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

A Appendix

A.1 Implementation Details

Network and Pre-training Settings. We adopt VGG-16 [3] pre-trained on ImageNet [2] as our feature extractor. To maintain stable representations and prevent feature space collapse during the self-supervised pretext task, the feature extractor is completely frozen during pre-training. Our VIC-SSL is pre-trained on the SenseCrowd dataset for 10 epochs with a batch size of 3. The input images are resized to 512×512, resulting in an extracted feature map resolution of 32×32. Both the feature map dimension c and the flow prompt dimension g are set to 512. We initialize the learning rate at 1×10^{-4} and apply a step decay of 0.95 per epoch. For the Foreground-driven ShiftMix (F-ShiftMix), the shifting ratios for the foreground and background elements are 100% and 1%, respectively. The radius for bounded offsets and the localized cost volume is fixed at $r = \gamma = 7$, and we utilize $S = 8$ attention heads in the cross-attention module. The weighting factor ω in the pre-training loss $\mathcal{L}^{\text{PT}}$ is set to 1.

Fine-tuning Settings. During the fine-tuning phase, we freeze the Cost-guided Flow Prompt (CFP) module to preserve its learned dynamic guidance and motion priors. The model is fine-tuned using a batch size of 2 with a higher input resolution of 768×768. We employ a one-cycle learning rate scheduler with a one-epoch warm-up. The learning rates for the trainable modules are set to 1×10^{-4}, 3×10^{-5}, and 1×10^{-6} for the SenseCrowd, CroHD, and CARLA datasets, respectively. The loss balancing factor α in $\mathcal{L}^{\text{FT}}$ is set to 1. Frame pairs are randomly sampled with time intervals ranging from 1s to 5s during training, while a fixed interval of 2s is utilized during inference to ensure consistent temporal evaluation.

*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

Evaluation Metrics Formulation. In addition to standard error metrics, we utilize the Weighted Relative Absolute Error (WRAE) to provide a balanced evaluation across varying video lengths and pedestrian densities. It is formally defined as:

$$WRAE = \sum_{m=1}^{M} \frac{T_m}{\sum_{n=1}^{M} T_n} \frac{|N_m - \hat{N}_m|}{N_m}, \tag{1}$$

where M denotes the total number of videos, T_m is the duration (length) of the m-th video, and N_m and $\hat{N}_m$ represent the ground-truth and predicted individual counts in the m-th video, respectively.

A.2 Further Ablation and Analysis

Computational Cost and Efficiency. To demonstrate the practical deployment value of VIC-SSL, we conduct a comprehensive analysis of the computational overhead compared to existing end-to-end approaches, as detailed in Table A1. Our complete framework significantly surpasses prior methods in both accuracy and efficiency. Specifically, VIC-SSL consumes 28.1% to 35.8% fewer GFLOPs than FMDC [4] and DRNet [1], while achieving a superior inference speed of 16.0 FPS. These efficiency gains are primarily attributed to performing our lightweight Distinction-aware Cross-Attention (DCA) on high-level, lower-resolution feature maps rather than directly on the image space. Furthermore, the introduction of DCA provides a substantial performance boost (MAE drops from 12.6 to 9.3) at virtually no additional computational expense. While integrating CFP slightly increases the parameter count, the resulting state-of-the-art accuracy justifies this minor trade-off, confirming that our method strikes an optimal balance between precision and computational efficiency.

Impact of Freezing Strategies. We investigate how different component freezing strategies during the pre-training and fine-tuning stages affect the final downstream performance, with results summarized in Table A2. Freezing the feature extractor during pre-training is crucial; it prevents the well-established spatial representations from degrading and forces the network to focus entirely on learning inter-frame correspondences. Conversely, freezing the CFP module during fine-tuning yields a significant performance

Table A1: Comparison of performance and computational cost on the SenseCrowd dataset.

Method	Config.	MAE↓	Param. (M)	GFLOPs↓	FPS↑
DRNet [1]	-	12.3	18.8	1555.9	12.4
FMDC [4]	-	16.6	20.9	1743.7	9.2
VIC-SSL (Ours)	Baseline	12.6	20.3	1119.3	17.2
	+ DCA	9.3	20.3	1119.3	17.2
	+ DCA + CFP	8.2	24.8	1154.3	16.0
	+ DCA + CFP + Pretrain	**7.6**	24.8	1154.3	16.0

gain (reducing MAE from 10.6 to 7.6). Because CFP distills generalized crowd dynamic cues from extensive unlabeled video sequences during pre-training, preserving these weights prevents catastrophic forgetting and avoids overfitting to the smaller fine-tuning dataset. This strategy ensures that the Distinction-aware Cross-Attention (DCA) can effectively leverage these robust motion priors to refine inter-frame relationships.

Table A2: Ablation of freezing strategies during pre-training and fine-tuning on SenseCrowd.

Frozen Component		MAE↓	MSE↓	WRAE(%)↓
Pre-training	Fine-tuning			
-	CFP	13.5	21.3	15.3
Feature Extractor	-	10.6	19.1	15.3
Feature Extractor	CFP	**7.6**	**12.6**	**10.4**

The Purpose of Background Shifting. In F-ShiftMix, distinct shifting ratios are applied to the foreground and background regions. While surveillance cameras are typically stationary, implying a static background, introducing minor background shifts is essential for feature-level augmentation. In convolutional neural networks, a feature's representation is deeply intertwined with its surrounding spatial context. If we only shift the foreground while keeping the background entirely static, the shifted pedestrian features will share nearly identical local contextual cues with their original positions. This allows the model to rely on trivial static cues rather than learning genuine temporal motion mapping. By applying a 1% bounded shift to the background, we introduce slight contextual perturbations. This subtle distortion forces the network to perform rigorous, distinct feature matching to correctly predict the transportation map. As demonstrated in Table A3, incorporating background shifts yields a notable 9.5% improvement in MAE compared to using a perfectly static background, confirming its necessity for robust representation learning.

Table A3: Analysis of the background shifting strategy in F-ShiftMix on SenseCrowd.

Background Shifting	MAE↓	MSE↓	WRAE(%)↓
✗	8.4	14.0	10.6
✓	**7.6**	**12.6**	**10.4**

References

[1] 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.

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

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

[4] 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.