Spatiotemporal Consensus with Scene Prior for Unsupervised Domain Adaptive Person Search

Yimin Jiang[†]
Dalian Maritime University
yimin_jiang@dlmu.edu.cn

Huibing Wang[*][†]
Dalian Maritime University
huibing.wang@dlmu.edu.cn

Jinjia Peng[†]
Hebei University
pengjinjia@hbu.edu.cn

Abstract

Person Search aims to locate query persons in gallery scene images, but faces severe performance degradation under domain shifts. Unsupervised domain adaptation transfers knowledge from the labeled source domain to the unlabeled target domain and iteratively rectifies the pseudo-labels. However, the pseudo-labels are inevitably contaminated by the source-biased model, which misleads the training process. This, in turn, reduces the quality of the pseudo-labels themselves and ultimately affects the search performance. In this paper, we propose a Spatiotemporal Consensus with Scene Prior (STCSP) framework that effectively eliminates the interference of noise on pseudo-labels, establishes positive feedback, and thus gradually bridging the domain gap. Firstly, STCSP uses a Spatiotemporal Consensus pipeline to suppress the noise from being mixed into the pseudo-labels. Secondly, leveraging the scene prior, STCSP employs our designed Iterative Bilateral Extremum Matching method to prevent the occurrence of some incorrect pseudo-labels. Thirdly, we propose a Scene Prior Contrastive Learning module, which encourages the model to directly acquire the scene prior knowledge from the target domain, thereby mitigating the generation of noise. By suppressing noise contamination, avoiding noise occurrence and mitigating noise generation, our framework achieves state-of-the-art performance on two benchmark datasets, PRW with 50.2% mAP and CUHK-SYSU with 87.0% mAP.

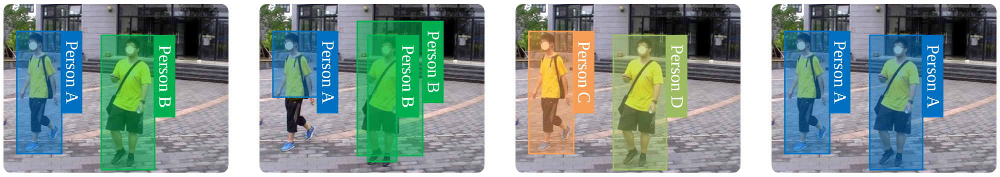
1 Introduction

Person search aims to localize and identify a query person from a gallery of scene images. It can be taken as a joint task of person detection and re-identification (re-id) in an end-to-end manner, where supervised learning has made significant advancements. However, notable performance degradation is observed in these methods when deployed to new application scenarios, primarily attributed to domain gaps induced by factors such as camera configuration. Additionally, the process of annotating an adequate amount of training data for a particular domain is both arduous and costly. Consequently, unsupervised domain adaptation (UDA) holds considerable promise in practical real-world scenarios for person search.

UDA in person search aims to achieve model generalization from labeled source data to unlabeled target deployment scenarios. DAPS [11], a pioneering work in applying UDA to person search, employs implicit domain alignment in conjunction with a pseudo-labeling method to effectively bridge the gap between the source and target domains. Building upon DAPS, Almansoori *et al.* proposed DDAM [1], which generates a hybrid domain and learns within it to minimize the distance between the two domains. Although these methods achieve satisfactory performance, they failed to

[*] Corresponding Author
[†] Equal Contribution
Our code is available at https://github.com/whbdmu/STCSP.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Correct labeling. (b) Inaccurate BBoxes. (c) Incorrect Identities. (d) Identity Conflict.

Figure 1: Display of incorrect pseudo-labels. (a) A correct pseudo-label has an accurate BBox and a correct person identity. (b) An inaccurate pseudo-BBox manifests in the form of either encapsulating an erroneous region or exhibiting the overlap of multiple boxes. (c) An incorrect identity label comes from situations where an identity is assigned to different individuals or where different instances of a person are assigned different identities. (d) Assigning the same identity label to intra-scene persons is a clear error.

effectively eliminate the interference of noise stemming from the source-biased models on pseudo-labels. As illustrated in Fig. 1b, the inaccurate bounding boxes (BBoxes) inevitably lead to detection failures, which can induce significant degradation in person search performance. Additionally, as shown in Fig. 1c, the incorrect labels will directly cause confusion in identifying the person identities. Furthermore, the noise accumulates with the increase of epochs, forming a negative feedback mechanism that will cause performance degradation of the model during the later training phase. Therefore, eliminating the interference of the noise on pseudo-labels is a major challenge. In addition, previous studies failed to realize that a scene can provide relationships for its inner persons. Specifically, the identities of persons within a scene image are inherently distinct, and this is referred to as scene prior. The omission of considering and exploiting this scene prior results in a conspicuous error as illustrated in Fig. 1d. Consequently, the utilization of the scene prior is another crucial issue.

To address these issues, we propose a novel Spatiotemporal Consensus with Scene Prior (STCSP) framework that eliminates the impact of the noise through suppressing noise contamination, preventing some erroneous pseudo-labels and mitigating noise generation. Our framework progressively bridges the domain gap via three core innovations. First, we devise a Spatiotemporal Consensus (STC) pipeline to suppress noise propagation in pseudo-labels. STC maintains a memory bank of the previous detection and re-id information. To filter out temporal jitters, STC leverages the temporal consensus between the information in the memory bank and the current state. In addition, for the detection and re-id subtasks respectively, STC exploits the spatial consensus in the scene space of the image and the latent space of clustering to eliminate spatial outliers. Second, leveraging the intrinsic scene prior, STCSP conducts bipartite matching for the instances in any two of all the images, thus averting the error depicted in Fig. 1d. However, deriving the optimal solutions for bipartite matching across thousands of images is computationally prohibitive. To tackle this challenge, we propose an Iterative Bilateral Extremum Matching (IBEM) method, leveraging GPU acceleration, to seek the approximate solutions. Third, we propose a Scene Prior Contrastive Learning (SPCL) module that encourages the model to learn discriminative person features within the target domain. Specifically, besides proxy learning, SPCL steers the model to learn identity heterogeneity within the same scene, enabling the model to directly absorb knowledge from the target domain. The existence of this knowledge effectively reduces the noise output of the model.

Our major contributions are summarized below:

- A Spatiotemporal Consensus pipeline is proposed to suppress the noise origination from the domain gap, thereby generating reliable pseudo-labels.
- An Iterative Bilateral Extremum Matching method is designed. Within a few seconds, it can match the instances in any two of all thousands of images.
- A Scene Prior Contrastive Learning module is proposed, which encourages the model to directly acquire knowledge from the target domain.

With the spatiotemporal consensus and the scene prior, STCSP achieves state-of-the-art (SOTA) performance on two benchmark datasets. Specifically, it attains a mean average precision of 50.2% on the PRW dataset and 87.0% on the CUHK-SYSU dataset.

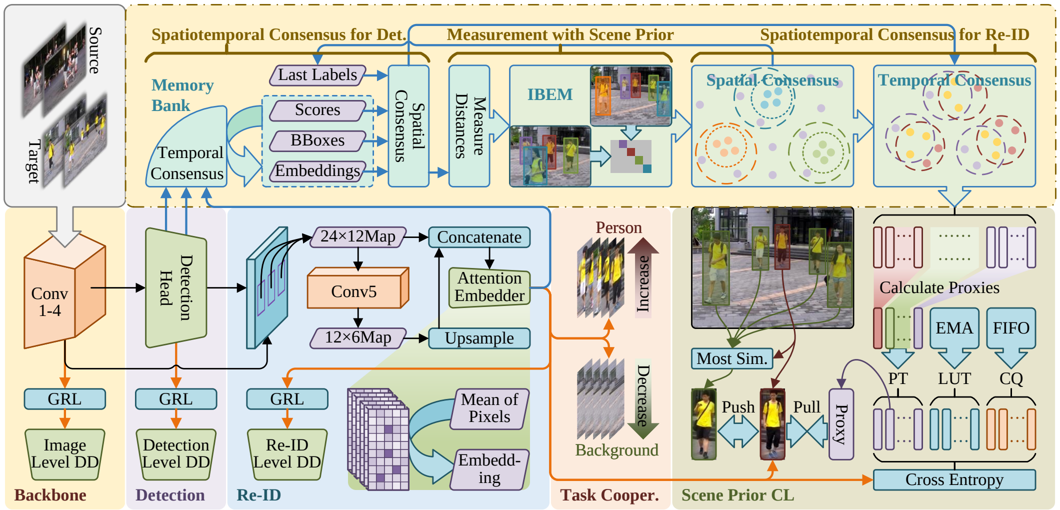


Figure 2: Architecture of the proposed framework. This figure is structured into two rows. The first row presents the input images and the spatiotemporal consensus pipeline. The second row, arranged from left to right, encompasses the backbone network, detection module, re-id module, task cooperation module, and scene prior contrastive learning module.

2 Method

2.1 Framework Overview

Our UDA person search model is an end-to-end architecture based on SeqNet [12]. Differently, we replaced the backbone network and person feature embedder with ConvNeXt-B [16] and our designed attention embedder, respectively. Following the approach of DAPS [11], we introduce gradient reversal layers (GRLs) and domain discriminators (DDs) for three level features. We adhere to the principle of "sample quality first" and use our STC pipeline to obtain reliable pseudo-labels. The SPCL and task cooperation (TC) modules are employed to steer our model to learn discriminative features.

Prior to the training of each epoch, the STC pipeline preprocesses the unlabeled target domain. As in the first row of Fig. 2, the temporal consensus updates the memory bank with model-inferred BBox scores, BBoxes, and embeddings. The spatial consensus then screens out pseudo-BBoxes from the memory bank. Subsequently, their embeddings are calculated by Re-Ranking [30] to form a cross-distance matrix. This matrix undergoes bipartite matching via the IBEM method, and the result is used for matrix sparsity regularization. The regularized matrix generates spatial consensus clusters, which combine with those from the previous epoch to form spatiotemporal ones. The center of each cluster serves as a proxy sent to the SPCL.

During the training phase, the model is fed source and target domain mini-batches in an alternating sequence, with parameters being updated every two mini-batches. As shown in the second row of Fig. 2, the first four ConvNeXt-B [16] stages are responsible for extracting scene features. The feature map is then fed into the Faster-RCNN-based [20] detection module for BBox generation, where RoI-Align pools the instance feature map for each BBox and the Attention Embedder encodes it into an embedding. Finally, the embedding is sent to the TC and SPCL modules for loss calculation.

2.2 Spatiotemporal Consensus for Detection

Temporal Consensus for Detection. For the estimated BBoxes $\mathcal{H}$ inferred by the model and the anchor BBoxes $\mathcal{G}$ stored in the memory bank, we use the intersection-over-union (IoU) scores as the basis for the matching between them. The unmatched BBoxes and their accompanying BBox scores and instance embeddings in the memory bank will be removed, and the remaining contents will be

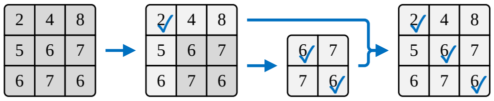


Figure 3: An example of IBEM. On the left and right are the cost matrix and the optimal solution of a minimum bipartite matching task, respectively. The first BEM extracts the primary pair, and the secondary BEM resolves the residuals.

updated with momentum η. The matching and updating operations are expressed as follows:

$$\widetilde{\mathcal{G}} = \left\{ \eta g + (1-\eta) h \,\middle|\, \begin{array}{l} g \in \mathcal{G},\ h \in \mathcal{H},\ \rho(g,h) > \theta, \\ \operatorname*{argmax}_{a \in \mathcal{G}} \rho\left(h,a\right) = g,\ \operatorname*{argmax}_{b \in \mathcal{H}} \rho\left(g,b\right) = h \end{array} \right\}, \tag{1}$$

where $\widetilde{\mathcal{G}}$ is a set of updated anchor BBoxes, ρ is a box IoU function and θ is a threshold for the IoU score. In this equation, the anchor BBox g and the estimated BBox h are mutually IoU-optimal, and the score also meets the threshold. Matching the anchor with the maximum IoU for each estimate is to avoid updating multiple anchors for one estimate. Based on the assumption that the pseudo-labels are true, we consider each anchor to be true. Therefore, selecting the estimate with the maximum IoU is equivalent to maximizing the likelihood probability of generating that anchor.

In addition, the estimated BBoxes in $\mathcal{H}$ that are independent of all anchor BBoxes are used to generate new anchors. Whose definition is as follows:

$$\widehat{\mathcal{H}} = \left\{ h \in \mathcal{H} \,\middle|\, \max_{g \in \mathcal{G}} \rho\left(g,h\right) \leq \theta \right\}. \tag{2}$$

The boxes in $\widehat{\mathcal{H}}$ with a high IoU score between them will be merged to ensure the independence among the new anchors. First, a variant of non-maximum suppression (NMS) operation with threshold θ is applied to divide $\widehat{\mathcal{H}}$ into multiple subsets, as follows:

$$\widehat{\mathcal{H}} \xrightarrow{\text{Split by NMS}_\theta} \left\{ \widehat{\mathcal{H}}_1, \widehat{\mathcal{H}}_2, \cdots \right\}. \tag{3}$$

Specifically, a box retained by the NMS and the boxes it suppresses are grouped together. Finally, the mean values of each subset from the new anchors $\widehat{\mathcal{G}}$. Notably, the score and embedding corresponding to each BBox will undergo the same operations during this process. Each update of the memory bank is a deep integration of historical and current knowledge. Therefore, the result $\mathcal{G} \leftarrow \widetilde{\mathcal{G}} \cup \widehat{\mathcal{G}}$ is a temporal consensus.

Spatial Consensus for Detection. The comprehensive analysis of Eqs. (1) to (3) reveals that the maximum value of the IoU scores among the anchor BBoxes in the memory bank is approximately θ. This indicates that there may be multiple BBoxes stacked on one person instance, as shown in the right of Fig. 1b. Therefore, we apply a NMS with threshold φ to screen out reliable BBoxes from the memory bank, thereby establishing spatial consensus on each image.

2.3 Measurement with Scene Prior

Following previous works, we adopt Re-Ranking [30] method to compute pairwise distances between instance embeddings. To optimize computational performance, we have developed a GPU-accelerated implementation using PyTorch. However, different from previous works, we find that scenes prior can sparsely regularize the distance matrix, thus avoiding the error as shown in Fig. 1d. Specifically, our approach constructs inter-scene cost matrices from the instance distances, performing minimum bipartite matching to set unmapped pair distances to infinity, thereby enforcing intra-scene identity uniqueness. However, the $O\left(n^3\right)$ omplexity of exact solvers like Hungarian algorithm [10] becomes prohibitive at scale ($\approx$ 2hours$/$10k images).

Iterative Bilateral Extremum Matching. To relieve this dilemma, we propose an IBEM method to calculate the approximate solution. Specifically, as shown in algorithm 1, the IBEM iteratively identifies entries that are extrema (*e.g.*, minima or maxima) in both their row and column, then removes the matched entries' rows and columns for further iterations until the conditions are met.

Algorithm 1 Iterative Bilateral Extremum Matching

```
Require: Cost matrix $\mathcal{C} = \{c_{1,1}, \cdots\}$ and its row and column indices $I_1$ and $J_1$.
Ensure: Matching pairs: $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_\delta$
  while $t = 1$ to $\delta$, $|I_t| \times |J_t| \,/\, |I_1| \times |J_1| < \vartheta$ do
      $\mathcal{M}_t \leftarrow \left\{ (i,j) \middle| c_{i,j} = \min_{k \in I_t} c_{k,j}, c_{i,j} = \min_{k \in J_t} c_{i,k} \right\}$
      $I_{t+1} \leftarrow \{ i \in I_t | i \neq p, (p,q) \in \mathcal{M}_t \}$
      $J_{t+1} \leftarrow \{ j \in J_t | j \neq q, (p,q) \in \mathcal{M}_t \}$
  end while
```

Although some rows and columns may remain unmatched during each iteration of bilateral extremum matching (BEM), we rigorously prove by contradiction that a scenario with complete unmatching of all rows and columns is theoretically impossible. Consequently, full instance matching is guaranteed via the iterative process. To balance computational efficiency and convergence speed, we introduce two hyperparameters, an unmatched proportion threshold ϑ and a maximum number of iterations δ as limitations. Fig. 3 shows an example of this method. Notably, by leveraging GPU acceleration, IBEM completes cross-image matching across thousands of instances within seconds.

2.4 Spatiotemporal Consensus for Re-ID

Spatial Consensus for Re-ID. Conceptually, a discriminative cluster requires sufficient separation from other instances. However, the uneven density of the latent space renders a fixed-distance threshold inadequate for the specificity evaluation. To address this challenge, we propose a spatial clustering consensus strategy, which identifies reliable clusters through dual-granularity pattern analysis. Specifically, we implement a dual-density DBSCAN strategy: A conservative density with radius ϵ and min sample size κ captures coarse-grained clusters, while an aggressive density with radius $\alpha \cdot \epsilon$ and min sample size $\alpha \cdot \kappa$ detects fine-grained clusters. Then, through the cluster consensus method, our strategy synthesizes spatial consensus clusters, as shown on the first row of Fig. 2. This cross-validation mechanism effectively distinguishes reliable clusters, ensuring robustness against the latent space with uneven density.

Temporal Consensus for Re-ID. As the instances iteratively update their latent representations, the temporally stable clusters maintain structural cohesion despite positional shifts. Therefore, to screen out the clusters with temporal robustness, we propose a temporal clustering consensus strategy, which integrates the spatial consensus results across previous and current epochs. This cross-epoch fusion process identifies the persistent clusters by the cluster consensus method, thereby forming temporal consensus results. Specifically, as shown on the first row of Fig. 2, the temporal consensus window is constrained to two adjacent epochs, enforcing gradual cluster evolution while preserving temporal consensus across training iterations.

Cluster Consensus. For a certain element ξ in a set $\Xi = \{\xi_1, \xi_2, \cdots, \xi_n\}$ containing n instances, there are two different strategies, Φ and Ψ, to obtain its cluster labels $\Phi(\xi, \Xi)$ and $\Psi(\xi, \Xi)$ respectively. So, in strategy Φ, the cluster which ξ belongs to is as follows:

$$\mathrm{F}(\xi, \Xi, \Phi) = \{ x \in \Xi | \Phi(x, \Xi) = \Phi(\xi, \Xi) \}. \tag{4}$$

Based on F, an intersection-over-union score of the clusters where ξ is located in Φ and Ψ can be calculated as follows:

$$\mathrm{f}(\xi, \Xi, \Phi, \Psi) = \frac{|\mathrm{F}(\xi, \Xi, \Phi) \cap \mathrm{F}(\xi, \Xi, \Psi)|}{|\mathrm{F}(\xi, \Xi, \Phi) \cup \mathrm{F}(\xi, \Xi, \Psi)|}, \tag{5}$$

where $\mathrm{F}(\xi, \Xi, \Phi)$ and $\mathrm{F}(\xi, \Xi, \Psi)$ represent the clusters where ξ is placed in Φ and Φ, respectively. Then, the maximum IoU score among all instances in $\mathrm{F}(\xi, \Xi, \Phi)$ is as follows:

$$\mathrm{g}(\xi, \Xi, \Phi, \Psi) = \max_{x \in \mathrm{F}(\xi, \Xi, \Phi)} \mathrm{f}(x, \Xi, \Phi, \Psi). \tag{6}$$

Finally, a cluster consensus label for ξ is obtained through the following conditional mapping:

$$\mathrm{h}(\xi, \Xi, \Phi, \Psi) = \begin{cases} \Phi(\xi, \Xi), & \mathrm{f}(\xi, \Xi, \Phi, \Psi) = \mathrm{g}(\xi, \Xi, \Phi, \Psi) > 0.5 \\ -1, & \text{else} \end{cases}, \tag{7}$$

where $\mathrm{f}(\xi, \Xi, \Phi, \Psi) = \mathrm{g}(\xi, \Xi, \Phi, \Psi)$ means that ξ lies in the intersection of $\mathrm{F}(\xi, \Xi, \Phi)$ and $\mathrm{F}(\xi, \Xi, \Psi)$, which has the maximum IoU score within $\mathrm{F}(\xi, \Xi, \Phi)$. Moreover, we enforce a 0.5 IoU threshold to validate high confidence in Φ and Ψ consensus. Besides, an instance with cluster label of -1 denotes an outlier.

2.5 Scene Prior Contrastive Learning

Existing works have achieved satisfactory results through contrastive learning using person feature proxies. However, they failed to realize and utilize scene prior to guide the model learning. Therefore, we design a triplet loss function to encourage the model to learn this prior knowledge. Specifically, we introduce a person proxy table (PT) $V = \{v_1, v_2, \cdots\}$ for the target domain. Before each training epoch, the PT will be re-formed by taking the mean value of embeddings in each cluster. And, during the training, it is updated online with a momentum λ. Given a set $\mathcal{I}$ of the instance feature embeddings in a scene image, the margin m of an embedding x in $\mathcal{I}$ is calculated as follows:

$$m = \left\langle v_{\mathrm{L}(x)}, x \right\rangle - \max_{a \in \mathcal{A}} \langle a, x \rangle \,, \ \mathcal{A} = \{a \in \mathcal{I} | \mathrm{L}(a) \neq \mathrm{L}(x)\} \,, \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two embeddings and $\mathrm{L}(\cdot)$ represents taking the label corresponding to the embedding. $v_{\mathrm{L}(x)}$ is the proxy in V corresponding to the x and $\mathcal{A}$ is all the embeddings in $\mathcal{I}$ that have different labels from x. Besides, we use a hyper-parameter M to evaluate the margin. And the triplet loss objective is to increase the margin m to M:

$$\mathcal{L}_{Triplet} = \max(M - m, 0) \,. \tag{9}$$

In addition, we also employ a cross-entropy loss in this module. It calculates the probability that the label of x is $\mathrm{L}(x)$, and uses cross-entropy to supervised the probability.

$$\mathcal{L}_{CE} = -\log p, \ p = \frac{\varepsilon\left(v_{\mathrm{L}(x)}, x\right)}{\sum\limits_{a \in V} \varepsilon(a, x) + \sum\limits_{b \in U} \varepsilon(b, x) + \sum\limits_{c \in Q} \varepsilon(c, x)}, \ \varepsilon(a, b) = \exp(\langle a, b \rangle / \tau), \tag{10}$$

where $\tau = 1/30$ is hyper-parameter which adjusts the softness of the probability distribution. U and Q are respectively the lookup table (LUT) and circular queue (CQ) proposed by OIM, which are updated by the embeddings generated from the source domain. Finally, the scene prior contrastive learning loss is obtained by adding two terms together:

$$\mathcal{L}_{SPCL} = \mathcal{L}_{Triplet} + \mathcal{L}_{CE}. \tag{11}$$

2.6 Task Cooperation

BNR [9] loss utilizes the foreground and background labels from the detection head to supervise the embeddings generated by the re-id head. It can encourage the backbone to enhance the foreground features and suppress the background features, thereby improving the model's detection and re-id performance simultaneously. We improved the equation for calculating the probability that an embedding x belongs to the foreground as follows:

$$p = \sigma\left[\mathrm{BN}\left(\|x\|_2^2 \big/ 2\right)\right], \tag{12}$$

where $\sigma(\cdot)$ is a sigmoid function, $\mathrm{BN}(\cdot)$ is an batch normalization [8] layer, and $\|\xi\|_2$ is the L_2-norm value of the feature embedding x. We employ the square L_2-norm as the regularization term to impose independent constraints on each dimension of x and prevent outliers in some dimensions from influencing others through cross-term interactions. And dividing by 2 is to simplify the derivative operation. Finally, Focal Loss [14] is calculated for p, which focus training on the hard negatives:

$$\mathcal{L}_{TC} = -(1 - p_t)^2 \log(p_t), \ p_t = \begin{cases} p, & y = 1 \\ 1 - p, & y = 0 \end{cases}, \tag{13}$$

where y is a label (either 0 or 1) indicating whether its corresponding embedding is marked as a background or person.

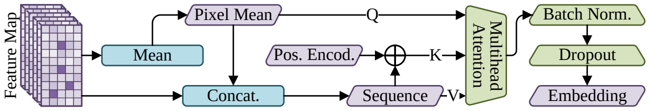


Figure 4: Details of the proposed attention embedder. “Pos. Encod.”, “Concat.” and “Batch Norm.” respectively represent “Positional Encodings”, “Concatenate” and “Batch Normalization”.

2.7 Attention Embedder

Inspired by the attention pool module in CLIP [19], we propose a novel attention embedder as shown in Fig. 4. Distinct from the attention pool module, our design introduces two critical enhancements to the multi-head attention output: Batch Normalization [8] (BN) and Dropout [21]. BN standardizes the distribution of each embedding dimension by centering it near zero with unit variance, mitigating its congregation in positive or negative ranges. This normalization encourages a discriminative embedding and enhance its expressivity in the latent space. Subsequently, Dropout injects stochasticity during training, preventing over-reliance on specific embedding dimensions. Through the synergism of these components, a discriminative and robust embedding is obtained.

3 Experiments

3.1 Experimental Setup and Implementation Details

CUHK-SYSU [25] is a large-scale person search dataset that contains 18,184 images collected from handheld cameras, movies, and TV shows, resulting in significant scene diversity. It encompasses 8,432 unique person identities and 96,143 annotated bounding boxes. The dataset is split into a training set with 5,532 identities and 11,206 images, and a test set with 2,900 queries and 6,978 gallery images. For each query, the dataset defines a gallery size ranging from 50 to 4,000, with a default gallery size of 100 images.

PRW [29] is composed of video frames captured by six fixed cameras on a university campus. It contains 11,816 scene images with 932 distinct person identities and 43,110 annotated bounding boxes. The training set consists of 483 identities with 5,704 images, and the test set contains 2,057 queries and 6,112 scene images. For each query, the dataset uses all images in the test set except for the query as the gallery.

Evaluation Metric. Following the established setting in previous work, the mean Average Precision (mAP) and top-1 accuracy (top-1) are employed to evaluate the performance for person search.

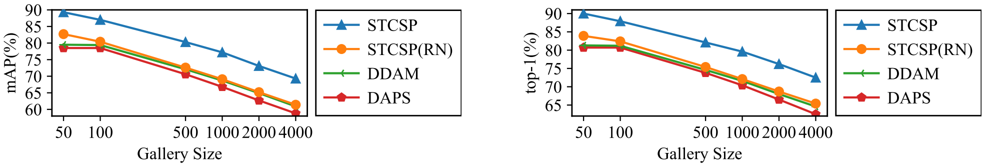
Implementation Details. We conduct all experiments on a NVIDIA A800 GPU and implement our model with PyTorch [18]. ConvNeXt [16] pre-trained on ImageNet [5] is adopted as the backbone network. During training, we set the batch size to 4, and use Automatic Mixed Precision (AMP). Adam is used to optimize our model for 20 epochs with an initial learning rate of 0.0001 and a weight decay of 0.01, which is warmed up during the first epoch and reduced by a factor of 0.1 at epochs 8 and 14. When PRW is used as the target domain, our model undergoes pre-training on the source domain CUHK-SYSU for 5 epochs before commencing joint training. Conversely, it pre-train on PRW for 3 epoch. The thresholds θ and φ in Sec. 2.2 are set to 0.7 and 0.4, respectively. For algorithm 1, the limitations δ and ϑ are set to 2 and 0.2, respectively. The hyper-parameter ϵ, κ and α in Sec. 2.4 is set to 0.5, 4 and 0.8, respectively. For Eq. (9), we initialize the hyper-parameter M as 0.35. For the attention embedder, the probability of the Dropout is set to 0.1.

3.2 Comparison to the State-of-the-Arts

Comparison on PRW. The 3rd column of Tab. 1 shows the performance of all methods on the PRW test set. As shown in the table, our framework outperforming all the other UDA and weakly-unsupervised methods, whether based on ConvNeXt or ResNet. Especially for the current best-performing DDAM [1], STCSP outperforms it by 13.5% and 6.1% in mAP and top-1 accuracy, respectively. And when both employ ResNet-50 as the backbone, our method still outperform it by 5.7% and 0.9%. Notably, our method outperforms most fully supervised methods, despite they using ground truth labels to train models.

Table 1: Comparison of mAP(%) and top-1(%) accuracy with the fully-supervised, weakly-supervised and UDA methods on the PRW and CUHK-SYSU test sets. For UDA methods, the best and second best scores are shown in bold and underlined, respectively.

Method	**Backbone**	**PRW**		**CUHK-SYSU**	
		mAP	top-1	mAP	top-1
Fully Supervised Learning					
OIM [25]	ResNet-50	21.3	49.4	75.5	78.7
IAN [24]	ResNet-50	23.0	61.9	76.3	80.1
NPSM [15]	ResNet-50	24.2	53.1	77.9	81.2
RCAA [2]	ResNet-50	-	-	79.3	81.3
CTXG [28]	ResNet-50	33.4	73.6	84.1	86.5
QEEPS [17]	ResNet-50	37.1	76.7	88.9	89.1
NAE+ [3]	ResNet-50	44.0	81.1	92.1	92.9
AlignPS+ [27]	ResNet-50	46.1	82.1	94.0	94.5
SeqNet [12]	ResNet-50	46.7	83.4	93.8	94.6
SEAS [9]	ResNet-50	52.0	85.7	96.2	97.1
SEAS [9]	ConvNeXt-B	60.5	89.5	97.1	97.8
Weakly Supervised Learning					
R-SiamNet[6]	ResNet-50	21.4	75.2	86.0	87.1
CGPS [26]	ResNet-50	16.2	68.0	80.0	82.3
SSL [22]	ResNet-50	30.7	80.6	87.4	88.5
KCD [13]	ResNet-50	40.5	83.6	86.8	88.2
DICL [23]	ResNet-50	35.5	80.9	87.4	88.8
Unsupervised Domain Adaptation					
DAPS [11]	ResNet-50	34.7	80.6	77.6	79.6
FOUS [4]	ResNet-50	35.4	80.8	78.7	80.5
DDAM [1]	ResNet-50	36.7	81.2	79.5	81.3
STCSP (ours)	ResNet-50	42.4	82.1	80.4	82.5
STCSP (ours)	ConvNeXt-B	**50.2**	**87.3**	**87.0**	**87.9**



(a) Comparison of mAP.

(b) Comparison of top-1 accuracy.

Figure 5: Comparison of mAP and top-1 accuracy on CUHK-SYSU across various gallery sizes.

Comparison on CUHK-SYSU. The performance of our method and the other methods on the CUHK-SYSU test set is shown in the 4th column of Tab. 1. STCSP based on ConvNeXt or ResNet outperforms all the other UDA methods. It achieves the best scores of 87.0% in mAP and 87.9% in top-1 accuracy, surpassing DDAM [1] by 7.5% and 6.6% in mAP and top-1 accuracy, respectively. The CUHK-SYSU training set boasts greater diversity and a larger quantity of scene images compared to the PRW training set. As a result, UDA from PRW to CUHK-SYSU presents a formidable challenge. Despite this, our method still outperforms several weakly-supervised and fully-supervised methods on the CUHK-SYSU test set.

We further perform a comparison between our STCSP and other methods on the CUHK-SYSU test set with gallery sizes ranging from 50 to 4,000. Fig. 5 shows the performance curve of all the methods in terms of mAP and top-1 as the gallery size increases. Since it is a challenge for all compared methods to consider more distracting persons in the gallery set, the performance of them is reduced as the gallery size increases. However, our method consistently outperforms all the other methods in

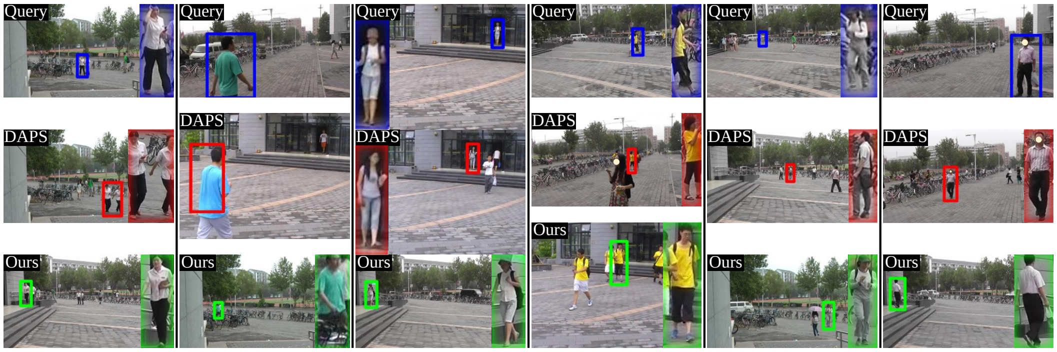


Figure 6: Qualitative comparison of our method with DAPS on the PRW test set. Each column represents a comparison result. The blue bounding boxes denote the queries, while the green and red bounding boxes denote correct and incorrect top-1 matches, respectively.

Table 2: Comparative results of the different strategies for our framework.

Strategy	**PRW**		**CUHK-SYSU**	
	mAP	top-1	mAP	top-1
w/o Detection Spatial Consensus	46.8	85.7	85.4	85.4
w/o Detection Temporal Consensus	47.7	86.0	86.4	87.5
w/o Re-ID Spatial Consensus	46.9	85.8	86.7	87.7
w/o Re-ID Temporal Consensus	49.0	87.4	86.6	87.5
w/o IBEM	48.2	85.9	85.9	86.6
w/o Triple Loss of SPCL	47.9	87.0	86.2	87.4
	50.2	**87.3**	**87.0**	**87.9**

mAP and top-1, whether based on ConvNeXt or ResNet. This indicates that our method has excellent generalization ability.

Qualitative Results. Some example person search results are illustrated in Fig. 6. The 5th column of the figure shows the result of using a low-resolution image as the query. When multiple highly similar characters appear in a scene, the search results are illustrated in the 1st and 4th columns. The 2nd column shows the result of the person's body being obstructed. The 3rd and 6th columns show the result when the scene and viewpoint change. It can be observed that our successfully handles low-resolution, similar-character, occlusion, cross-scene and viewpoint variation, while other state-of-the-art methods such as DAPS fail in these scenarios. This demonstrates the robustness of our method.

3.3 Ablation Study

Spatiotemporal Consensus for Detection. To verify the effectiveness of our STC strategy for detection, we evaluate separately the performance of our framework in two scenarios: detection without spatial consensus and without temporal consensus, and report the results in Tab. 2. We find that STCSP exhibits significant performance degradation when lacking the spatial or temporal consensus for detection. This results indicate that the STC strategy for detection significantly elevates the quality of pseudo-BBoxes.

Spatiotemporal Consensus for Re-ID. Tab. 2 shows the performance of our framework being employed with different strategies on the two benchmarks. First. to validate the effectiveness of the spatial consensus, we replace it with a DBSCAN method. The results reveal that our strategy of employing a loose and compact clustering space to form spatial consensus can generate reliable clusters. Furthermore, The results after removal of the temporal consensus are also reported in Tab. 2. Through comparison, we can see that the temporal consensus improve the performance of our model

Table 3: Comparative results of the different components for our framework. RN, CN, NAE and AE indicate ResNet-50, ConvNeXt-B, Norm-Aware Embedding and Attention Embedder, respectively.

Backbone		**Embedder**		**PRW**		**CUHK-SYSU**	
RN	CN	NAE	AE	mAP	top-1	mAP	top-1
✓	✗	✓	✗	41.1	82.2	80.3	82.3
✓	✗	✗	✓	42.4	82.2	80.4	82.5
✗	✓	✓	✗	45.8	85.6	85.6	85.5
✗	✓	✗	✓	**50.2**	**87.3**	**87.0**	**87.9**

on both of PRW and CUHK-SYSU. We attribute these improvements to the stable clusters formed by the temporal consensus.

Scene Prior. We evaluated the influence exerted by the lack of scene priors within the related components on the performance manifestations of our framework, and report the results in Tab. 2. As presented in the 5th row of the table, when the IBEM is absent from our framework, entailing the potential occurrence of the error depicted in Fig. 1d, the performance of our framework experiences a substantial decline across both datasets. Notably, this decline is more pronounced on the PRW dataset. Similarly, as indicated in the 6th row of the table, a comparable scenario transpires when the triplet loss within the SPCL is lacking. The significant performance degradation observed on PRW can be attributed to the existence of persons with a high similarity within some scene images in this dataset. This indicates that scene prior effectively steers the model to learn discriminative features.

Model Architecture. As shown in Tab. 3, in order to assess the influence of diverse model architectures on our framework, we report the performance outcomes under all cross-combination modes of ResNet-50 (RN) [7] and ConvNeXt-B (CN) [16] with Norm-Aware Embedding (NAE) [3] and Attention Embedder (AE). When comparing backbone networks, ConvNeXt-B significantly outperforms ResNet-50 across all configurations. Furthermore, the AE leads to a further performance improvement compared to the NAE, especially when it is integrated with CN. Notably, the performance gains of AE are relatively less pronounced when combined with RN, suggesting that the compatibility between the backbone network and the embedder affects the performance of the framework. The results indicate that the combination of CN and AE is the optimal configuration.

4 Conclusion

In this paper, we focus on UDA Person Search, where existing methods suffer from noise accumulation caused by source-biased pseudo-labels. To bridge the domain gap, we propose the STCSP framework, which systematically suppresses noise contamination, avoids noise occurrence, and mitigates noise generation. With STC pipeline, IBEM method and SPCL module, our framework achieves SOTA performance. For future research, the IBEM algorithm can be extended to other tasks that demand the rapid completion of extensive bipartite matchings.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China Grant 62576067 and 62501226, National Key Research and Development Program of China Grant 2024YFB4710800, Liaoning Provincial Natural Science Foundation Grant 2025-YQ-01 and 2024-MS-012, Dalian Science and Technology Talent Innovation Support Plan Grant 2024RY010, Natural Science Foundation of Hebei Province Grant F2025201037.

References

[1] Mohammed Khaleed Almansoori, Mustansar Fiaz, and Hisham Cholakkal. Ddam-ps: Diligent domain adaptive mixer for person search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6688–6697, 2024.

[2] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. Rcaa: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 84–100, 2018.

[3] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12615–12624, 2020.

[4] Tianxiang Cui, Huibing Wang, Jinjia Peng, Ruoxi Deng, Xianping Fu, and Yang Wang. Fast one-stage unsupervised domain adaptive person search. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 713–721. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Chuchu Han, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang, and Changhu Wang. Weakly supervised person search with region siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12006–12015, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[9] Yimin Jiang, Huibing Wang, Jinjia Peng, Xianping Fu, and Yang Wang. Scene-adaptive person search via bilateral modulations. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 938–946. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[11] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Domain adaptive person search. In *European conference on computer vision*, pages 302–318. Springer, 2022.

[12] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2011–2019, 2021.

[13] Zongyi Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Runsheng Wang, Chengxin Zhao, Qian Wang, and Shijuan Huang. Knowledge consistency distillation for weakly supervised one step person search. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[15] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 493–501, 2017.

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[17] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[22] Benzhi Wang, Yang Yang, Jinlin Wu, Guo-jun Qi, and Zhen Lei. Self-similarity driven scale-invariant learning for weakly supervised person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1813–1822, 2023.

[23] Jiabei Wang, Yanwei Pang, Jiale Cao, Hanqing Sun, Zhuang Shao, and Xuelong Li. Deep intra-image contrastive learning for weakly supervised one-step person search. *Pattern Recognition*, 147:110047, 2024.

[24] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87:332–340, 2019.

[25] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017.

[26] Yichao Yan, Jinpeng Li, Shengcai Liao, Jie Qin, Bingbing Ni, Ke Lu, and Xiaokang Yang. Exploring visual context for weakly supervised person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3027–3035, 2022.

[27] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2021.

[28] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2158–2167, 2019.

[29] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.

[30] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduced the research content and listed the contributions of this paper in Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Due to the length limitation of the paper, we do not discuss these in detail.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We derived the theoretical results in Sec. 2 and validated them in Sec. 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have elaborated on the proposed method in Sec. 2 and provided implementation details in Sec. 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: If this paper fortunate enough to be accepted, we will provide open access to the data and code in the final version.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided implementation details in 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental data reported in the chart is the mean of several experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Due to the length limitation of the paper, we are unable to report these data in detail.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have conducted an inspection.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. 1 introduces the positive impact of this study on society, but no negative impact has been found so far.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The method proposed in this paper is only applicable in certain specific fields and currently has no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper provides detailed annotations for all of these.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.