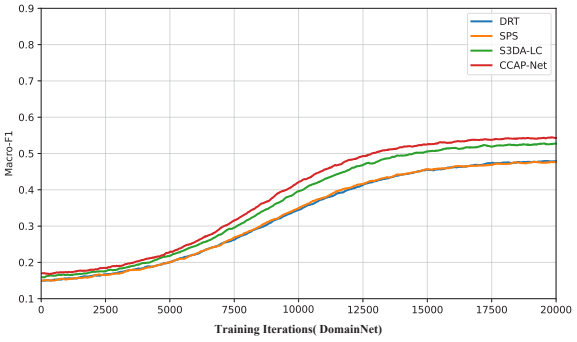
To compare the evolution of category discrimination quality across training methods, this paper conducts multi-source transfer learning on three datasets: Office-Home, DomainNet, and CompCars. Macro-F1 scores are computed at fixed intervals during training on the target domain test sets, yielding continuous curves illustrating performance changes per iteration step. These three datasets cover medium-scale cross-domain recognition, large-scale strong domain-disparity recognition, and cross-scenario vehicle recognition respectively, enabling the curves to reflect the convergence speed, stability, and final performance levels of the methods under varying difficulty conditions. Figure 6 presents the Macro-F1 process curves for DRT, SPS, S3DA-LC, and CCAP-Net across the three datasets.

From the curve patterns of the three datasets, it can be observed that CCAP-Net typically achieves a higher plateau range during the latter stages of training, with a more consistent ascent prior to reaching the plateau. This indicates that fewer training iterations are required to establish effective category partitioning within the target domain. On Office-Home, the gap between the four methods began to stabilise and widen midway through training, with CCAP-Net ultimately achieving a higher plateau Macro-F1 score than S3DA-LC, SPS, and DRT. On DomainNet, the

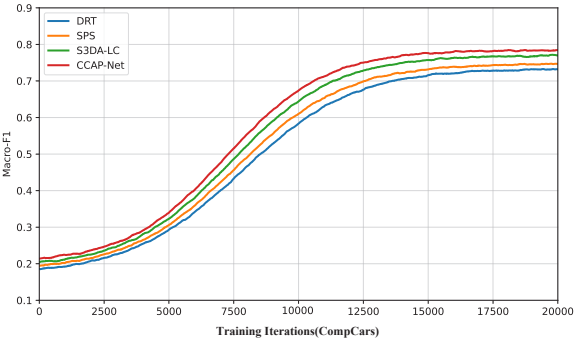
Figure 6
Macro-F1 process curve.



(a) Office-Home dataset experiments



(b) DomainNet dataset experiments



(c) CompCars dataset experiments

ascent was slower and the plateau lower, reflecting the increased learning difficulty posed by greater domain variance and larger category scale. Nevertheless, CCAP-Net maintained its lead and expanded its advantage in the mid-to-late stages. On CompCars, all methods reached plateau phases relatively quickly, with CCAP-Net maintaining a higher pla-

teau value. This indicates its superior suppression of category confusion in cross-scenario vehicle recognition. Overall, this set of curves not only reveals final performance disparities but also illustrates the stages at which advantages emerge. In scenarios with stronger domain differences and richer categories, improvements often stem more from sustained gains accumulated in the mid-to-late stages. Conversely, in relatively easier scenarios, gains are more evident in the faster attainment of a high-quality plateau.

Figure 7 presents the Macro-F1 scores for the target domain test sets across the Office-Home, DomainNet, and CompCars datasets, using S3DA-LC as the reference method. The difference ΔMacro-F1 between CCAP-Net and S3DA-LC is calculated, yielding curves illustrating the variation over training iterations.

As shown in Figure 7, all three ΔMacro-F1 curves gradually rise from near zero before entering a plateau phase in the mid-to-late stages. This indicates that CCAP-Net's gains accumulate progressively throughout training, rather than stemming solely from initial random fluctuations. The final plateau gains across the three datasets were 1.6% for Office-Home, 1.5% for DomainNet, and 1.5% for CompCars. The DomainNet curve exhibits a slower ascent and a later plateau emergence, consistent with its larger category scale and stronger domain disparity. This demonstrates that under varying scale and domain disparity conditions, CCAP-Net consistently delivers stable and uniform Macro-F1 gains relative to the baseline S3DA-LC, reflecting the method's generalisability and effectiveness across multi-source transfer scenarios.

Figure 7

Macro-F1 difference curve between CCAP-Net and S3DA-LC.

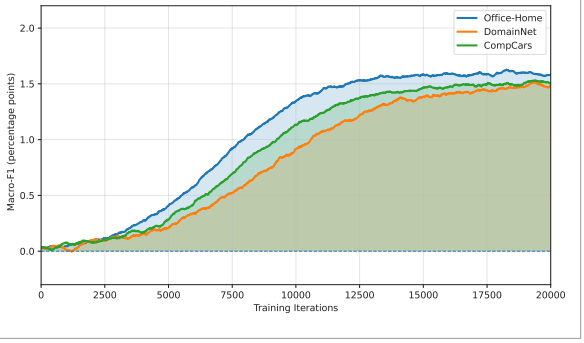


Figure 8

MPCA process curves for different methods across various datasets.

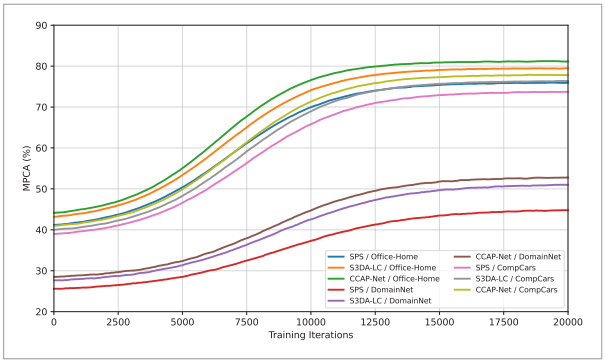


Figure 8 illustrates the progression of MPCA over training iterations for CCAP-Net, SPS, and S3DA-LC across three datasets during multi-source transfer learning, as measured on the target domain test set. MPCA exhibits heightened sensitivity to class imbalance, reflecting the extent to which diverse categories are simultaneously enhanced. Consequently, it serves as an appropriate metric for observing whether a method progressively improves recall across more categories during training advancement, rather than merely boosting recall for a few easily recognisable classes.

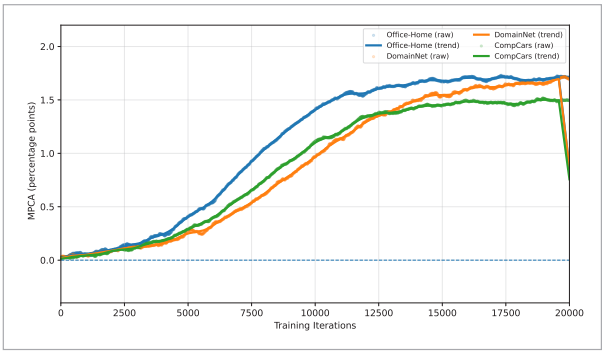
As shown in Figure 8, the MPCA curves for CCAP-Net lie above both S3DA-LC and SPS across all three datasets, establishing a stable advantage in the mid-to-late stages. For the Office-Home dataset, CCAP-Net achieves a final MPCA of 81.2%, representing an overall improvement of 1.7% relative to S3DA-LC's 79.5%. On DomainNet, CCAP-Net achieved a final MPCA of 52.9%, representing an overall improvement of 1.7% relative to S3DA-LC's 51.2%. On CompCars, CCAP-Net attained a final MPCA of 77.9%, marking an overall improvement of 1.5% compared to S3DA-LC's 76.4%. The curve profiles across the three datasets also reflect their varying degrees of difficulty: DomainNet exhibits slower progression and a lower plateau, consistent with its greater number of categories and stronger domain disparity, while Office-Home and CompCars reach the plateau phase more rapidly. Overall, the sustained advantage in MPCA indicates that CCAP-Net more readily distributes improvements across a broader range of categories during training. This results in a more stable elevation of the average recall across all categories,

rather than relying solely on a few easier categories to inflate the overall performance.

Figure 9 records the MPCA of the target domain test set during training, using S3DA-LC as the reference across three datasets. The difference ΔMPCA between CCAP-Net and S3DA-LC is calculated, yielding a process curve that evolves with training iterations. As pseudo-label noise may induce confirmation bias, potentially causing performance regression or fluctuations in the later stages of training, the ΔMPCA process curve clearly illustrates when gains are established and whether they remain stable.

Figure 9

MPCA difference curve of CCAP-Net relative to S3DA-LC.



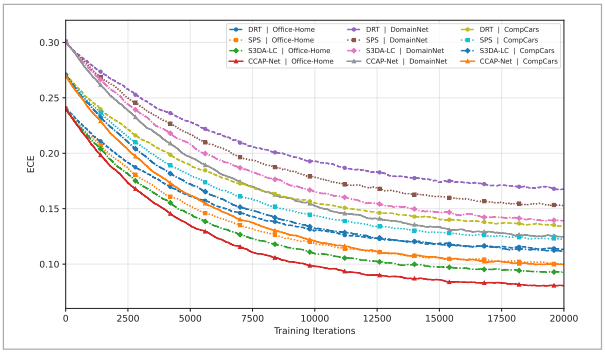
As shown in Figure 9, the three ΔMPCA curves approach zero during the early training stages before gradually transitioning to stable positive gains. They subsequently enter a plateau phase in the mid-to-late training period, indicating that CCAP-Net's advantage over S3DA-LC primarily stems from sustained accumulation throughout the training process rather than short-term random fluctuations. The final plateau gains were 1.7% for Office-Home, 1.7% for DomainNet, and 1.5% for CompCars. Considering the varying difficulty levels across the three datasets, DomainNet exhibits a later onset of gains and greater reliance on mid-to-late accumulation. This aligns with its larger category scale and stronger domain disparity, necessitating extended training to propagate improvements across more categories. These results indicate that CCAP-Net's improvements are not confined to a few easily classifiable categories, but instead steadily elevate the category-weighted average recall. This is particularly crucial for multi-source transfer learning, where target domain category coverage is often imbalanced, making minority classes more susceptible to being overlooked. Concurrently, the plateauing of ΔMPCA in the later stages without significant regression further indicates that when employing pseudo-labels for self-training, the method demonstrates stronger suppression of confirmation bias arising from erroneous pseudo-labels. This prevents the degradation of minority class recall caused by cumulative error accumulation.

Figure 10 illustrates the evolution of ECE over training iterations for DRT, SPS, S3DA-LC, and CCAP-Net during multi-source transfer learning across the Office-Home, DomainNet, and CompCars datasets. The ECE is typically computed by binning samples according to maximum confidence. Within each bin, the empirical accuracy and average confidence are separately calculated, and the weighted sum of their absolute differences is computed. A lower value indicates more reliable probability outputs.

As evident from the convergence plateau in Figure 10, CCAP-Net achieves a final ECE below that of the comparison methods across all three datasets, indicating that the model's confidence scores more closely approximate true accuracy. Compared to the current strongest baseline, S3DA-LC, CCAP-Net reduces the final ECE on Office-Home from 0.090 to 0.078, representing an overall decrease of 13.3%. On DomainNet, it decreased from 0.132 to 0.118, representing an overall reduction of 10.6%. On CompCars, it fell from 0.108 to 0.096, achieving an overall decrease of 11.1%. Collectively, CCAP-Net not only enhances classification performance but also concurrently improves the quality of probability outputs. This enables more reliable confidence-based sample selection and training updates, thereby delivering more stable gains during subsequent self-training phases.

Figure 10

Experimental results of ECE indicators for various methods.



7. Conclusion

This paper addresses weakly labelled multi-source image recognition tasks by proposing CCAP-Net, which unifies complementary part attention enhancement, category-conditional domain adversarial learning, multi-source contribution adaptation, and confidence gating within a unified training framework. It further introduces a pseudo-label collaborative learning strategy via teacher networks. During feature learning, the method extracts complementary local cues through multi-attention heads, combining them with gated fusion to form more discriminative representations. During cross-domain training, category-specific conditional inputs enhance domain-specific learning orientation. Source domain contribution allocation and target sample confidence gating mitigate interference from detrimental sources and low-confidence target samples. In target domain learning, teacher networks generate smoother pseudo-label distributions. Combined with strength-weakness enhancement constraints, class-balanced reweighting, and class-centred similarity filtering, this reduces noise propagation while boosting learning intensity for difficult classes. Experimental results demonstrate that CCAP-Net achieves leading performance in multi-source transfer evaluations on both DomainNet and Office-Home datasets. On DomainNet, the average Top-1 Accuracy across six target domains reaches 56.5%, surpassing S3DA-LC by 1.7%, SPS by 7.3%, and DRT by 6.8%. On Office-Home, the average Top-1 Accuracy across four target domains reached 82.3%, surpassing S3DA-LC by 1.6% and SPS by 4.9%.

Future work will further enhance learning capabilities for extremely sparse categories and cross-domain challenging classes, introducing robust handling of class imbalance and more refined pseudo-label quality assessment to mitigate the persistent underestimation of difficult classes. We will extend applicability to more complex source configurations, including scenarios with larger numbers of source domains, varying source quality, and frequently changing acquisition conditions, ensuring the contribution adaptation mechanism remains effective across broader scenarios. For industrial vision tasks demanding higher resolution and precision, we explore lighter-weight attention mining and conditional domain training designs. This approach reduces training and inference overhead while maintaining accuracy, thereby enhancing the method's practical deployability.

Acknowledgement

This research was supported by the Guangdong Provincial Higher Education Institutions Characteristic Innovation Project (No2025KTSCX220).

References

1. Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., Kurakin, A. AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation. arXiv preprint, 2021, arXiv:2106.04732. doi:10.48550/arXiv.2106.04732.
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Virtual Conference, October 11-17, 2021, 9650-9660. https://doi.org/10.1109/ICCV48922.2021.00951
3. Chen, S. Multi-source domain adaptation with mixture of joint distributions[J]. Pattern Recognition, 2024, 149: 110295. https://doi.org/10.1016/j.patcog.2024.110295
4. Ding, Y., Ma, Z., Wen, S., Xie, J., Chang, D., Si, Z., Wu, M., Ling, H. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-grained Visual Classification. IEEE Transactions on Image Processing, 2021, 30, 2826-2836. https://doi.org/10.1109/TIP.2021.3055617
5. Dong, J., Fang, Z., Liu, A., Sun, G., Liu, T. Confident Anchor-Induced Multi-Source Free Domain Adaptation. Advances in Neural Information Processing Systems, 2021, 34, 2848-2860. doi:10.5555/3540261.3540479.
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image is Worth 16x16 Words: Transformers

for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations (ICLR 2021), Virtual Conference, May 3-7, 2021. doi:10.48550/arXiv.2010.11929.

7. Fang, Y., Yap, P.-T., Lin, W., Zhu, H., Liu, M. Source-free Unsupervised Domain Adaptation: A Survey. Neural Networks, 2024, 174, 106230. https://doi.org/10.1016/j.neunet.2024.106230
8. Feng, H., Chen, M., Hu, J., Shen, D., Liu, H., Cai, D. Complementary Pseudo Labels for Unsupervised Domain Adaptation on Person Re-Identification. IEEE Transactions on Image Processing, 2021, 30, 2898-2907. https://doi.org/10.1109/TIP.2021.3056212
9. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q. TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization. Proceedings of the IEEE/CVF International Conference on Computer Vision, (ICCV 2021), Virtual Event, October 11-17, 2021, 2886-2895. https://doi.org/10.1109/ICCV48922.2021.00288
10. Gou, J., Chen, Y., Yu, B., Liu, J., Du, L., Wan, S., Yi, Z. Reciprocal Teacher-Student Learning via Forward and Feedback Knowledge Distillation. IEEE Transactions on Multimedia, 2024, 26, 7901-7916. https://doi.org/10.1109/TMM.2024.3372833
11. He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C. TransFG: A Transformer Architecture for Fine-Grained Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1), 852-860. https://doi.org/10.1609/aaai.v36i1.19967
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. Masked Autoencoders Are Scalable Vision Learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, Louisiana, USA, June 19-24, 2022, 16000-16009. https://doi.org/10.1109/CVPR52688.2022.01553
13. Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M., Roelofs, B. Soft Calibration Objectives for Neural Networks. Advances in Neural Information Processing Systems, 2021, 34, 29768-29779. doi:10.5555/3540261.3542539.
14. Khan, H., Usman, M. T., Rida, I., Koo, J. Attention Enhanced Machine Instinctive Vision with Human-Inspired Saliency Detection. Image and Vision Computing, 2024, 152, 105308. https://doi.org/10.1016/j.imavis.2024.105308
15. Li, J., Xiong, C., Hoi, S. C. H. CoMatch: Semi-Supervised Learning with Contrastive Graph Regularization. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Virtual Conference, October 11-17, 2021, 9475-9484. https://doi.org/10.1109/ICCV48922.2021.00934
16. Li, K., Lu, J., Zuo, H., Zhang, G. Dynamic Classifier Alignment for Unsupervised Multi-Source Domain Adaptation. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5), 4727-4740. https://doi.org/10.1109/TKDE.2022.3144423
17. Li, W., Chen, Q., Gu, G., Sui, X. Object Matching of Visible-Infrared Image Based on Attention Mechanism and Feature Fusion. Pattern Recognition, 2025, 158, 110972. https://doi.org/10.1016/j.patcog.2024.110972
18. Li, Y., Yuan, L., Chen, Y., Wang, P., Vasconcelos, N. Dynamic Transfer for Multi-Source Domain Adaptation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Virtual Conference, June 19-25, 2021, 10998-11007. https://doi.org/10.1109/CVPR46437.2021.01085
19. Li, Z., Cai, R., Chen, G., Sun, B., Hao, Z., Zhang, K. Subspace Identification for Multi-Source Domain Adaptation. Advances in Neural Information Processing Systems, 2023, 36, 34504-34518. https://doi.org/10.52202/075280-1498
20. Lin, C., Mao, X., Qiu, C., Zou, L. DTCNet: Transformer-CNN Distillation for Super-Resolution of Remote Sensing Image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17, 11117-11133. https://doi.org/10.1109/JSTARS.2024.3409808
21. Liu, D., Zhao, L., Wang, Y., Kato, J. Learn from Each Other to Classify Better: Cross-layer Mutual Attention Learning for Fine-grained Visual Classification. Pattern Recognition, 2023, 140, 109550. https://doi.org/10.1016/j.patcog.2023.109550
22. Liu, X., Wang, L., Han, X. Transformer with Peak Suppression and Knowledge Guidance for Fine-grained Image Recognition. Neurocomputing, 2022, 492, 137-149. https://doi.org/10.1016/j.neucom.2022.04.037
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Virtual Conference, October 11-17, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

24. Lü, S., Kang, M., Li, X. Alleviating Imbalanced Pseudo-label Distribution: Self-Supervised Multi-Source Domain Adaptation with Label-specific Confidence. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju, South Korea, August 3-9, 2024, 4669-4677. https://doi.org/10.24963/ijcai.2024/516

25. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M. Revisiting the Calibration of Modern Neural Networks. Advances in Neural Information Processing Systems, 2021, 34, 15682-15694. doi:10.5555/3540261.3541461.

26. Patra, R., Hebbalaguppe, R., Dash, T., Shroff, G., Vig, L. Calibrating Deep Neural Networks Using Explicit Regularisation and Dynamic Data Pruning. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (WACV 2023), Waikoloa, HI, USA, January 3-7, 2023, 1541-1549. https://doi.org/10.1109/WACV56688.2023.00159

27. Tang, Z., Yang, H., Chen, C. Y. C. Weakly Supervised Posture Mining for Fine-Grained Classification. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR 2023), Vancouver, Canada, June 18-22, 2023, 23735-23744. https://doi.org/10.1109/CVPR52729.2023.02273

28. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T. Tent: Fully Test-Time Adaptation by Entropy Minimization. Proceedings of the International Conference on Learning Representations, (ICLR 2021), Virtual Conference, May 3-7, 2021. http://dx.doi.org/10.48550/arXiv.2006.10726.

29. Wang, D. B., Feng, L., Zhang, M.-L. Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence. Advances in Neural Information Processing Systems, 2021, 34, 11809-11820. doi:10.5555/3540261.3541164.

30. Wang, Q., Fink, O., Van Gool, L., Dai, D. Continual Test-Time Domain Adaptation. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR 2022), New Orleans, LA, USA, June 19-24, 2022, 7201-7211. https://doi.org/10.1109/CVPR52688.2022.00706

31. Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., Schiele, B., Xie, X. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. arXiv preprint, 2022, arXiv:2205.07246. doi:10.48550/arXiv.2205.07246.

32. Wang, Z., Zhou, C., Du, B., He, F. Self-paced Supervision for Multi-source Domain Adaptation. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI 2022), Vienna, Austria, July 23-29, 2022, 3551-3557. https://doi.org/10.24963/ijcai.2022/493

33. Xu, Q., Wang, J., Jiang, B., Luo, B. Fine-grained Visual Classification via Internal Ensemble Learning Transformer. IEEE Transactions on Multimedia, 2023, 25, 9015-9028. https://doi.org/10.1109/TMM.2023.3244340

34. Yu, Y. C., Lin, H. T. Semi-Supervised Domain Adaptation with Source Label Adaptation. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR 2023), Vancouver, Canada, June 18-22, 2023, 24100-24109. https://doi.org/10.1109/CVPR52729.2023.02308

35. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. Advances in Neural Information Processing Systems, 2021, 34, 18408-18419. doi:10.5555/3540261.3541668.