Flowing Crowd to Count Flows: A Self-Supervised Framework for Video Individual Counting

Feng-Kai Huang
National Taiwan University
Taipei, Taiwan
leonelhuang@cmlab.csie.ntu.edu.tw

Bo-Lun Huang
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
kevin503.ee12@nycu.edu.tw

Li-Wu Tsao
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
lwtsao.ee09@nycu.edu.tw

Jhih-Ciang Wu
National Taiwan Normal University
Taipei, Taiwan
jcwu@csie.ntnu.edu.tw

Hong-Han Shuai*
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
hhshuai@nycu.edu.tw

Wen-Huang Cheng
National Taiwan University
Taipei, Taiwan
wenhuang@csie.ntu.edu.tw

Abstract

Video Individual Counting (VIC), which seeks to count unique individuals across video sequences without duplication, has broader applications than traditional Video Crowd Counting (VCC), including urban planning, event management, and safety monitoring. However, although current VIC approaches have demonstrated strong capabilities, their reliance on identity-level or group-level annotations necessitates substantial labeling effort and expense. To reduce the high costs of manual annotation, we introduce VIC-SSL, a novel self-supervised learning approach that utilizes unlabeled data along with the innovative feature-level augmentation technique called Foreground-driven ShiftMix (F-ShiftMix). By blending and shifting in the feature space rather than the image space, F-ShiftMix generates realistic crowd motion without explicit annotations, while preserving global semantic coherence. Furthermore, VIC-SSL integrates the Cost-guided Flow Prompt (CFP) and the Distinction-aware Cross-Attention (DCA) to enhance flow-aware localization and inter-frame correspondence learning. Our extensive experiments across three datasets, including SenseCrowd, CroHD, and CARLA, demonstrate that VIC-SSL substantially outperforms existing methods, achieving state-of-the-art results with significantly reduced data requirements. These results showcase VIC-SSL's potential to dramatically lower annotation costs and improve the deployment feasibility of VIC systems in complex scenarios. The project website is available at https://leohuang0511.github.io/vic-ssl.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; **Computer vision problems**.

*Corresponding author, hhshuai@nycu.edu.tw

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755385

Keywords

Video Individual Counting; Self-Supervised Learning

ACM Reference Format:
Feng-Kai Huang, Bo-Lun Huang, Li-Wu Tsao, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng. 2025. Flowing Crowd to Count Flows: A Self-Supervised Framework for Video Individual Counting. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755385

1 Introduction

Video Individual Counting (VIC) has emerged as a significant extension of the traditional Video Crowd Counting (VCC) paradigm. Unlike VCC, which focuses on determining the number of pedestrians in individual frames, VIC extends this capability to count distinct individuals across a video sequence, ensuring no individual is counted more than once. This distinction not only enhances accuracy but also improves the utility and applicability of crowd counting technologies in practical settings. For instance, VIC can be applied to monitor the number of people passing through a specific region, providing valuable data for footfall analysis [14], which is crucial for urban planning, event safety monitoring, crowd management, security surveillance, and public transportation optimization.

Multi-object tracking (MOT) methods [18, 43, 55] have been employed in VIC tasks to maintain consistent pedestrian identities across frames. However, these methods encounter challenges such as identity switches and tracking inaccuracies, particularly in scenarios with dense crowds and significant occlusions. These limitations hinder their practical deployment. Addressing these issues, recent advancements segment the VIC problem into estimating initial pedestrian counts and predicting subsequent inflows, with approaches like DRNet [11], PDTR [21], and FMDC [39] employing strategies that range from utilizing head descriptors in density maps to localization regression and inflow mask integration. CGNet [25] further reduces annotation demands through a weakly supervised model that utilizes identity-agnostic labels to distinguish between different pedestrian inflows, easing scalability constraints.

Despite these advances, labeling in crowded scenarios remains costly and labor-intensive. Furthermore, the robustness and transferability of these models still require massive datasets. We thus

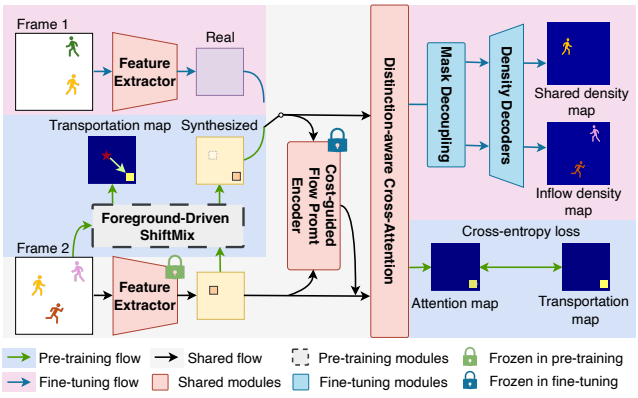

Figure 1: Overview of VIC-SSL. Foreground-driven ShiftMix simulates crowd motion by modifying feature maps, generating transportation maps that enable Cost-guided Flow Prompt to learn flow-aware localization during pretraining. While Distinction-aware Cross-Attention learns the inter-frame correspondences. This approach enhances downstream VIC result and reduces reliance on labeled data.

pose the question: "Can we automatically generate effective training data for VIC without exhaustive manual annotations?" Our core insight is to bypass laborious labeling by synthetically generating frame-to-frame correspondences. We achieve this by realistically *moving* individuals in a scene to create the subsequent frame. A naive approach is to simply shift an entire image to simulate flow. However, such flow is homogeneous, *i.e.*, all individuals move in the same direction, which diverges from realistic scenarios. Another potential strategy is to leverage an off-the-shelf object detector and adopt a copy-and-paste strategy [9] to cut-and-paste individuals. Yet, this approach can unintentionally include unrelated people or objects within bounding boxes and demands computationally expensive image inpainting to repair areas vacated by moved entities.

To overcome these annotation and data scarcity challenges, we propose a novel self-supervised learning strategy, VIC-SSL. Our approach drastically reduces reliance on labeled data by exploiting the inherent information present within unlabeled video sequences. Figure 1 shows the overview of our VIC-SSL. At the heart of VIC-SSL lies a novel feature-level data augmentation technique called Foreground-driven ShiftMix (F-ShiftMix). This method synthesizes realistic crowd motions by blending and shifting high-level feature representations, thereby generating semantically coherent synthesized reference feature maps. Unlike traditional image-level augmentations, our F-ShiftMix maintains global semantic integrity and simulates realistic crowd dynamics, which is essential for training robust and reliable VIC models without explicit annotations.

Furthermore, we introduce the Cost-guided Flow Prompt (CFP) and the Distinction-aware Cross-Attention (DCA) to enhance inter-frame correspondence learning. CFP dynamically generates a motion-aware prompt by computing a localized cost volume, effectively guiding the model's attention toward relevant regions exhibiting significant inter-frame motion. Concurrently, DCA explicitly captures the subtle yet significant distinctions between consecutive frames, thereby enhancing the model's capability to detect and interpret pedestrian inflows accurately. Collectively, these modules equip VIC-SSL with powerful self-supervised capabilities, allowing the model to learn temporal correspondences directly from unlabeled data. The contributions are summarized as follows.

- We introduce VIC-SSL, a self-supervised learning strategy for Video Individual Counting (VIC) that significantly reduces the dependency on labeled data by exploiting unlabeled video sequences with the proposed data augmentation method, Foreground-driven ShiftMix (F-ShiftMix).
- We propose a new architecture for VIC-SSL, including two new modules, the Cost-guided Flow Prompt (CFP) and the Distinction-aware Cross-Attention (DCA), which enhance the model's capability to learn accurate temporal correspondences and adapt to complex crowd dynamics.
- Experimental results on the SenseCrowd, CroHD and CARLA datasets show that VIC-SSL outperforms the state-of-the-art methods by at least 14.6% in terms of MSE.

2 Related Work

2.1 Video Individual Counting

Video Individual Counting (VIC) extends conventional Video Crowd Counting (VCC) [13, 19, 24, 29, 45, 47], which counts pedestrians in each frame, into a more valuable task that ensures no individual is counted more than once. Multi-object tracking (MOT) methods [2, 18, 34, 41, 43, 54, 55] often suffer from ID switches in crowded or occluded scenes. To overcome this, recent methods [11, 21, 25, 39] decompose VIC into initial counting and inflow estimation across frames. For instance, DRNet [11] uses head descriptors for inflow detection, while PDTR [21] and FMDC [39] reformulate inflow prediction via localization regression or mask integration. CGNet [25] reduces annotation cost using group-level labels, yet still requires manual supervision. In contrast, our VIC-SSL captures inter-frame correspondences from unlabeled video, reducing reliance on annotations while maintaining performance.

2.2 Self-supervised Representation Learning

Self-supervised learning (SSL) learns from unlabeled video using pretext tasks that exploit spatiotemporal cues. Video-level SSL [10, 20, 40, 52] reconstructs or models global motion across multiple frames. Masked modeling methods [33, 36] learn general temporal features, while contrastive learning [7, 8, 26, 44] enforces temporal consistency between clips. However, these approaches require careful sampling and overlook fine-grained motion. In contrast, frame-level SSL [4, 37, 58] learns short-term correspondences for tasks like tracking or flow estimation. Techniques based on cycle-consistency [42] or motion coherence [1] aim to recover reliable pairwise mappings but are limited to simple scenes. Our approach targets challenging scenarios involving dense crowds, characterized by subtle inter-frame variations and frequent occlusions.

To support this, we employ feature-level data augmentation. While image-level strategies [16, 48–50] like CutMix [53] and jigsaw [3, 5, 6] improve spatial robustness, they fail to simulate realistic temporal dynamics. Recent works on feature-level augmentation [12, 27, 31, 46, 51] like TokenMix [23] and AutoMix [28] enable semantically coherent mixing in ViT-based models. We extend

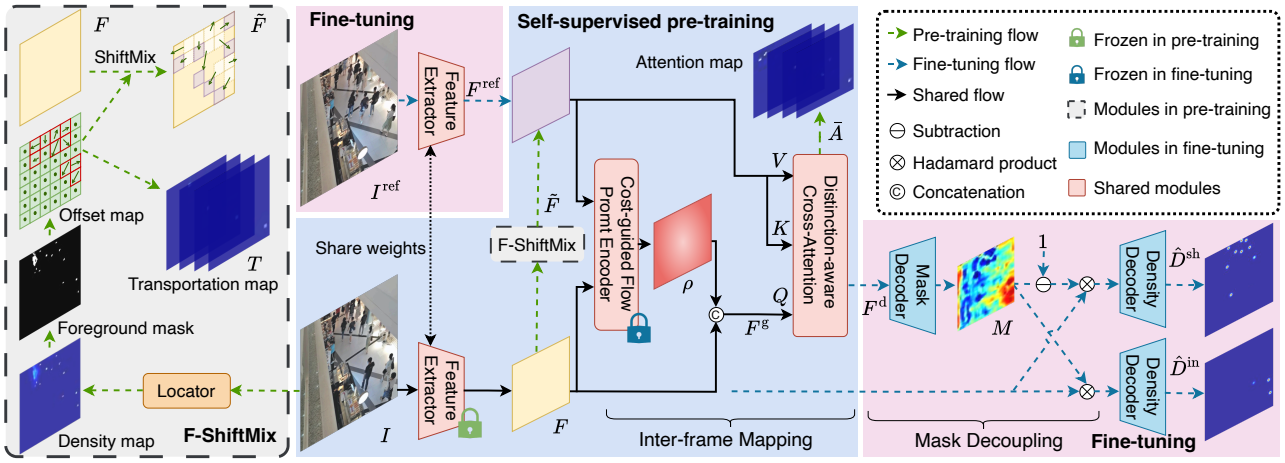

Figure 2: Architecture of VIC-SSL. VIC-SSL consists of two stages: self-supervised pre-training and fine-tuning. In pre-training, a feature extractor processes an input frame I, and the F-ShiftMix module simulates crowd motion by blending and shifting this feature map F to create a pseudo-reference feature map $\tilde{F}$. A Cost-guided Flow Prompt (CFP) ρ is then generated, guiding the Distinction-aware Cross-Attention (DCA) to emphasize frame differences. During fine-tuning, a real reference frame is used. After distinction-aware features are extracted by DCA, a decoupling mask M generated by the mask decoder identifies the distinct regions and decouples the feature map, predicting inflow and shared crowd densities by the density decoders.

these ideas temporally by applying structured shifts and blending in the feature space across frames, allowing it to synthesize motion-consistent representations for learning without manual labels.

3 Method

3.1 Overview

Given a video $\mathcal{I} = \{I_t\}_{t=0}^{L-1}$ of L frames, VIC estimates the total number of unique individuals $\hat{N}$. Following [11], the task involves estimating the initial count n_0 and the subsequent inflows $n^{\text{in}}_{t,t+\tau}$ between frames I_t and $I_{t+\tau}$ over a temporal gap τ. VIC-SSL learns fine-grained inter-frame correspondences from unlabeled crowd videos, significantly reducing annotation requirements. As illustrated in Figure 2, VIC-SSL operates in two stages: self-supervised pre-training and fine-tuning. During pre-training, a feature extractor processes each frame $I \in \mathbb{R}^{H\times W\times 3}$ into a feature map $F \in \mathbb{R}^{h\times w\times c}$, where $h = H/16$ and $w = W/16$. The F-ShiftMix simulates motion by blending and shifting F to generate a reference feature map $\tilde{F} \in \mathbb{R}^{h\times w\times c}$ and its corresponding transportation map. A Cost-guided Flow Prompt Encoder then computes a localized cost volume between F and $\tilde{F}$ to produce a CFP, which is concatenated with F and passed to the DCA module. DCA performs distinction-aware cross-attention to model inter-frame correspondences, supervised by the transportation map, thus avoiding manual labels.

In the fine-tuning stage, a real reference frame $I^{\text{ref}} \in \mathbb{R}^{H\times W\times 3}$ is introduced. Its feature map $F^{\text{ref}} \in \mathbb{R}^{h\times w\times c}$ is paired with F, and DCA, guided by CFP, extracts distinction-aware features. These are processed by a mask decoder to predict a decoupling mask $M \in \mathbb{R}^{h\times w}$ that highlights inflow regions. The masked feature maps $F \otimes M$ and $F \otimes (1-M)$ are passed through two separate density decoders to produce the inflow density map $\hat{D}^{\text{in}}$ and the shared density map $\hat{D}^{\text{sh}}$, respectively. These maps estimate newly appearing individuals and those persisting across frames, completing the VIC prediction pipeline. In the following, we present (i) the design of F-ShiftMix, (ii) the CFP and DCA modules, and (iii) the training objectives employed in both the pre-training and fine-tuning stages.

3.2 Self-supervised Foreground-driven ShiftMix

To learn inter-frame correspondences without relying on manual annotations, we propose a novel approach to synthesize a reference feature map $\tilde{F}$ by applying random patch blending and shifting directly on the highest-level feature map F. This design choice is motivated by the limitations of performing such operations at the image level, where blending and shifting patches can fragment crowded scenes and disrupt visual continuity. By augmenting at the feature level [12, 27, 46], we leverage the abstracted representations in the given frame I, which encapsulate semantic information to ensure that the synthesized reference map $\tilde{F}$ emulates realistic crowd motion while preserving global semantics.

The procedures for generating reference feature maps are designed to imitate the movements observed in the VIC task, with each operation addressing distinct aspects of crowd dynamics. The blending operation combines patches from different spatial locations within the feature map, effectively simulating the interactions and overlaps that are characteristic of crowded scenes. This operation introduces spatial diversity, encouraging the model to learn robust representations that account for complex spatial relationships. Formally, the blending operation can be expressed as

$$\tilde{F}_{i'j'} = F_{i'j'} + F_{ij}, \tag{1}$$

where $i' = i + \Delta y_{ij}$ and $j' = j + \Delta x_{ij}$. The magnitudes $(\Delta y_{ij}, \Delta x_{ij})$ for (i, j) are spatial variations. We note that the updated position is restricted within the boundary, *i.e.*, $i' < h$, and $j' < w$ are

satisfied for each location. Precisely, elements that move beyond the boundary are removed from $\tilde{F}$ and do not blend into any others.

The subsequent shifting operation presents temporal variations by removing small regions from the original feature map. This process mimics the natural movement of individuals in a crowd, capturing the essential temporal dynamics needed for learning inter-frame correspondences. The formulation for updating the feature in place through a shifting operation is defined as

$$\tilde{F} \leftarrow \tilde{F} - F. \tag{2}$$

These operations in (1) and (2) create a diverse and realistic set of augmented feature maps, allowing the model to comprehend and generalize the spatial-temporal complexities inherent in the VIC task. To accurately capture the simulated movements, we define a pseudo label to represent the concept of *transportation* in this pretext task, which is formulated as

$$T_{ijkl} = \begin{cases} 1, & \text{if } (k,l) = (i',j') \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $T_{ijkl} \in \mathbb{R}^{h\times w\times h\times w}$ (the expression is similar to [35]) represents the mapping between the original feature map F and the augmented reference feature map $\tilde{F}$. Specifically, $T_{ijkl} = 1$ indicates that the element at position (i, j) in F is transported to position (k, l) in $\tilde{F}$. To ensure consistency, the constraint $\sum_{k=0}^{h-1}\sum_{l=0}^{w-1} T_{ijkl} = 1$ holds everywhere, guaranteeing that each element in F corresponds to exactly one position in $\tilde{F}$. This transportation map explicitly encodes how each element in the original feature map is shifted or blended to its counterpart in the reference feature map.

The most straightforward approach to synthesizing motion is to depict a shifting ratio determining the proportion of elements in F to be randomly shifted and blended. However, this naive approach raises two influential issues that fail to reflect realistic crowd motion. We refer to these two subjects as *Random Shifting* and *Unbounded Offsets*. The former comes from substantial displacements typically originating from moving individuals (foreground) rather than static background elements. Randomly shifting all elements without distinguishing between foreground and background regions disrupts the semantic coherence of the synthesized motion. Besides, the latter is due to the individuals in a crowd rarely exhibiting large, abrupt movements. Allowing unbounded offsets may result in unrealistic displacements that deviate from genuine crowd dynamics. To address these limitations, we propose a foreground-driven shifting strategy and introduce bounded offsets, as elaborated in the following sections. These enhancements ensure that the proposed synthesized motion more accurately emulates real-world crowd behavior while maintaining spatial-temporal consistency.

3.2.1 Foreground-driven Shifting Strategy. We propose a foreground-driven shifting strategy that treats the foreground and background regions of the feature map differently by applying distinct shifting ratios to better approximate real-world motion. As illustrated in the left part of Figure 2, the input image I is first processed using an off-the-shelf locator [22] to generate a crowd density map. This density map is then binarized to produce a foreground mask that highlights the crowd regions and resized to align with the offsets grid, enabling us to assign higher shifting ratios to the foreground elements, which simulate the movement of pedestrians, and lower

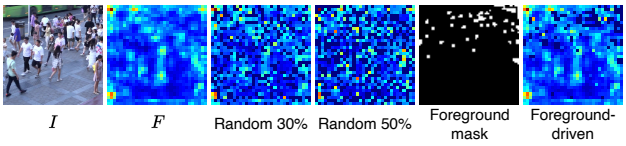

Figure 3: Visualization of different shifting strategies. Randomly shifting does not correspond to movements in a natural scene, while our F-ShiftMix blends and shifts the features of foreground and background with different ratios.

shifting ratios to the background elements, which represent relatively static objects. This strategy ensures that the synthesized motion more closely reflects the dynamics of crowded scenes. The examples of different shifting strategies are shown in Figure 3.

3.2.2 Bounded Offsets. To maintain realism and avoid large, unrealistic displacements, we introduce bounded offsets to constrain the magnitude of the shifts. Specifically, we predefine a radius r to limit the range of the randomly sampled offsets $\Delta y_{ij}, \Delta x_{ij} \sim U(-r, r)$, enforcing the constraints $|\Delta y_{ij}| \le r$ and $|\Delta x_{ij}| \le r$. This restriction ensures that the generated displacements remain within a plausible range, preserving the authenticity of the simulated crowd motion.

By integrating the foreground-driven shifting strategy with bounded offsets, our proposed F-ShiftMix effectively leverages an unlabeled frame to emulate realistic crowd motion. By preserving spatial-temporal consistency and mimicking real-world crowd dynamics, F-ShiftMix provides a robust foundation for self-supervised learning. The next section elaborates on how the model utilizes this enhanced reference frame to perform inter-frame mapping, enabling the capture of accurate spatial-temporal correspondences.

3.3 Inter-frame Mapping

In the VIC task, models are asked to predict pedestrian inflow between consecutive frames by capturing shared and distinct spatio-temporal cues, requiring precise inter-frame correspondence modeling. To achieve this, we employ a cross-attention mechanism to correlate the original feature map and the synthesized reference map through query-key interaction. However, standard cross-attention relies solely on static feature similarities, limiting its ability to capture dynamic motion. As a result, we introduce CFP, a conditional and learnable flow prompt $\rho \in \mathbb{R}^{h\times w\times g}$ that dynamically adapts to inter-frame motion by conditioning on the cost map between F and $\tilde{F}$. Additionally, conventional cross-attention operation struggles to extract distinguishing features between queries and keys when tackling VIC, which is critical for identifying unique spatio-temporal signals. To resolve this, we propose DCA, which models differences in representations to capture distinct information well.

3.3.1 Cost-guided Flow Prompt. Motivated by optical flow techniques [15, 35, 56], we leverage the cost volume between F and $\tilde{F}$ to guide the learning of the flow prompt, as shown in Figure 4 (a). While the cost volume measures the similarity between every element of F and $\tilde{F}$, the actual crowd motion rarely spans such large spatial ranges, and computing the dense matrix is computationally expensive. Consequently, we employ a radius γ and derive $p, q = \{0, 1, \ldots, 2\gamma\}$ to compute a localized cost volume

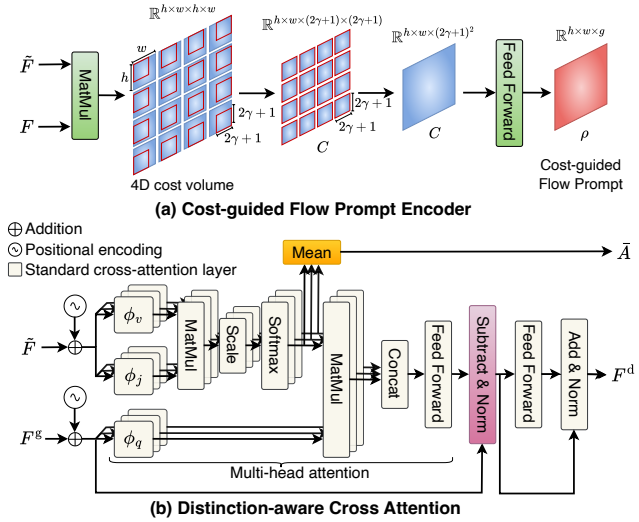

Figure 4: Illustration of (a) the Cost-guided Flow Prompt Encoder and (b) the Distinction-aware Cross-Attention (DCA). The Cost-guided Flow Prompt Encoder constructs a flow prompt conditioned on the localized cost volume between two feature maps. The DCA captures inter-frame correspondences and further extracts distinction-aware information.

$C \in \mathbb{R}^{h\times w\times(2\gamma+1)\times(2\gamma+1)}$. The description of the rigor for indexing the component in C is defined as

$$C_{ijpq} = \sum_{d=0}^{c-1} \frac{F_{ijd} \cdot \tilde{F}_{kld}}{\sqrt{c}}, \tag{4}$$

where $k = i - \gamma + p$ and $l = j - \gamma + q$. We reshape C and pass it through a feed-forward layer to encode the flow prompt ρ. The flow prompt is then concatenated with F, resulting in the guided feature map $F^{\mathrm{g}} = \mathrm{Concat}(F, \rho) \in \mathbb{R}^{h\times w\times(c+g)}$. The guided feature map directs the subsequent cross-attention mechanism to focus on regions with significant motion, thereby improving temporal coherence and enhancing prediction accuracy.

3.3.2 Distinction-aware Cross-attention. To capture correspondences between F^{g} and $\tilde{F}$ while emphasizing distinct information, we employ the DCA mechanism, as illustrated in Figure 4 (b). Specifically, we flatten F^{g} and $\tilde{F}$ into shapes $hw \times (c+g)$ and $hw \times c$, respectively. We compute multi-head cross-attention (MCA) and denote the attention map at s-th head as

$$A_s = \mathrm{softmax}\left(\frac{Q_s K_s^{\mathsf{T}}}{\sqrt{c}}\right), \tag{5}$$

where $Q_s = \phi_q(F^{\mathrm{g}})$ and $K_s = \phi_k(\tilde{F})$. Here, ϕ represents a combination of positional embedding and linear projection, simplifying the attention mechanism. The completed MCA is obtained by multiplying A_s from (5) with $V_s = \phi_v(\tilde{F})$, resulting in the representation z_s. The MCA can be expressed as $z = \mathrm{MCA}(Q, K, V)$, where z is derived by concatenating the outputs of all attention heads.

As we aim to emphasize distinct information, we modify the traditional cross-attention layer by replacing the original residual connection with a subtraction operation $z' = Q - z$, formulated as

$$F^{\mathrm{d}} = \mathrm{LN}\big(\mathrm{FFN}(\mathrm{LN}(z')) + z'\big), \tag{6}$$

where FFN is a feed-forward network, LN denotes LayerNorm, and F^{d} represents the distinction-aware feature map. This design allows the model to explicitly extract the differences between $\tilde{F}$ and F^{g}, improving its ability to identify newly appearing individuals.

3.4 Two-stage Training

We detail our two-stage training protocol, integrating the F-ShiftMix, CFP, and DCA designs introduced earlier. The self-supervised pre-training phase focuses on learning robust inter-frame correspondences from unlabeled data, while the fine-tuning phase leverages a few labeled frame pairs to refine the network specifically for VIC.

3.4.1 Self-supervised Pre-training. As depicted by the green and black arrows in Figure 2, our self-supervised pre-training pipeline teaches the model to capture correspondences between the feature map of I and the pseudo-reference feature map generated by F-ShiftMix. To maintain consistent feature representations and avoid training collapse, we freeze the feature extractor during this phase. We then employ the transportation map T from (3) as supervision for the attention map A_s in (5). Specifically, we compute the mean of all multi-head attention maps (we denote as $\bar{A}$) and reshape both $\bar{A}$ and T into $hw \times hw$. The pre-training loss is formulated as a weighted Cross-Entropy (CE) loss:

$$\mathcal{L}^{\mathrm{PT}} = -\frac{1}{hw}\sum_{a=1}^{hw}\sum_{b=1}^{hw} \Omega_{ab} \cdot T_{ab} \cdot \log \bar{A}_{ab}, \tag{7}$$

where the matrix $\Omega \in \mathbb{R}^{hw\times hw}$ lessens the influence of diagonal elements—stationary features under F-ShiftMix—and is defined by

$$\Omega_{ab} = \begin{cases} \omega, & \text{if } a = b, \\ 1, & \text{otherwise}, \end{cases} \tag{8}$$

where $\omega < 1$. By optimizing $\bar{A}$ to match T, the model leverages CFP to guide DCA in a self-supervised manner, thereby improving its ability to align corresponding features in downstream VIC tasks.

3.4.2 Fine-tuning. Once the model has learned robust inter-frame correspondences, we fine-tune it for the specific goal of VIC. As shown by the blue and black arrows in Figure 2, we now sample a frame pair (I, I^{ref}) separated by a fixed time interval. Instead of the pseudo-reference $\tilde{F}$ from pre-training, we feed the real reference feature map F^{ref} extracted from I^{ref}. The distinction-aware feature F^{d} between F and F^{ref} is then produced via DCA, guided by CFP. Notably, the Cost-guided Flow Prompt Encoder is frozen here to preserve its learned motion awareness (further discussed in Section 4.3). Next, we feed F^{d} into a mask decoder—two residual blocks plus a 1×1 convolution and a sigmoid—to produce a decoupling mask $M \in \mathbb{R}^{h\times w}$. This mask highlights newly appearing or distinct regions and splits F into two streams via Hadamard products: $F \otimes M$ and $F \otimes (1 - M)$. We then pass these into two density decoders to estimate the inflow density map $\hat{D}^{\mathrm{in}}$ and the shared density map $\hat{D}^{\mathrm{sh}}$, respectively. Both maps are supervised by ground-truth inflow/shared density labels (D^{in}, D^{sh}), generated

Table 1: Performance comparison on SenseCrowd dataset. $\mathcal{D}_i$ denotes varying crowd density levels. Methods with the † symbol utilize group-level annotations. Bolded values represent the best performance, while underlines values indicate the second-best.

Method	Venue	Overall			Density (MAE)				
		MAE↓	MSE↓	WRAE(%)↓	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$
LOI [57]	ECCV'16	24.7	33.1	37.4	12.5	25.4	39.3	39.6	86.7
HeadHunter-T [34]	CVPR'21	30.0	50.6	38.6	11.8	25.7	56.0	92.6	131.4
FairMOT [55]	IJCV'21	35.4	62.3	48.9	13.5	22.4	67.9	84.4	145.8
Deep OC-SORT [30]	ICIP'23	26.0	43.6	29.5	9.5	25.1	29.1	60.3	156.5
SMILEtrack [43]	AAAI'24	27.2	36.7	32.5	9.2	21.0	33.8	76.5	203.5
DRNet [11]	CVPR'22	12.3	24.7	12.7	4.1	8.0	23.3	50.0	77.0
FMDC [39]	WACV'24	16.6	36.5	16.5	5.8	10.6	26.6	52.6	125.1
CGNet† [25]	CVPR'24	8.9	17.7	12.6	5.0	**5.8**	**8.5**	25.0	63.4
PDTR [21]	MM'24	9.6	17.6	11.4	4.6	6.8	14.7	23.6	60.6
(Ours) w/o Pretrain†	-	8.2	14.3	10.7	3.7	7.4	14.3	24.8	31.8
(Ours) w/ Pretrain†	-	**7.6**	**12.6**	**10.4**	**3.6**	6.7	14.8	**15.4**	**30.8**

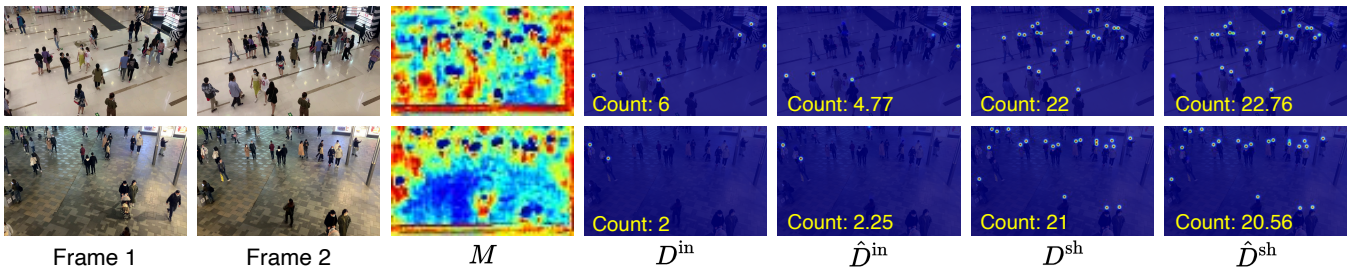

Figure 5: Visualization of VIC-SSL's predictions on SenseCrowd. We visualize the decoupling mask M, the ground-truth/predicted inflow density map $D^{\text{in}}/\hat{D}^{\text{in}}$ and the ground-truth/predicted shared density map $D^{\text{sh}}/\hat{D}^{\text{sh}}$.

by placing Gaussian kernels on head keypoints categorized at the group level [25]. The fine-tuning loss is thus defined as

$$\mathcal{L}^{\text{FT}} = ||\hat{D}^{\text{in}} - D^{\text{in}}||_2^2 + \alpha\, ||\hat{D}^{\text{sh}} - D^{\text{sh}}||_2^2, \tag{9}$$

where α balances the two terms.

3.4.3 Inference. Following prior work, we decompose VIC into counting the initial population n_0 and the inflow $n^{\text{in}}_{t,t+\tau}$ for subsequent intervals. Let I_t and $I_{t+\tau}$ be I^{ref} and I, respectively. Our model predicts the inflow density $\hat{D}^{\text{in}}_{t,t+\tau}$ and the shared density $\hat{D}^{\text{sh}}_{t,t+\tau}$, and calculates the total number of individuals in video $\mathcal{I}$ by

$$\hat{N} = n_0 + \sum_{k=1}^{L/\tau} n^{\text{in}}_{(k-1)\tau,k\tau} = \left(\hat{D}^{\text{in}}_0 + \hat{D}^{\text{sh}}_0\right) + \sum_{k=1}^{L/\tau} \hat{D}^{\text{in}}_{(k-1)\tau,k\tau}. \tag{10}$$

4 Experiments

4.1 Experiments Setting

Dataset. We evaluate our method on three datasets—SenseCrowd, CroHD, and CARLA—each offering diverse scenes and crowd densities to ensure robust performance evaluation. SenseCrowd [19] is a large-scale video crowd dataset comprising 634 video sequences and a total of 62,938 frames. It captures a wide range of scenarios with annotated attributes such as crowd density and spatial distributions. The dataset is split into training, validation, and testing sets following the standard VIC protocol [11, 39]. CroHD [34] features congestly crowd and contains 11,463 frames across 9 video sequences, with 4 sequences used for training and 5 for testing. CARLA [39] is a synthetic dataset composed of 10 video sequences captured under varying weather conditions and camera perspectives, evenly divided into 5 training and 5 testing videos.

Evaluation Metrics. Mean Absolute Error (MAE), Mean Square Error (MSE), and Weighted Relative Absolute Errors (WRAE) [11] are used for evaluation. MAE and MSE are widely used in traditional crowd counting, while WRAE proposes to balance the performance in various video lengths and pedestrian count.

Implementation Details. We adopt VGG-16 [32] pretrained on ImageNet [17] as the feature extractor. Our VIC-SSL is pre-trained on SenseCrowd for 10 epochs with a batch size of 3. We set the initial learning rate to 1×10^{-4} and apply a step decay of 0.95 each epoch. The shifting ratios for foreground and background elements are 100% and 1%, respectively. We fix the radius for bounded offsets and the localized cost volume to $r = \gamma = 7$, and use 8 attention heads. The weighting factor ω in $\mathcal{L}^{\text{PT}}$ is set to 0.1. During fine-tuning, we freeze Cost-guided Flow Prompt to preserve its learned dynamic guidance. We employ a batch size of 2. The learning rates for the

Table 2: Comparisons for VIC on CroHD.

Method	Venue	MAE↓	MSE↓	WRAE(%)↓
LOI [57]	ECCV'16	305.0	371.1	46.0
HeadHunter-T [34]	CVPR'21	253.2	351.7	32.7
FairMOT [55]	IJCV'21	256.2	300.8	44.1
PHDTT [38]	IW-FCV'22	2130.4	2808.3	401.6
Deep OC-SORT [30]	ICIP'23	165.2	195.9	33.1
SMILEtrack [43]	AAAI'24	181.6	235.1	36.1
DRNet [11]	CVPR'22	141.1	192.3	27.4
FMDC [39]	WACV'24	54.2	61.7	10.7
CGNet† [25]	CVPR'24	75.0	95.1	14.5
PDTR [21]	MM'24	60.6	73.7	12.7
(Ours) w/o Pretrain†	-	56.4	78.4	10.4
(Ours) w/ Pretrain†	-	**40.6**	**50.9**	**8.8**

Table 3: Comparisons for VIC on CARLA.

Method	Venue	MAE↓	MSE↓	WRAE(%)↓
HeadHunter-T [34]	CVPR'21	325.2	516.9	103.4
FairMOT [55]	IJCV'21	330.0	521.1	105.6
DRNet [11]	CVPR'22	86.1	105.9	33.2
FMDC [39]	WACV'24	59.1	90.0	20.7
(Ours) w/o Pretrain†	-	56.4	63.2	27.6
(Ours) w/ Pretrain†	-	**35.8**	**36.5**	**18.2**

remaining modules are 1×10^{-4}, 3×10^{-5}, and 1×10^{-6} when fine-tuning on SenseCrowd, CroHD, and CARLA, respectively, alongside a one-cycle scheduler with one epoch of warm-up. We set $\alpha = 1$ in $\mathcal{L}^{\text{FT}}$. Frame pairs are randomly sampled with time intervals from 1s to 5s during training and fixed at 3s in the inference stage.

Comparison Approaches. We compare our method against several representative approaches, which can be grouped into three categories. 1) The cross-line method, LOI [57] estimates pedestrian flow by counting individuals as they cross a predefined line. 2) Tracking-based methodd include HeadHunter-T [34], FairMOT [55], Deep OC-SORT [30] and SMILEtrack [43]. These methods detect and track individuals throughout the video, with the total count obtained by summing non-repetitive identity predictions across frames. 3) Approaches designed explicitly for the VIC task, such as DRNet [11], FMDC [39], CGNet [25] and PDTR [21], estimate per-frame inflow and accumulate these predictions, adding the number of individuals in the initial frame to produce the final count.

4.2 Quantitative Results

Results on SenseCrowd. Table 1 shows various models' performance on SenseCrowd, where VIC-SSL sets new state-of-the-art scores across multiple metrics. Even without pre-training, VIC-SSL surpasses CGNet, which also uses group-level labels, on all metrics. Once our self-supervised pre-training is added, VIC-SSL further gains the performance, having the improvement of 14.6%, 28.8%, and 17.5% in overall metric scores against to CGNet. We attribute this improvement to VIC-SSL's enhanced inter-frame correspondence

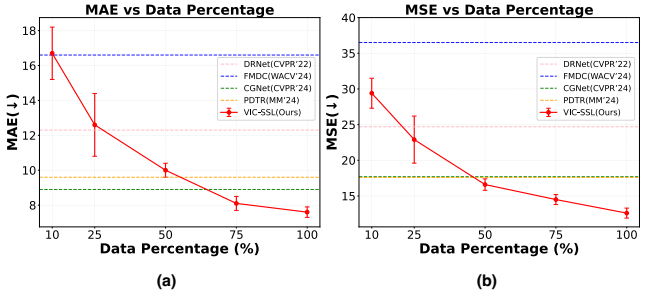

Figure 6: The (a) MAE and (b) MSE analysis of fine-tuning on different amounts of data on SenseCrowd.

Table 4: Effectiveness of each component on SenseCrowd.

Method	MAE↓	MSE↓	WRAE(%)↓
Baseline	12.6	19.1	16.1
+ DCA	9.3	15.9	11.5
+ DCA+ CFP	8.2	14.3	10.7
+ DCA+ CFP+ Pretrain	**7.6**	**12.6**	**10.3**

learning, leveraging abundant unlabeled data via our proposed strategy to better capture spatial and contextual crowd dynamics.

Moreover, VIC-SSL exhibits strong performance across varying crowd densities. In the highest-density setting (*i.e.*, $\mathcal{D}_4$), it reduces MAE by 29.8 compared to PDTR, demonstrating its ability to model fine-grained motion. Specifically, DCA improves feature alignment across frames, while CFP discerns local movement rather than relying solely on global cues. To illustrate VIC-SSL's adaptability, Figure 5 visualizes predictions for two different scenes, showing accurate inflow estimates under differing conditions. The decoupling mask M pinpoints distinct areas, splitting the feature map into distinct and shared regions that closely match actual inflow locations, further highlighting VIC-SSL's robust performance.

Results on CroHD. Table 2 presents the results on CroHD. Our proposed VIC-SSL achieves the best overall performance, attaining 40.6, 50.9, and 8.8 in MAE, MSE, and WRAE, respectively. When trained from scratch, VIC-SSL performs comparably to FMDC, which leverages identity-level annotations for inter-frame contrastive learning in highly congested scenes, whereas we only use group-level labels. However, once our self-supervised pre-training is applied, VIC-SSL improves by 25.1% and 17.5% in MAE and MSE, respectively. This significant gain validates how simulating crowd movement via F-ShiftMix during pre-training enables superior fine-grained inter-frame feature matching—even surpassing methods that rely on identity-level supervision. These results highlight the robustness of our approach in densely crowded scenarios.

Results on CARLA. Table 3 compares the performance on CARLA, where VIC-SSL shows clear superiority, reducing MAE and MSE by 39.4% and 59.4%, respectively, relative to the second-best FMDC. Notably, our self-supervised pre-training contributes a substantial performance boost even if CARLA is a synthetic dataset with a domain gap between the pre-trained real-world dataset, SenseCrowd. Furthermore, CARLA presents scenes of diverse weather conditions

Table 5: Shifting strategy on SenseCrowd.

Strategy	Radius r	MAE↓	MSE↓	WRAE(%)↓
Train from scratch	-	8.2	14.3	10.7
Random 30%	7	8.6	15.2	11.0
Random 50%	7	10.1	16.7	12.3
Foreground-driven	2	9.5	16.0	12.0
Foreground-driven	7	**7.6**	**12.6**	**10.4**
Foreground-driven	15	8.7	16.1	10.6

and crowd distributions. Despite these domain gaps, VIC-SSL still adapts effectively on CARLA by focusing on inter-frame dynamics learned through CFP in the pre-training. Overall, the impressive performance in downstream fine-tuning with CARLA shows that VIC-SSL exhibits strong robustness and adaptability to new, complex conditions, emphasizing its practical deployment value.

4.3 Ablation Study

Data Efficiency. We investigate the effect of using different proportions of labeled data (100%, 75%, 50%, 25%, and 10%) for fine-tuning, as shown in Figure 6. Notably, our method requires only 50% of the data to outperform state-of-the-art methods in MSE while remaining comparable in MAE. With as little as 25% labeled data, we still match DRNet's performance and surpass FMDC. These strong results stem from our self-supervised strategy via F-ShiftMix, which effectively learns to identify and align frames from abundant unlabeled data, thereby capturing robust inter-frame correspondences. Consequently, the model acquires CFP during pre-training, enabling more accurate recognition of inflow individuals in downstream VIC tasks. As a result, only minimal fine-tuning is needed, significantly reducing data requirements while preserving high performance.

Effectiveness of Each Component. We conduct an ablation study (Table 4) to quantify the impact of each module in our architecture. The baseline mirrors our framework but omits Cost-guided Flow Prompt, replaces DCA with a standard cross-attention layer, and trains from scratch. First, we assess DCA, which targets frame-specific distinctions. This module alone provides over a 15% gain across all metrics, highlighting the importance of identifying unique features between frames in VIC. Next, integrating CFP directs the attention mechanism toward prominent motion cues, enabling more precise inter-frame dynamics capture. Finally, self-supervised pre-training via F-ShiftMix further boosts performance on all metrics by equipping the model with generalized pedestrian representations. Together, these components enhance both the effectiveness and adaptability of VIC-SSL, enabling more accurate tracking of pedestrian inflows. We also visualize the attention maps of DCA in Figure 7 to compare training with and without self-supervised pre-training. When trained from scratch, the model struggles to align the same individual's tokens across frames, leading to mismatched attention values. In contrast, with pre-training in place, VIC-SSL accurately associates individuals in different frames, which substantially improves downstream VIC performance. This outcome further validates our proposed pre-training design via F-ShiftMix.

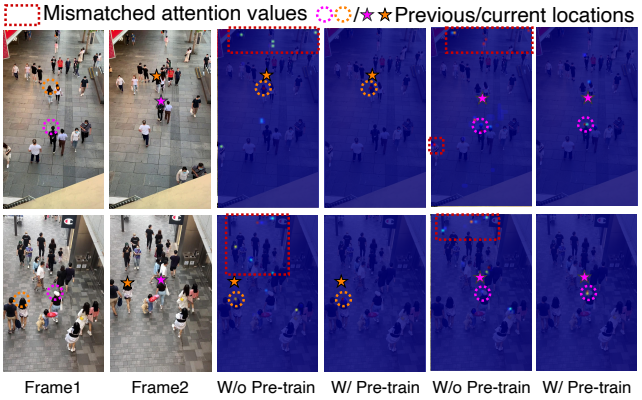

Figure 7: Comparison of DCA attention maps learned with and without pre-training.

Augmentation Design. We assess different augmentation strategies by varying the shifting ratio and the movement radius r in Table 5. In the *Random* setting, a fixed proportion of elements is randomly shifted and blended during pre-training. By contrast, the *Foreground-driven* setting follows the strategy described in Section 3.2.1, focusing on realistic crowd movement where the foreground shifts more than the background. As shown, random shifting actually lowers model performance compared to training from scratch, likely because it disrupts the foreground-background distinction critical to learning real-world crowd dynamics. This finding validates the importance of alignment between our augmentation mechanism and true crowd motion. We also examine how different values of r, which governs the magnitude of feature displacement, affect performance. Experiments indicate that both excessively small and overly large displacements degrade the accuracy, confirming our intuition that real populations move within a bounded range. The best performance is achieved at $r = 7$.

4.4 Limitations

Our method is effective but has some limitations. The fixed radius for the localized cost volume may not be suitable for scenes with varying motion scales. Additionally, the generation of reference frames may encounter difficulties in highly complex or non-linear motion patterns, such as abrupt movements or occlusions.

5 Conclusion

We present VIC-SSL, a novel self-supervised framework for Video Individual Counting (VIC) that addresses the high cost and scalability limitations of annotation-heavy methods. By introducing Foreground-driven ShiftMix, Cost-guided Flow Prompt (CFP), and the Distinction-aware Cross-Attention (DCA) module, our approach effectively learns fine-grained inter-frame correspondences from unlabeled videos. This enables the model to accurately estimate pedestrian inflow. Extensive evaluations across diverse benchmarks demonstrate the superior accuracy, data efficiency, and generalization capabilities of VIC-SSL. In the future, we aim to explore the interplay between self-supervised learning and video-specific temporal structures to further enhance model generalization.

Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan, under Grant: NSTC-113-2640-B-005-001, NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Görkay Aydemir, Weidi Xie, and Fatma Guney. 2023. Self-supervised object-centric learning for videos. *Advances in Neural Information Processing Systems* (2023).

[2] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9686–9696.

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[4] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. 2022. Frame-wise action representations for long videos via sequence contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[5] Pengguang Chen, Shu Liu, and Jiaya Jia. 2021. Jigsaw clustering for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. 2023. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters* (2023).

[7] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Dongliang He, and Weipinng Wang. 2022. Mamico: Macro-to-micro semantic correspondence for self-supervised video representation learning. In *ACM International Conference on Multimedia*.

[8] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 2021. A large-scale study on unsupervised spatiotemporal representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[9] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. 2021. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2917–2927.

[10] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. 2023. Siamese masked autoencoders. *Advances in Neural Information Processing Systems* (2023).

[11] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. 2022. Dr. vic: Decomposition and reasoning for video individual counting. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3083–3092.

[12] Qi He, Zhaoquan Yuan, Xiao Wu, and Jun-Yan He. 2022. Domain-Specific Conditional Jigsaw Adaptation for Enhancing transferability and Discriminability. In *ACM International Conference on Multimedia*.

[13] Mohammad Asiful Hossain, Kevin Cannons, Daesik Jang, Fabio Cuzzolin, and Zhan Xu. 2020. Video-based crowd counting using a multi-scale optical flow pyramid network. In *Asian Conference on Computer Vision*.

[14] Auke Hunneman, Tammo H.A. Bijmolt, and J. Paul Elhorst. 2023. Evaluating store location and department composition based on spatial heterogeneity in sales potential. *Journal of Retailing and Consumer Services* 73 (2023), 103355.

[15] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. 2021. Learning to estimate hidden motions with global motion aggregation. In *International Conference on Computer Vision*. 9772–9781.

[16] Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. 2022. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review* (2022).

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25 (2012).

[18] Yi Lei, Huilin Zhu, Jingling Yuan, Guangli Xiang, Xian Zhong, and Shengfeng He. 2024. DenseTrack: Drone-based Crowd Tracking via Density-aware Motion-appearance Synergy. In *ACM International Conference on Multimedia*. 2050–2058.

[19] Haopeng Li, Lingbo Liu, Kunlin Yang, Shinan Liu, Junyu Gao, Bin Zhao, Rui Zhang, and Jun Hou. 2022. Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark. *IEEE Transactions on Image Processing* 31 (2022), 6032–6047.

[20] Qiankun Li, Xiaolong Huang, Zhifan Wan, Lanqing Hu, Shuzhe Wu, Jie Zhang, Shiguang Shan, and Zengfu Wang. 2023. Data-efficient masked video modeling for self-supervised action recognition. In *ACM International Conference on Multimedia*.

[21] Rui Li, Yishu Liu, Huafeng Li, Jinxing Li, and Guangming Lu. 2024. Prototype-Guided Dual-Transformer Reasoning for Video Individual Counting. In *ACM International Conference on Multimedia*. 10258–10267.

[22] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. 2024. Gramformer: learning crowd counting via graph-modulated transformer. In *Association for the Advancement of Artificial Intelligence*, Vol. 38. 3395–3403.

[23] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. 2022. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *European Conference on Computer Vision*.

[24] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2021. Counting people by estimating people flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8151–8166.

[25] Xinyan Liu, Guorong Li, Yuankai Qi, Ziheng Yan, Zhenjun Han, Anton Van Den Hengel, Ming-Hsuan Yang, and Qingming Huang. 2024. Weakly Supervised Video Individual Counting. In *IEEE Conference on Computer Vision and Pattern Recognition*. 19228–19237.

[26] Yunze Liu, Changxi Chen, Zifan Wang, and Li Yi. 2024. CrossVideo: Self-supervised Cross-modal Contrastive Learning for Point Cloud Video Understanding. In *IEEE International Conference on Robotics and Automation*.

[27] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*.

[28] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. 2022. Automix: Unveiling the power of mixup for stronger classifiers. In *European Conference on Computer Vision*.

[29] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. 2021. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. *IEEE Transactions on Multimedia* 24 (2021), 261–273.

[30] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *IEEE International Conference on Image Processing*.

[31] Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Gustavo A Vargas Hakim, David Osowiechi, Ismail Ben Ayed, and Christian Desrosiers. 2024. Tfs-vit: Token-level feature stylization for domain generalization. *Pattern Recognition* (2024).

[32] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015).

[33] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. 2023. Masked motion encoding for self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. 2021. Tracking pedestrian heads in dense crowd. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3865–3875.

[35] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*. Springer, 402–419.

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems* (2022).

[37] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. 2024. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*.

[38] Xuan-Thuy Vo, Van-Dung Hoang, Duy-Linh Nguyen, and Kang-Hyun Jo. 2022. Pedestrian head detection and tracking via global vision transformer. In *International Workshop on Frontiers of Computer Vision*.

[39] Chang-Lin Wan, Feng-Kai Huang, and Hong-Han Shuai. 2024. Density-Based Flow Mask Integration via Deformable Convolution for Video People Flux Estimation. In *IEEE Winter Conference on Application of Computer Vision*. 6573–6582.

[40] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. 2023. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[41] Shuai Wang, Da Yang, Yubin Wu, Yang Liu, and Hao Sheng. 2022. Tracking game: Self-adaptative agent based multi-object tracking. In *ACM International Conference on Multimedia*. 1964–1972.

[42] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[43] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung-Hin So, and Xin Li. 2024. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. In *Association for the Advancement of Artificial Intelligence*, Vol. 38. 5740–5748.

[44] Rukai Wei, Yu Liu, Jingkuan Song, Heng Cui, Yanzhao Xie, and Ke Zhou. 2023. Chain: Exploring global-local spatio-temporal information for improved self-supervised video hashing. In *ACM International Conference on Multimedia*.

[45] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. 2021. Detection, tracking, and counting meets drones in crowds: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7812–7821.

[46] Han Xiao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2023. Token-label alignment for vision transformers. In *International Conference on Computer Vision*.

[47] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. 2017. Spatiotemporal modeling for crowd counting in videos. In *International Conference on Computer Vision*. 5151–5159.
[48] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. 2023. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition* (2023).
[49] Wenhan Yang, Rizhao Cai, and Alex Kot. 2022. Image inpainting detection via enriched attentive pattern with near original image augmentation. In *ACM International Conference on Multimedia*.
[50] Hao Yin, Dongyu Cao, and Ying Zhou. 2022. Randommix: An effective framework to protect user privacy information on ethereum. In *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion*.
[51] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. 2022. S2-mlp: Spatial-shift mlp architecture for vision. In *IEEE Winter Conference on Application of Computer Vision*.
[52] Liangzhe Yuan, Rui Qian, Yin Cui, Boqing Gong, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. 2022. Contextualized spatio-temporal contrastive learning with self-supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[53] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision*.
[54] Wei Zhang, Lingxiao He, Peng Chen, Xingyu Liao, Wu Liu, Qi Li, and Zhenan Sun. 2021. Boosting end-to-end multi-object tracking and person search via knowledge distillation. In *ACM International Conference on Multimedia*. 1192–1201.
[55] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision* (2021), 3069–3087.
[56] Yang Zhao, Gangwei Xu, and Gang Wu. 2024. Hybrid Cost Volume for Memory-Efficient Optical Flow. In *ACM International Conference on Multimedia*. 8740–8749.
[57] Zhuoyi Zhao, Hongsheng Li, Rui Zhao, and Xiaogang Wang. 2016. Crossing-line crowd counting with two-phase deep neural networks. In *European Conference on Computer Vision*.
[58] Jiquan Zhong, Xiaolin Huang, and Xiao Yu. 2023. Multi-frame self-supervised depth estimation with multi-scale feature fusion in dynamic scenes. In *ACM International Conference on Multimedia*.