Self-Supervised Partial Cycle-Consistency for Multi-View Matching

Fedor Taggenbrock[1,2] [a], Gertjan Burghouts[1] [b] and Ronald Poppe[2] [c]

[1]*Utrecht University, Utrecht, Netherlands*
[2]*TNO, The Hague, Netherlands*

Abstract: Matching objects across partially overlapping camera views is crucial in multi-camera systems and requires a view-invariant feature extraction network. Training such a network with cycle-consistency circumvents the need for labor-intensive labeling. In this paper, we extend the mathematical formulation of cycle-consistency to handle partial overlap. We then introduce a pseudo-mask which directs the training loss to take partial overlap into account. We additionally present several new cycle variants that complement each other and present a time-divergent scene sampling scheme that improves the data input for this self-supervised setting. Cross-camera matching experiments on the challenging DIVOTrack dataset show the merits of our approach. Compared to the self-supervised state-of-the-art, we achieve a 4.3 percentage point higher F1 score with our combined contributions. Our improvements are robust to reduced overlap in the training data, with substantial improvements in challenging scenes that need to make few matches between many people. Self-supervised feature networks trained with our method are effective at matching objects in a range of multi-camera settings, providing opportunities for complex tasks like large-scale multi-camera scene understanding.

1 INTRODUCTION

Matching people and objects across cameras is essential for multi-camera understanding (Hao et al., 2023; Loy et al., 2010; Zhao et al., 2020). Matches are commonly obtained by solving a multi-view matching problem. One crucial factor that determines the quality of the matching is the feature extractors' generalization to varying appearances as a result of expressiveness and view angle (Ristani and Tomasi, 2018). Feature extractors can be trained in a supervised setting, which requires labor-intensive data labeling (Hao et al., 2023). The lack or scarcity of labeled data for novel domains is a limiting factor. Self-supervised techniques thus offer an attractive alternative because they can be trained directly on object and person bounding boxes without labels.

Effective, view-invariant feature networks have been learned with self-supervision through cycle-consistency, for use in multi-view matching, cross-view multi-object tracking, and re-identification (Re-ID) (Gan et al., 2021; Wang et al., 2020). Training

[a] https://orcid.org/0009-0002-6166-0865
[b] https://orcid.org/0000-0001-6265-7276
[c] https://orcid.org/0000-0002-0843-7878

these networks only requires sets of objects where there is a sufficient amount of overlap between sets of objects between views. For multi-person matching and tracking, sets are typically detections of people from multiple camera views (Gan et al., 2021; Hao et al., 2023). When the overlapping field of view between cameras decreases in the training data, self-supervised cycle-consistency methods have a diluted learning signal.

In this work, we address this situation and extend the theory of cycle-consistency for partial overlap with a new mathematical formulation. We then implement this theory to effectively handle partial overlap in the training data through a pseudo-mask, and introduce trainable cycle variations to obtain a richer learning signal, see Figure 1. Consequently, we can get more out of the training data, thus providing a stronger cycle-consistency learning signal. Our method is shown to be robust in more challenging settings, with less overlap between cameras and fewer matches in the training data. It is especially effective for challenging scenes where few matches need to be found between many people. The additional information from partial cycle-consistency thus leads to substantial improvements, as shown in the experimental
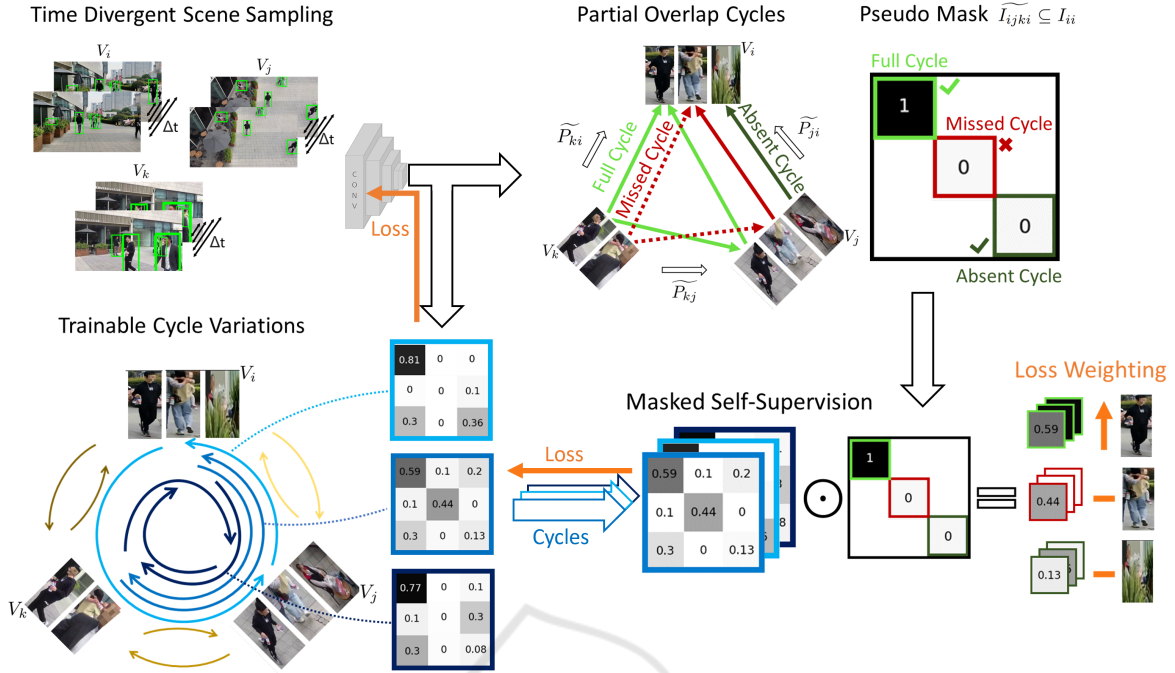
Figure 1: Overview of our self-supervised cycle-consistency training method. Trainable cycle variations (left bottom) are constructed from sampled batches (left top). Cycle matrices represent chains of matches starting and ending in the same view. With partial overlap, however, we construct a pseudo-mask of the identity matrix (top right) to determine which specific cycles should be trained due to partial overlap. This pseudo-mask is then used to provide a weighted loss signal with more emphasis on the positive predicted cycles (right bottom).

section. The code is also made open source[1]

Our contributions are as follows:

1. We extend the mathematical formulation of cycle-consistency to handle partial overlap, leading to a new formulation for partial cycle-consistency.

2. We use pseudo-masks to implement partial cycle-consistency and introduce several cycle variants, motivating how these translate to a richer self-supervision learning signal.

3. We experiment with cross camera matching on the challenging DIVOTrack dataset, and obtain systematic improvements. Our experiments highlight the merits of using a range of cycle variants, and indicate that our approach is especially effective in more challenging scenarios.

Section 2 covers related works on self-supervised feature learning. Section 3 summarizes our mathematical formulation and derivation of cycle-consistency with partial overlap. Section 4 details our self-supervised method. We discuss the experimental validation in Section 5 and conclude in Section 6.

[1]For the open source code and theoretical analysis, see the Supplementary Materials available at Github.

2 RELATED WORK

We first address the general multi-view matching problem, and highlight its application areas. Section 2.2 summarizes supervised feature learning, whereas Section 2.3 details self-supervised alternatives.

2.1 Multi-View Matching

Many problems in computer vision can be framed as a multi-view matching problem. Examples include keypoint matching (Sarlin et al., 2020), video correspondence over time (Jabri et al., 2020), shape matching (Huang and Guibas, 2013), 3D human pose estimation (Dong et al., 2019), multi-object tracking (MOT) (Sun et al., 2019), re-identification (Re-ID) (Ye et al., 2021), and cross-camera matching (CCM) (Han et al., 2022). Cross-view multi-object tracking (CVMOT) combines CCM with a tracking algorithm (Gan et al., 2021; Hao et al., 2023). The underlying problem is that there are more than two views of the same set of objects, and we want to find matches between the sets. For MOT, detections between two subsequent time frames are matched (Wo-

jke et al., 2017). Instead, in CCM, detections from different camera views should be matched. One particular challenge is that the observations have significantly different viewing angles. Such invariancies should be handled effectively through a feature extraction network. Such networks can be trained using identity label supervision but obtaining consistent labels across cameras is labor-intensive (Hao et al., 2023), highlighting the need for good self-supervised alternatives.

2.2 Supervised Feature Learning

Supervised Re-ID methods (Wieczorek et al., 2021; Ye et al., 2021) work well for CCM. With labels, feature representations from the same instance are metrically moved closer, while pushing apart feature representations from different instances. Other approaches such as joint detection and Re-ID learning (Hao et al., 2023), or training specific matching networks (Han et al., 2022) have been explored. Supervised methods for CCM typically degrade in performance when applied to unseen scenes, indicating issues with overfitting. Self-supervised cycle-consistency (Gan et al., 2021) has been shown to generalize better (Hao et al., 2023).

2.3 Self-Supervised Feature Learning

Self-supervised feature learning algorithms do not exploit labels. Rather, common large-scale self-supervised contrastive learning techniques (Chen et al., 2020) rely on data augmentation. We argue that the significant variations in object appearance across views cannot be adequately modeled through data augmentations, meaning that such approaches cannot achieve view-invariance. Clustering-based self-supervised techniques (Fan et al., 2018) are also not designed to deal with significant view-invariance. Another alternative is to learn self-supervised features through forcing dissimilarity between tracklets within cameras while encouraging association with tracklets across cameras (Li et al., 2019). Early work on self-supervised cycle-consistency has shown that this framework significantly outperforms clustering and tracklet based self-supervision methods (Wang et al., 2020). Self-supervision with cycle-consistency is especially suitable for multi-camera systems because it enables learning to associate consistently between the object representations from different cameras and at different timesteps. Trainable cycles can be constructed as series of matchings that start and end at the same object. Each object should be matched back to itself as long as the object is visible in all views. If

an object is matched back to a different one, a cycle-inconsistency has been found which then serves as a learning signal (Jabri et al., 2020; Wang et al., 2020).

Given the feature representations of detections in two different views, a symmetric cycle between these two views can be constructed by combining two soft-maxed similarity matrices, matching back and forth. The feature network can then be trained by forcing this cycle to resemble the identity matrix with a loss (Wang et al., 2020). This approach can be extended to transitive cycles between three views, which is sufficient to cover cycle-consistency between any number of views (Gan et al., 2021; Huang and Guibas, 2013). With little partial overlap in the training data, forcing cycles to resemble the full identity matrix (Gan et al., 2021; Wang et al., 2020) provides a diluted learning signal that trains many non-existent cycles without putting proper emphasis on the actual cycles that should be trained. To effectively handle partial overlap, it is therefore important to differentiate between possibly existing and absent cycles in each batch. To this end, we implement a strategy that makes this differentiation. A work that was developed in parallel to ours (Feng et al., 2024) has also found improvements with a related partial masking strategy. Our work confirms their observations that considering partial overlap improves matching performance. In addition, we provide a rigid mathematical underpinning for our method, introduce more cycle variations, and trace back improvements to characteristics of the scene including the amount of overlap between views.

Learning with cycle-consistency is not exclusive to CCM. Cycles between detections at different timesteps can be employed to train a self-supervised feature extractor for MOT (Bastani et al., 2021), and cycles between image patches or video frames can serve to learn correspondence features at the image level (Dwibedi et al., 2019; Jabri et al., 2020; Wang et al., 2019). This highlights the importance of a rigid mathematical derivation of partial cycle-consistency in a self-supervised loss.

3 PARTIAL CYCLE-CONSISTENCY

We summarize the main contributions from our theoretical extension of partial cycle-consistency, which appears in full in the supplementary materials[1]. Given are pairwise similarities $S_{ij} \in \mathbb{R}^{n_i \times n_j} \ \forall i, j$ between the views V_i, V_j, that contain n_i, n_j bounding boxes. Partial multi-view matching aims to obtain the optimal partial matching matrices $P_{ij} \in \{0, 1\}^{n_i \times n_j} \ \forall i, j$, given
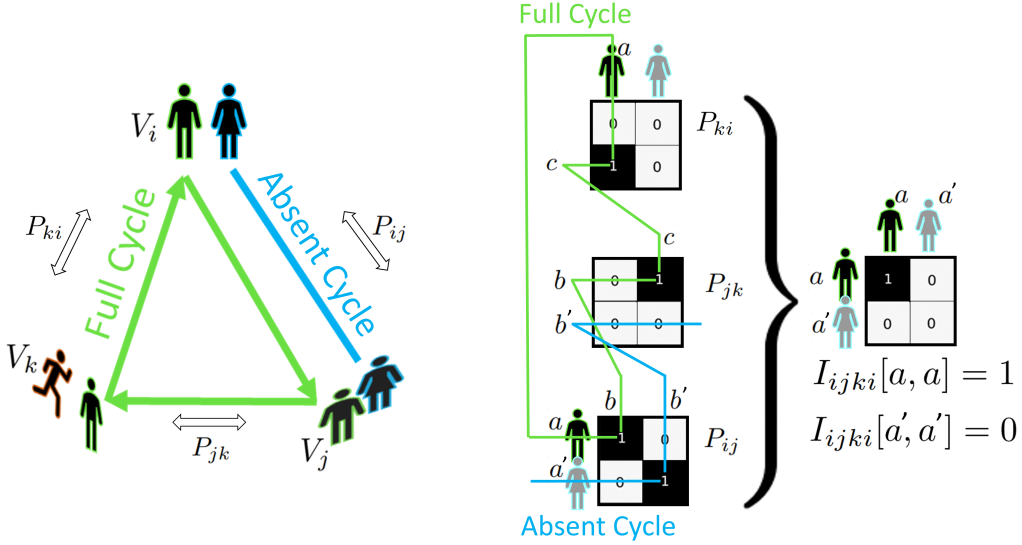
Figure 2: Partial cycle-consistency and an interpretation of Equation 5. $I_{ijki}[a,a] = 1$ because a is matched to b, matched to c which is then matched back to a. The same does not hold for a', so this cycle is absent.

the S_{ij} that are partially cycle-consistent with each other. See also Figure 2. Partial cycle-consistency implies that, among others, matching from view V_i to view V_j and then to view V_k should be a subset of the direct matching between V_i and V_k. We make this subset relation explicit, pinpointing which matches get lost through view V_j by inspecting the pairwise matches, proving equivalence to the original definition. We then prove that partial cycle-consistency in general implies the most usable form of self-supervision cycle-consistency, where matches are combined into full cycles that start and end in the same view and should thus be a subset of the identity matrix. We are able to explicitly define this usable form of partial cycle-consistency in proposition 1. Based on this insight, in Section 4, we construct subsets of the identity matrix during training to serve as pseudo-masks, improving the training process with partially overlapping views. Our explicit cycle-consistency proposition is as follows:

Proposition 1 (Explicit partial cycle-consistency).
If a multi-view matching $\{P_{ij}\}_{\forall i,j}$ is partially cycle-consistent, it holds that:

$$P_{ii} = I_{n_i \times n_i} \quad \forall i \in \{1,\dots,N\}, \tag{1}$$

$$P_{ij}P_{ji} = I_{iji} \quad \forall i,j \in \{1,\dots,N\}, \tag{2}$$

$$P_{ij}P_{jk}P_{ki} = I_{ijki} \quad \forall i,j,k \in \{1,\dots,N\}, \tag{3}$$

where $I_{iji} \subseteq I_{n_i \times n_i}$ is the identity map from view i back to itself, filtering out matches that are not seen in view V_j:

$$I_{iji}[a,c] = \begin{cases} 1 & \text{if } a = c \ \& \ \exists b \text{ s.t. } P_{ij}[a,b] = 1. \\ 0 & \text{else}, \end{cases} \tag{4}$$

and where $I_{ijki} \subseteq I_{n_i \times n_i}$ is the identity mapping from view i back to itself, filtering out all matches that are not seen in views V_j and V_k:

$$I_{ijki}[a,d] = \begin{cases} 1 & \text{if } a = d \ \& \ \exists b,c \text{ s.t. } P_{ij}[a,b] \\ & = P_{jk}[b,c] = P_{ki}[c,d] = 1. \\ 0 & \text{else}. \end{cases} \tag{5}$$

The notation $X[\cdot,\cdot]$ is used for indexing a matrix X.

The intuition behind Equation 5 can be best understood through the visualization in Figure 2. Here, $I_{ijki}[a',a'] = 0$ because there is a detection of a' absent in view V_k, while $I_{ijki}[a,a] = 1$ because a full cycle is formed from the corresponding pairwise matches. The proofs with detailed explanations are given in the supplementary materials[1].

4 SELF-SUPERVISION WITH PARTIAL CYCLE-CONSISTENCY

The theory of cycle-consistency and its relation to partial overlap can be translated into a self-supervised feature network training strategy. The main challenges are to determine which cycles to train, which loss to use, and how to implement the findings from Proposition 1 to handle partial overlap. Section 4.1 explores what cycles to train and how to construct them. Section 4.2 explores how to obtain partial overlap masks for the cycles that approximate the $I_{iji}, I_{ijki} \subseteq I_{n_i \times n_i}$ from Proposition 1. It also explores

how these masks can be incorporated in a loss to deal with partial overlap during training.

4.1 Trainable Cycle Variations

Given are the pairwise similarities S_{ij} between all view pairs, obtained from the feature extractor ϕ that we wish to train. The idea is to combine softmax matchings of the S_{ij} into cycles, similar to Equations 2 and 3. For this we use the temperature-adaptive row-wise softmax f_τ (Wang et al., 2020) on a similarity matrix S to perform a soft row-wise partial matching. This function has the differentiability needed to train a feature network and the flexibility to make non-matches for low similarity values. We get:

$$f_\tau(S[a,b]) = \frac{\exp(\tau S[a,b])}{\sum_{b'} \exp(\tau S[a,b'])}, \qquad (6)$$

where the notation $S[\cdot,\cdot]$ is used for matrix indexing. The temperature τ depends on the size of S as in (Wang et al., 2020).

4.1.1 Pairwise Cycles

The pairwise cycles need to be constructed from just $S_{ij} = S_{ji}^T$. To this end, we take:

$$A_{ij} = f_\tau(S_{ij}), \quad A_{ji} = f_\tau(S_{ij}^T), \quad A_{iji} = A_{ij}A_{ji}. \quad (7)$$

The pairwise cycle A_{iji}, originally proposed in (Wang et al., 2020), represents a trainable variant of the pairwise cycle $P_{ij}P_{ji}$ from Equation 2, and so a learning signal is obtained by forcing it to resemble I_{iji}. Note that the A_{ij} and A_{ji} differ because they match the rows and columns of S_{ij}, respectively. This is important because the loss then forces these different soft matchings to be consistent with each other, modelling the partial cycle-consistency constraint in Equation 2. The loss will be the same for A_{iji} and $A_{iji}^T = A_{ji}^T A_{ij}^T$, so just Equation 7 suffices.

4.1.2 Triplewise Cycles

The triplewise cycles are constructed from S_{ij}, S_{jk} and S_{ki}, and should resemble the $P_{ij}P_{jk}P_{ki}$ from Equation 3. The authors in (Feng et al., 2024) propose to use:

$$A_{ijki}^0 = A_{ij}A_{jk}A_{ki}, \qquad (8)$$

while in (Gan et al., 2021), the similarities are combined first so that:

$$S_{ijk} = S_{ij}S_{jk}, \quad A_{ijk} = f_\tau(S_{ijk}), \qquad (9)$$

with which their triplewise cycle is created as:

$$A_{ijki}^1 = A_{ijk}A_{kji}. \qquad (10)$$

We discovered that using multiple triplewise cycle constructions in the training improves the results. Each of the constructed cycles exposes a different inconsistency in the extracted features, so that a combination of cycles provides a robust training signal. We propose to use the following four triplewise cycles:

$$A_{ijki}^0 = A_{ij}A_{jk}A_{ki}, \qquad (11)$$

$$A_{ijki}^1 = A_{ijk}A_{kji}, \qquad (12)$$

$$A_{ijki}^2 = A_{ijk}A_{ki}, \qquad (13)$$

$$A_{ijki}^3 = A_{ijk}A_{kij}A_{jki}. \qquad (14)$$

The cycles from Equation 12-14 are also visualized in Figure 1 as the three blue swirls. In the following, A_{ijki} can be used to refer to any of the four triplewise cycles in Equations 11-14, and additionally A_{iji} when assuming $j = k$. The symmetric property of the loss makes transposed versions of Equations 11 - 14 redundant.

4.2 Masked Partial Cycle-Consistency Loss

The A_{ijki} can be directly trained to resemble the identity matrix I_{ii}, by training each diagonal element in A_{ijki} to be a margin m greater than their corresponding maximum row and column values, similar to the triplet loss (Wang et al., 2020; Gan et al., 2021). This is achieved through:

$$L_m(A_{ijki}) = \sum_{i=1}^{n_i} \text{relu}(\max_{b \neq a}(A_{ijki}[a,b]) - A_{ijki}[a,a] + m). \qquad (15)$$

The following loss enforces this margin over both the rows and columns:

$$\mathcal{L}_m(A_{ijki}) = \frac{1}{2}(L_m(A_{ijki}) + L_m(A_{ijki}^T)). \qquad (16)$$

This loss, however, does not distinguish between absent and existing cycles that occur with partial overlap. Note that the ground truth I_{ijki} are masks (or subsets) of the I_{ii} that exactly filter out such absent cycles, while keeping the existing cycles, according to Equation 5 and visualized in Figure 2. In this figure, detections of the blue person form an absent cycle because the pairwise matches are not connected. The I_{ijki} are constructed based on the ground truth matches P_{ij}. We therefore propose to construct pseudo-masks $\widetilde{I}_{ijki}$ from pseudo-matches $\widetilde{P}_{ij}$ that are available during self-supervised training. For this we use:

$$\widetilde{P}_{ij} = \begin{cases} [f_\tau(S_{ij}) > 0.5] & \text{if } |V_i| < |V_j|, \\ \left[f_\tau(S_{ij}^T)^T > 0.5\right] & \text{if } |V_j| < |V_i|, \end{cases} \qquad (17)$$

where the Iverson bracket $[Predicate(X)]$ binarizes matrix X, with elements equal to 1 for which the predicate is true, and 0 otherwise. In $\widetilde{P}_{ij}$, each element in a view with fewer elements can be matched to at most one element in the other view, as desired for a partial matching. We construct the pseudo-masks as:

$$\widetilde{I}_{ijki}[a,a] = \begin{cases} 1 & \text{if } \exists b,c \text{ s.t. } \widetilde{P}_{ij}[a,b] \\ & = \widetilde{P}_{jk}[b,c] = \widetilde{P}_{ki}[c,a] = 1. \\ 0 & \text{else.} \end{cases} \quad (18)$$

$\widetilde{I}_{ijki}$ is invariant to the order in the i,j,k sequence, and independent of the cycle variant for which it is used as a mask. Equation 18 can be vectorized as:

$$\widetilde{I}_{ijki} = \left[\widetilde{P}_{ij}\widetilde{P}_{jk}\widetilde{P}_{ki} \odot I_{ii} \geq 1 \right]. \quad (19)$$

Our masked partial cycle-consistency loss extends the loss from Equation 16 with the pseudo-masks $\widetilde{I}_{ijki}$, for which only the diagonal elements of predicted existing cycles are 1. The absent cycles have diagonal elements of 0. The loss uses two different margins $m_+ > m_\emptyset > 0$, where m_+ is used for cycles that are predicted to exist with $\widetilde{I}_{ijki}$, and $m_\emptyset$ is used for the cycles predicted to be absent:

$$\mathcal{L}_{explicit} = \frac{\mathcal{L}_{m_+}(\widetilde{I}_{ijki} \odot A_{ijki}) + \mathcal{L}_{m_\emptyset}((I_{ii} - \widetilde{I}_{ijki}) \odot A_{ijki})}{2}.$$
$$(20)$$

5 RESULTS AND EXPERIMENTS

We demonstrate the merits of a stronger self-supervised training signal from the addition of our cycle variations and partial cycle-consistency mask. We introduce the training setting, before detailing our quantitative and qualitative results.

Dataset and Metrics. DIVOTrack (Hao et al., 2023) is a large and varied dataset of time-aligned overlapping videos with consistently labeled people across cameras. The train and testset are disjoint sets with 9k frames from three overlapping camera's each. Three time-aligned overlapping frames are one matching instance. Frames from 10 different scenes are used, equally distributed over the train and test set. Our self-supervised feature network trains with the 9k matching instances of the trainset without its labels. We report the average cross-camera matching precision, recall and F1 score (Han et al., 2022) over the 9k matching instances of the test set, averaged over five training runs with standard deviation. The average number of people per matching instance is around 19, but this varies per scene[1].

Implementation Details. Our contributions extends the state-of-the-art in self-supervised cycle-consistency (Gan et al., 2021). Our cycle variations from Equations 11-14 are used instead of theirs, providing a diverse set of cycles to capture different cycle-inconsistencies. Previous methods without masking (Wang et al., 2020) (Gan et al., 2021) use the loss from Equation 16. Our partial masking strategy instead constructs pseudo-masks with Equation 18 and uses these in our explicit partial masking loss from Equation 20, with $m_+ = 0.7$ and $m_\emptyset = 0.3$. We use the same training setup as (Gan et al., 2021) for fair comparison. Specifically, we use annotated bounding boxes without identity labels to extract features and train a ResNet-50 (He et al., 2016) for 10 epochs with an Adam optimizer with learning rate $1e-5$. Matching inference uses the Hungarian algorithm between all view pairs, with an optimized partial overlap parameter to handle non-matches.

Time-Divergent Scene Sampling. Detections from multiple cameras at two timesteps are used in a batch such that cycles are constructed and trained between the pairs and triples for $2C$ views of the same scene, with C the number of cameras (Feng et al., 2024; Gan et al., 2021). Time-divergent scene sampling gradually increases the interval Δt between timesteps during training, with Δt equal to the current epoch number. It also uses fractional sampling to obtain a balanced batch order, such that the local distribution of scenes resembles the average global distribution of scenes.

5.1 Main Results

We show the overall effectiveness of our cycle variations and partial masking as additions to the existing SOTA within the framework of self-supervised cycle-consistency (Gan et al., 2021) in Table 1.

The first paper in this framework (Wang et al., 2020) used a simple baseline approach with just pairwise cycles, and showed the effectiveness compared to multiple other self-supervised feature learning methods using clustering (Fan et al., 2018) and tracklet based techniques (Li et al., 2019) among others. The authors in (Gan et al., 2021) and (Feng et al., 2024) expanded upon this framework, where (Feng et al., 2024) is not open source. We report the results in our paper both with and without time-divergent scene sampling, as this simply makes the data input richer, improving performance regardless of which cycle-consistency method is used. We find that combining cycle variations, partial masking and Time Divergent Scene Sampling boosts the F1 matchings score of the previous SOTA by 4.3 percentage

Table 1: Cycle variations and partial masking together improve the overall matching performance by 2.5-2.1 percentage points. Every method benefits from time-divergent scene sampling, and combining everything boosts the previous SOTA by 4.3 percentage points, also improving stability.

Model	Standard			Time-Div. Scene Sampling		
	Precision	Recall	F1	Precision	Recall	F1
MvMHAT (Gan et al., 2021)	66.3	60.1	63.1±1.7	68.0	62.8	65.3±1.3
Cycle variations (CV)	68.8	**61.1**	64.7±1.9	70.4	62.3	66.1±1.4
CV + Partial masking	**71.0**	61.0	**65.6±1.1**	71.7	63.6	**67.4±0.9**

Table 2: Results per scene. Our methods improve the average F1 score on every scene. Crowded challenging test scenes like **Ground**, **Side** and **Shop** benefit most, with improvements of 9.1, 5.6 and 4.7 percentage points respectively.

Methods	Gate2	Square	Moving	Circle	Gate1
MvMHAT (Gan et al., 2021)	88.1	73.3	73.1	67.4	67.2
Ours w\o Masking	**88.3(+0.2)**	**74.9(+1.6)**	74.9(+1.8)	68.7(+1.3)	69.6(+2.4)
Ours	**88.3(+0.2)**	**74.9(+1.6)**	**76.2(+3.1)**	**69.9(+2.5)**	**70.4(+3.2)**

Methods	Floor	Park	Ground	Side	Shop
MvMHAT (Gan et al., 2021)	64.7	58.2	56.9	56.0	42.1
Ours w\o Masking	65.2(+0.5)	58.4(+0.2)	64.5(+7.6)	58.9(+2.9)	45.5(+3.4)
Ours	**66.8(+2.1)**	**60.4(+2.2)**	**66.0(+9.1)**	**61.6(+5.6)**	**46.8(+4.7)**

points, and that this combination is also the most consistent of all approaches. To put the results of Table 1 in perspective, we report that the F1 matching score of a Resnet pretrained on ImageNet is 16.8, while a supervised SOTA Re-ID model (Ye et al., 2021) with an optimized network architecture and hard negative mining is able to obtain a matching score of 82.28. This illustrates the strength of self-supervised cycle-consistency in general, showcasing its ability to significantly improve the feature quality of a pretrained ResNet. It also shows that our unoptimized self-supervised method is not to far from an optimized supervised baseline.

5.1.1 Results per Scene

The 10 scenes in the train and test data provide different challenges. During training, scenes with little overlap provide a worse learning signal for the overall model. During testing, scenes that require few matchings between many people are significantly more challenging. Insights into the overlap and number of people per scene is provided in the supplementary materials[1]. The scenes Ground, Side and Shop contain the highest number of people, around 24-32 per frame on average. The scenes Side and Shop also have little overlap, so that few matches needed to be correctly found from many possible ones. These scenes can thus be considered as the most challenging test set scenes. Table 2 reports the matching results per scene. Our methods outperform (Gan et al., 2021) on every test set, with the largest (relative) gains on

Ground, Side and Shop, with 9.1, 5.6 and 4.7 percentage points, respectively, highlighting the improved expressiveness of our feature network.

5.1.2 Partial Overlap Experiments

We experiment with artificially reducing the field of view in the training data by 20-40%. We implement this by reducing the actual width of each camera view starting from the right, throwing away the bounding boxes outside this reduced field of view. We train on these reduced overlap datasets and measure the robustness for each method, because self-supervision through cycle-consistency learns from overlap. An overlap analysis for the original and reduced datasets is provided in Table 3, and the evaluation results when training with the reduced data are shown in Table 4. We observe that our method is robust and contributes to the performance even in these harder training scenarios.

5.1.3 Cycle Variations Ablation

Our cycle variations use Equations 11- 14 to construct multiple trainable cycles to obtain a richer learning signal. We perform an ablation study on the effectiveness of each cycle, with and without masking, in Table 5. We find that our new $A_{ijk}A_{ki}$ and $A_{ijk}A_{kij}A_{jki}$ cycles from Equations 13 and 14 perform well on their own and even better when combined with the cycles from Equations 11 and 12. We observe that multiple cycle variations work especially well in the pres-
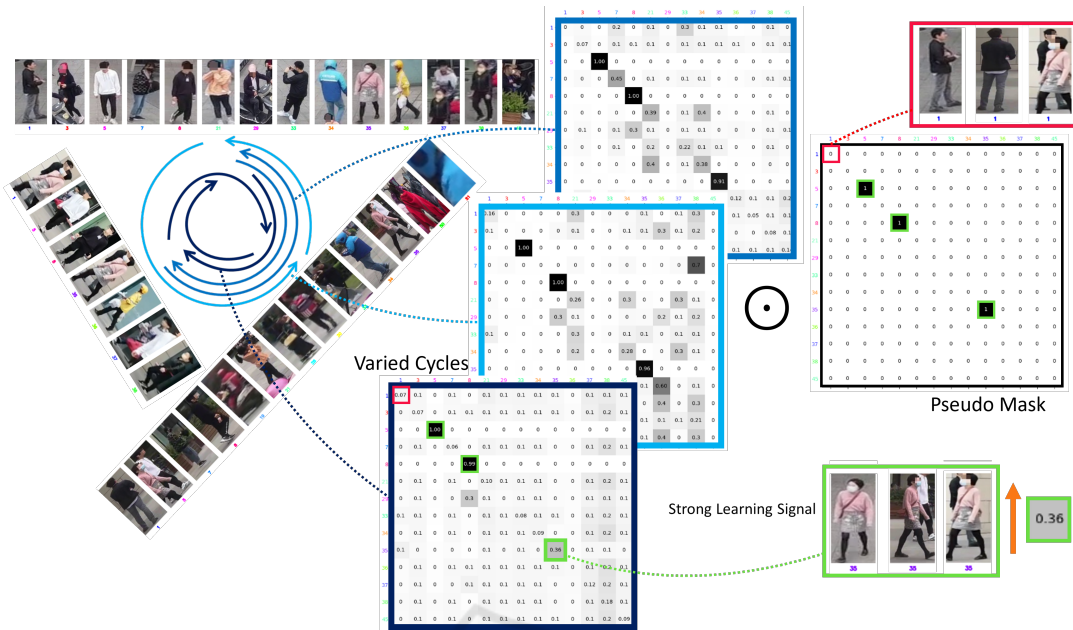
Figure 3: Qualitative example during training. Each of the blue swirls, representing Equations 12-14, constructs a cycle matrix with various cycle-inconsistencies. Partial overlap requires that only some of the diagonal elements are trained as cycles. The pseudo-mask correctly finds the existing cycles, except for a heavily occluded one. A strong learning signal is obtained from one of the diagonals of the dark blue cycle.

Table 3: The original train dataset has an average of 40% IoU between any two cameras, 26% people visible in all three cameras, and 18.4 unique people per frame. We reduce the FOV to simulate harder train data with less overlap.

Jaccard Index	Full Train|Test	80% Train Overlap	60% Train Overlap
Two Cameras	0.40|0.38	0.37	0.29
Three Cameras	0.26|0.23	0.24	0.15
Num People	18.4|19.4	16.5	14.0

Table 4: Our methods consistently improve performance, even with sparser training data that is reduced in partial overlap.

	Full Train	80% Train Overlap	60% Train Overlap
Methods	test set F1 score		
MvMHAT (Gan et al., 2021)	63.1±1.7	60.6±1.6	55.0±2.3
Ours w\o Masking	66.1±1.4	63.0±1.9	56.5±2.3
Ours	**67.4±0.9**	**63.8 ± 1.2**	**57.9 ± 1.5**

ence of masking, showing that these methods partly complement each other.
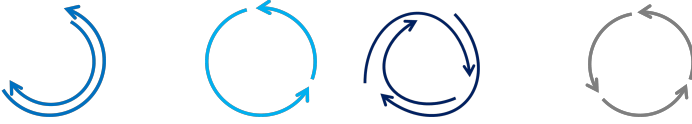
5.2 Qualitative Results

Figure 3 illustrates the contribution of the various cycles and pseudo-mask during training. In this specific example, it can be seen how the varied cycle constructions are cycle-inconsistent in different ways. Consequently, a robust learning signal is obtained from combining these cycle variants. The figure also shows the pseudo-mask I_{ijki} that is constructed for this batch,

where the existing cycles are correctly found with the exception of a severely occluded one in the top left, which you would not want to train anyway. The low value of 0.36 on the diagonal of the dark blue cycle means that a strong self-supervised learning signal is obtained from the masking, forcing the model to output more similar features for the different views of the person in pink.
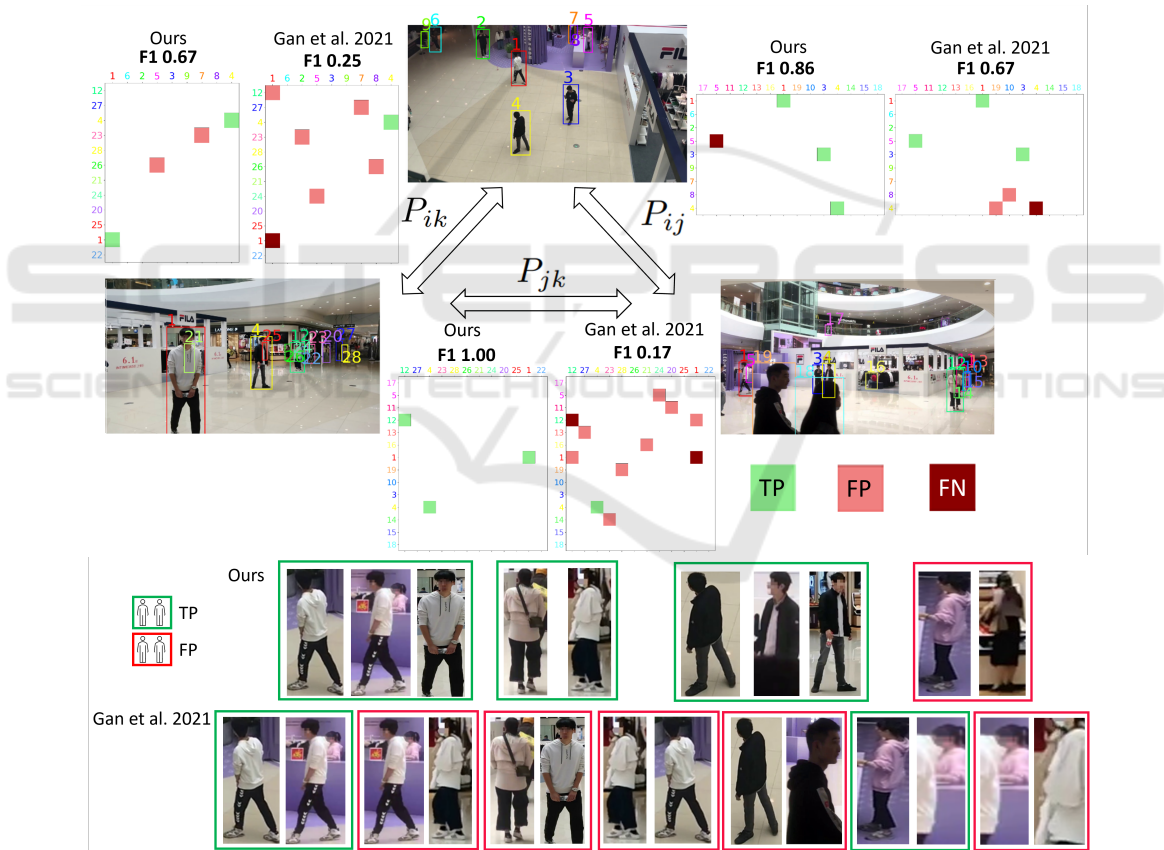
Figure 4 provides insight into the test set matching performance of our model compared to (Gan et al., 2021). It shows how our model effectively finds the pairwise matches at test time in a crowded

Table 5: Ablation of the cycle variations, also linking the Equations with illustrations. Our new $A_{ijk}A_{ki}$ and $A_{ijk}A_{kij}A_{jki}$ cycles work well individually, and even better in combination with the other cycle variations. Partial masking is also most effective when combined with multiple cycle variations. Our final method uses the setup from the bottom row with masking.

$A_{ijk}A_{kji}$ Eq 12 (Gan et al., 2021)	$A_{ijk}A_{ki}$ Eq 13 [Ours]	$A_{ijk}A_{kij}A_{jki}$ Eq 14 [Ours]	$A_{ij}A_{jk}A_{ki}$ Eq 11 (Feng et al., 2024)	w\o Masking	with Masking
✓				65.1 ± 0.9	66.7 ± 0.6
	✓			$\mathbf{66.4 \pm 1.0}$	66.2 ± 1.5
		✓		65.6 ± 1.8	66.4 ± 1.2
			✓	57.7 ± 1.5	55.9 ± 1.2
✓	✓			65.6 ± 1.5	66.7 ± 0.9
✓	✓	✓		66.2 ± 1.2	66.9 ± 0.7
✓	✓	✓	✓	66.3 ± 1.0	$\mathbf{67.2 \pm 1.1}$

Figure 4: Qualitative example during matching inference for a difficult frame in the test set. Our model is able to match with significantly fewer false positives. The matches found with our method are based on subtle clothing details, and have been correctly found in the presence of significant view angle differences and occlusion, significantly improving over the previous SOTA

scene. Note the difficulty of the matching problem, and how our method has significantly fewer false positive matches. The figure also demonstrates that our method is able to match significantly different representations of the same person across cameras, while differentiating between very similar looking people based on subtle clothing details.

6 CONCLUSIONS

We have extended the mathematical formulation of cycle-consistency to partial overlaps between views. We have leveraged these insights to develop a self-supervised training setting that employs multiple new cycle variants and a pseudo-masking approach to steer the loss function. The cycle variants expose different cycle-inconsistencies, ensuring that the self-supervised learning signal is more diverse and therefore stronger. We also presented a time divergent batch sampling approach for self-supervised cycle-consistency. Our methods combined improve the cross-camera matching performance of the current self-supervised state-of-the-art on the challenging DIVOTrack benchmark by 4.3 percentage points overall, and by 4.7-9.1 percentage points for the most challenging scenes.

Our method is effective in other multi-camera downstream tasks such as Re-ID and cross-view multi-object tracking. One limitation of self-supervision with cycle-consistency is its dependence on bounding boxes in the training data. Detections from an untrained detector could be used to train with instead, but this would likely degrade performance. Another area for improvement is to take location and relative distances into account both during training and testing, as this provides informative identity information.

Self-supervision through cycle-consistency is applicable to many more settings than just learning view-invariant object features. We believe the techniques introduced in this paper also benefit works that use cycle-consistency to learn image, patch, or keypoint features from videos or overlapping views.

REFERENCES

Bastani, F., He, S., and Madden, S. (2021). Self-supervised multi-object tracking with cross-input consistency. *Advances in Neural Information Processing Systems*, 34:13695–13706.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Dong, J., Jiang, W., Huang, Q., Bao, H., and Zhou, X. (2019). Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2019). Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810.

Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18.

Feng, W., Wang, F., Han, R., Qian, Z., and Wang, S. (2024). Unveiling the power of self-supervision for multi-view multi-human association and tracking. *arXiv preprint arXiv:2401.17617*.

Gan, Y., Han, R., Yin, L., Feng, W., and Wang, S. (2021). Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 282–290.

Han, R., Wang, Y., Yan, H., Feng, W., and Wang, S. (2022). Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing*, 31:1830–1840.

Hao, S., Liu, P., Zhan, Y., Jin, K., Liu, Z., Song, M., Hwang, J.-N., and Wang, G. (2023). Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, pages 1–16.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, Q.-X. and Guibas, L. (2013). Consistent shape maps via semidefinite programming. *Computer graphics forum*, 32(5):177–186.

Jabri, A., Owens, A., and Efros, A. (2020). Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560.

Li, M., Zhu, X., and Gong, S. (2019). Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782.

Loy, C. C., Xiang, T., and Gong, S. (2010). Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90:106–129.

Ristani, E. and Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Sun, S., Akhtar, N., Song, H., Mian, A., and Shah, M. (2019). Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):104–119.

Wang, X., Jabri, A., and Efros, A. A. (2019). Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576.

Wang, Z., Zhang, J., Zheng, L., Liu, Y., Sun, Y., Li, Y., and Wang, S. (2020). Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 72–88. Springer.

Wieczorek, M., Rychalska, B., and Dąbrowski, J. (2021). On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 212–223. Springer.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893.

Zhao, J., Han, R., Gan, Y., Wan, L., Feng, W., and Wang, S. (2020). Human identification and interaction detection in cross-view multi-person videos with wearable cameras. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2608–2616.