SCTrans: Multi-scale scRNA-seq Sub-vector Completion Transformer for Gene-selective Cell Type Annotation

Lu Lin[1*], **Wen Xue**[1*], **Xindian Wei**[2], **Wenjun Shen**[3], **Cheng Liu**[4], **Si Wu**[1] and **Hau San Wong**[2]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Department of Computer Science, City University of Hong Kong
[3]Department of Bioinformatics, Shantou University Medical College
[4]Department of Computer Science, Shantou University
{cslinlu, csxuewen}@mail.scut.edu.cn, xindiawei2-c@my.cityu.edu.hk,
{wjshen, cliu}@stu.edu.cn, cswusi@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Cell type annotation is pivotal to single-cell RNA sequencing data (scRNA-seq)-based biological and medical analysis, e.g., identifying biomarkers, exploring cellular heterogeneity, and understanding disease mechanisms. The previous annotation methods typically learn a nonlinear mapping to infer cell type from gene expression vectors, and thus fall short in discovering and associating salient genes with specific cell types. To address this issue, we propose a multi-scale scRNA-seq Sub-vector Completion Transformer, and our model is referred to as SCTrans. Considering that the expressiveness of gene sub-vectors is richer than that of individual genes, we perform multi-scale partitioning on gene vectors followed by masked sub-vector completion, conditioned on unmasked ones. Toward this end, the multi-scale sub-vectors are tokenized, and the intrinsic contextual relationships are modeled via self-attention computation and conditional contrastive regularization imposed on an encoding transformer. By performing mutual learning between the encoder and an additional lightweight counterpart, the salient tokens can be distinguished from the others. As a result, we can perform gene-selective cell type annotation, which contributes to our superior performance over state-of-the-art annotation methods.

1 Introduction

The rapid advancements in single-cell RNA sequencing (scRNA-seq) technologies have enabled high-resolution characterization of tissue heterogeneity at cellular-level[Ziegenhain *et al.*, 2017]. As a critical application of scRNA-seq, automatic cell type annotation enables the precise identification of unique and shared genetic expressions across different cell types. This process elucidates the molecular characteristics and functional regulation of cells. Furthermore, cell type annotation facilitates the comparison of cellular typology alterations under pathological states, including cells' stress responses and adaptive mechanisms. This sheds light on cellular functional dysregulation and the mechanisms of disease.

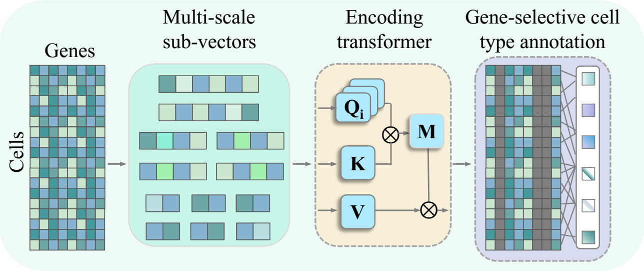

Figure 1: Illustrating SCTrans performs gene sub-vector representation learning and gene-selective cell type annotation via multi-scale tokenization and attention computation at each transformer block.

There are a number of attempts made in applying machine learning algorithms to cell type annotation [Wang *et al.*, 2022]. For example, scBERT [Yang *et al.*, 2022] leverages pretrained bidirectional transformer networks to learn the contextual relationships between gene expression vectors. Another representative strategy is to adopt an optimal transport model to identify the known cell types while at the same time discovering new ones, such as scPOT [Zhai *et al.*, 2023]. However, the existing methods mainly focus on learning a nonlinear mapping to infer cell types from individual genes, often overlooking the interdependency between the sub-vectors of gene expression. Compared to individual genes, we consider that gene sub-vectors provide richer information, and performing attention computation among them potentially identities salient ones for downstream biological research and analysis as shown in Figure 1.

More specifically, we propose a multi-scale gene expression Sub-vector Completion Transformer (SCTrans) for cell type annotation over scRNA-seq. SCTrans aims to learn an effective representation of the sub-vectors, such that the interdependency and the association with cell type can be better captured, than dealing with individual genes. As shown in Figure 2, we adopt a multi-scale partitioning strategy to yield gene sub-vectors across different scales. Next, the resulting sub-vectors are tokenized, followed by self-attention computation in each transformer block. One of our training goals is to perform sub-vector completion across different scales: predicting the masked sub-vectors from the unmasked ones. We further impose conditional contrastive regularization on the transformer to learn cell-type-aware representation. By modeling intrinsic contextual relationships, another advantage is to distinguish salient tokens from the others, based on attention scores. As a result, we can perform gene-selective cell type annotation, and provide meaningful and interpretable results. We perform extensive experiments on multiple benchmark dataset to highlight the effectiveness of our design elements and the superior performance over state-of-the-art cell type annotation methods.

In summary, the main contribution of this work are as follows: (a) We design a multi-scale scRNA-seq encoding transformer to learn the representation of gene sub-vectors, whose expressiveness is richer than that of individual genes. (b) Based on the nature of gene co-expression, we perform gene sub-vector completion to capture the intrinsic contextual relationships across different scales. (c) By performing mutual learning between the transformer and a lightweight counterpart, salient tokens of various scales can be distinguished from the others, and lead to significant improvement in terms of annotation accuracy and interpretability.

2 Related Work

2.1 Transformer Based Representation Learning

The Transformer architecture [Xu *et al.*, 2019][Dosovitskiy *et al.*, 2021][Wu *et al.*, 2021][Liu *et al.*, 2021] has emerged as a powerful framework, drawing inspiration from transformer research in natural language processing (NLP) [Devlin *et al.*, 2018][Vaswani *et al.*, 2017]. By utilizing the self-attention mechanism, Transformer can model the long time relation, particularly effective in analyzing complex relationships in biological data[Abnar and Zuidema, 2020][Alsaigh *et al.*, 2022]. In single-cell RNA sequencing (scRNA-seq) analysis, Transformer has led to significant advancements. For instance, scBERT [Yang *et al.*, 2022] utilizes the Transformer to annotate cell types in scRNA-seq data, learning from large datasets to understand gene interactions. Similarly, the Exceiver model [Connell *et al.*, 2022], based on the Perceiver IO framework, adapts to new datasets for gene expression analysis. TOSICA [Chen *et al.*, 2023], another Transformer-based tool, excels in integrating diverse datasets for cell type annotation. The DeepMAPS [Ma *et al.*, 2023] model uses a unique approach combining cells and genes to infer biological networks, streamlining the training process. Moreover, scFoundation [Hao *et al.*, 2023] and scAAGA [Meng *et al.*, 2023] models enhance single-cell analysis by transforming gene expression data into informative vectors and employing gene attention modules for feature extraction.

These models are exemplary in their integration of computational techniques with biological data, providing profound insights into genomic analysis. However, a limitation of existing methods is their focus either on single genes, failing to fully reveal the expressions across the gene sub-vectors. This oversight can potentially neglect the deeper implications and nuances within the genomic data.

2.2 Cell Type Annotation

Cell type annotation methods can be broadly categorized into two approaches: cluster-then-annotate and supervised cell type classification. The cluster-then-annotate approach seeks to categorize cells into distinct clusters, and these identified clusters are subsequently manually annotated by experts who examine cluster-specific gene expression patterns [Kiselev *et al.*, 2019]. In this learning paradigm, which falls under the unsupervised learning branch, various clustering approaches have been developed to identify cell types. For example, SC3 accomplishes cell clustering by consistently integrating diverse clustering solutions through a consensus approach [Kiselev *et al.*, 2017]. SIMLR strives to acquire a more accurate metric that better reflects the similarity between samples, capturing the underlying structure of the data through a multiple kernel learning approach [Wang *et al.*, 2017]. Recently, deep learning methods have emerged as powerful tools for clustering scRNA-seq data, overcoming challenges posed by high dropout rates and noise [Tian *et al.*, 2021]. For example, DESC is an unsupervised deep embedding algorithm designed to accurately cluster scRNA-seq data through an iterative self-learning process [Li *et al.*, 2020b]. scziDesk utilizes a convolutional autoencoder to learn representations of the single cell data, followed by a regularized soft k-means approach to identify cell clusters [Hu *et al.*, 2022]. scNAME method incorporates mask estimation task, in conjunction with a neighborhood contrastive learning framework to achieve effective clustering [Wan *et al.*, 2022]. These deep clustering techniques can characterize complex nonlinear relationships in single cell data to identify the cell groupings.

In contrast, the supervised classification approach involves training a model on data where cell types are pre-annotated. Subsequently, this model is utilized to predict cell type labels for new, unlabeled single cells [Abdelaal *et al.*, 2019]. While supervised methods can leverage expert knowledge during the training process, their effectiveness may be constrained by the quantity and quality of labeled data available. ScType leverages a comprehensive database of cell type-specific marker genes to facilitate the automated high-throughput annotation of cell types [Ianevski *et al.*, 2022]. scDeepSort leverages a pre-trained weighted graph neural network (GNN) for cell-type annotation [Shao *et al.*, 2021]. MARS, a meta-learning approach, is designed for the identification and annotation of both known and novel cell types [Brbić *et al.*, 2020]. scPOT is an end-to-end algorithm specifically designed for the annotation of cell types, including the discovery of novel cell types. This is achieved through the utilization of the optimal transport (OT) framework [Zhai *et al.*, 2023]. scBERT is a pre-trained deep cell-type annotation model that leverages the Bidirectional Encoder Representations from Transformers (BERT [Devlin *et al.*, 2018]).

Significantly different from existing methods, this work proposes a biologically interpretable cell annotation method that considers biological information. The proposed model not only accurately achieves cell annotation but also exploits the contextual relationships that exist between gene expression, providing meaningful biological gene sub-vector selection for downstream biological analysis.

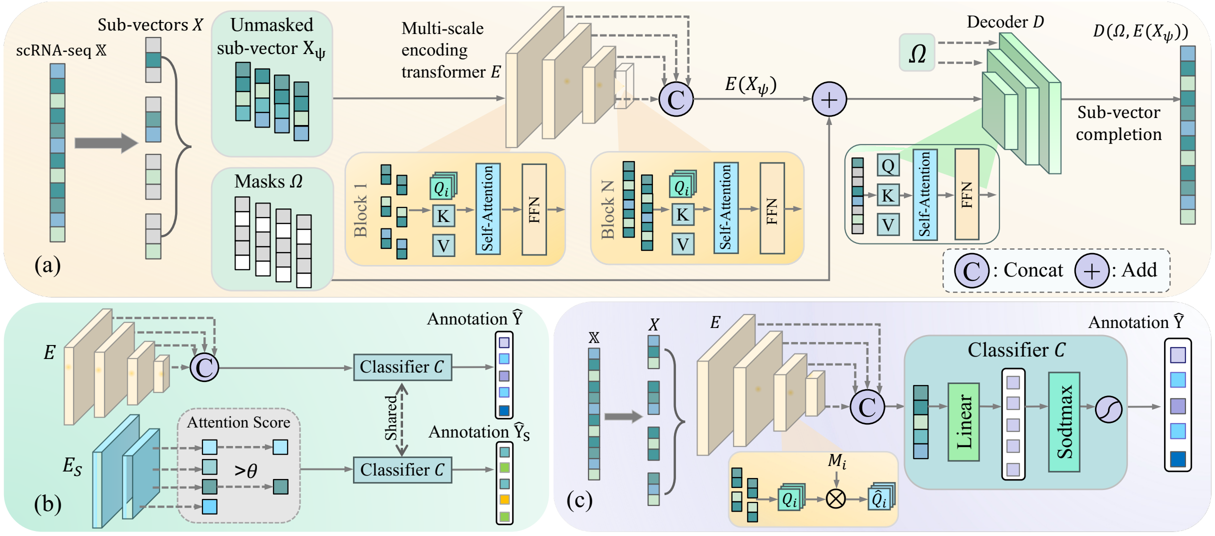

Figure 2: Our proposed multi-scale gene expression Sub-vector Completion Transformer (SCTrans). As shown in (a), SCTrans leverage random sub-vector completion to learn well-defined representations. This multi-scale encoder E aims to reconstruct multi-scale masked gene sub-vector to mitigate the impact of severe sparsity for scRNA-seq data. (b) represents the attention score based gene selection module, by incorporating the idea of mutual learning, a lightweight gene selection Encoder E_S will use the attention scores to assess the importance of the gene sub-vector and select salient tokens. Importantly, (c) performs our novel gene-selective cell type annotation mechanism, we leverage the coefficient matrix from (b) to focus on the salient tokens, utilize E and classifier C to annotate cell type gene-selectively.

3 Methodology

In this section, we present the details of our proposed SCTrans tailored for cell type annotation. As illustrated in Figure 2, we design an encoding transformer $\mathbb{E}$ to learn the representation of gene expression sub-vectors, and adopt an effective sub-vector selection mechanism to identify the salient ones, and facilitate gene-selective cell type annotation.

3.1 Random Sub-vector Completion

Based on the nature of gene co-expression, we adopt a self-supervised training strategy of random gene sub-vector completion to learn the representation of gene sub-vector for cell type annotation. This process is designed to effectively explore the intrinsic contextual relationships among gene expression patterns at various scales. As illustrated in Figure 2.(a), the gene sequencing data $\mathbb{X}$ can be sub-vectorified into multiple gene expression sub-vectors $\mathbf{X}$, and random masks $\mathbf{\Omega}$ are generated to mask the majority of these sub-vectors, the masked regions are denoted as $\mathbf{X}_{\mathbf{\Omega}}$. Inspired by Masked Auto-Encoder [He *et al.*, 2022], one of the training goal is to predict the masked sub-vectors, conditioned on the unmasked ones. To achieve this, the unmasked sub-vectors $\mathbf{X}_{\mathbf{\Psi}}$ are fed into the multi-scale encoding transformer E ($\mathbf{\Psi}$ denotes the unmasked position index). Specifically, we perform attention computation over multi-scale Queries $\{Q^{(i)}\}$ and global Key-Value pairs $\{K, V\}$ to explore the interdependency and learn the representation of gene sub-vectors across different scales, formulated as follows:

$$\begin{aligned}\{Q\}^{(i)} &= W_q^{(i)} \otimes \mathcal{I}(\mathbf{X}_{\mathbf{\Psi}}, i) + b_q^{(i)}, \quad i = 1, \dots, N,\\ \{K, V\} &= W_{\{k,v\}} \otimes \mathbf{X}_{\mathbf{\Psi}} + b_{\{k,v\}},\end{aligned} \tag{1}$$

where $W_{q,k,v}$ and $b_{q,k,v}$ are learnable model parameters, N denotes the number of sub-vector sizes, $\mathcal{I}(\cdot, i)$ represents the function to produce gene sub-vectors with size i. In the above equation, $Q^{(i)}$ involves the gene sub-vector representations with difference sizes. With the aggregation of multiple transformer blocks, E explores the interdependency of gene sub-vector with different scales, and the output of multiple blocks are concatenated together to build the final output $E(\mathbf{X}_{\mathbf{\Psi}})$:

$$\begin{aligned}\mathcal{A}^{(i)} &= \text{softmax}\left(Q^i \cdot K/\sqrt{\Lambda}\right) \cdot V,\\ E(\mathbf{X}_{\mathbf{\Psi}}) &= \text{concat}(\mathcal{A}^{(1)}, ..., \mathcal{A}^{(N)}),\end{aligned} \tag{2}$$

where Λ denotes the number of feature channels. The resulting representation $E(\mathbf{X}_{\mathbf{\Psi}})$ together with the mask Ω are fed into a simple decoding transformer D to predict the masked gene sub-vectors as follows:

$$\mathcal{L}_{comp} = \mathbb{E}_{\mathbf{X},\Omega}\left[\, \|(\mathbf{X}_{\mathbf{\Omega}} - D(\{\mathbf{\Omega}, E(\mathbf{X}_{\mathbf{\Psi}})\}))\|_2^2 \,\right]. \tag{3}$$

By minimizing the above mean squared error (MSE), the multi-scale encoding transformer is trained to complete the unmasked gene sub-vectors, such that their intrinsic contextual relationships are explored and leveraged.

3.2 Gene-selective Cell Type Annotation

To induce our model to concentrate on salient gene sub-vectors specific to cell annotation and provide interpretable results, we incorporate a lightweight encoding transformer denoted as E_S and perform mutual learning with E. The idea behind our selection module is to utilize the attention scores among multi-scaled Queries and global Keys to distinguish the gene sub-vectors playing more important role than the others, and the formulation is expressed as follows:

$$\begin{aligned}S^{(i)} &= Q^{(i)} \cdot K,\\ \mathbf{M}^{(i)} &= \text{sigmoid}(T \cdot (S^{(i)} - \theta))),\end{aligned} \tag{4}$$

where the attention score $S^{(i)}$ is calculated by aggregating the multi-scale Queries $\{Q^i\}$ and the global Key K. To indicate salient gene sub-vectors, a coefficient matrix $\mathbf{M}^{(i)}$ is generated in the above equation. The sigmoid function is used in conjunction with a temperature factor T and a learnable threshold θ to modulate the attention score. The value of the elements in $\mathbf{M}^{(i)}$ ranges from 0 to 1 to indicate the importance of each gene sub-vector at different positions.

To identify the meaningful gene sub-vectors for cell type annotation, $\mathbf{M}^{(i)}$ is further used to guide the lightweight encoding transformer E_S. Specifically, we apply $\mathbf{M}^{(i)}$ to weight the multi-scale gene sub-vector, and obtain the corresponding Queries as follows:

$$\hat{Q}^{(i)} = \hat{W}_q^{(i)} \otimes (\mathcal{I}(\mathbf{X}_{\mathbf{\Psi}}, i) \otimes \mathbf{M}^{(i)}) + \hat{b}^{(i)}, \tag{5}$$

where $\hat{Q}^{(i)}$ denotes the weighted queries with sub-vector size i. Different from E, we perform attention computation at each transformer block of E_S to mainly explore the relationship of the highlighted gene sub-vectors.

On the other hand, we incorporate a linear classifier C on top of E to infer the cell types: $\hat{\mathbf{Y}} = C(E(\mathbf{X}_{\mathbf{\Psi}}))$, where $\hat{\mathbf{Y}}_j$ denotes the predicted probability distribution of the j-th cell. The predictions are evaluated by using the cross-entropy-based loss function $\mathcal{L}_{eval}$ as follows:

$$\mathcal{L}_{eval} = \mathbb{E}_{\mathbf{X},\Omega}\left[\sum_j \left(-\mathbf{Y}_j \cdot \log(\hat{\mathbf{Y}}_j) - (1 - \mathbf{Y}_j) \cdot \log(1 - \hat{\mathbf{Y}}_j)\right)\right]. \tag{6}$$

Minimizing $\mathcal{L}_{eval}$ enforces E and C work together to produce the cell type annotation as accurate as possible. To associate gene sub-vector selection with specific cell types, we further perform mutual learning between E and E_S by imposing prediction consistency regularization as follows:

$$\mathcal{L}_{kl} = \mathbb{E}_{\mathbf{X},\Omega}\left[\sum_j \hat{\mathbf{Y}}_j \cdot \left(-\log(\hat{\mathcal{Y}}_j / \hat{\mathbf{Y}}_j)\right)\right], \tag{7}$$

where $\hat{\mathcal{Y}}_j = C(E_S(\mathbf{X}_{\mathbf{\Psi}}))$ denotes the cell type prediction from the representation learnt by E_S. As will be shown in the experiments, the integration of multi-scale encoding and re-weighting strategies contributes to our cell type annotation performance.

3.3 Model Optimization

Considering that the representation discrepancy among cell types is expected to be significant. For this purpose, we further incorporate a conditional contrastive learning term to regularize the encoding transformer. We can construct semantically congruent and incongruent pairs according to cell type. We aim to push the gene representations of the cells of the same type closer, while moving away from other types of cells. The conditional contrastive loss function is defined as follows:

$$\mathcal{L}_{ctrst} = \mathbb{E}_{\mathbf{X},\Omega}\left[-\sum_j log\left(\frac{\sum_{\mathbf{Y}_j = \mathbf{Y}_k} exp(cos(\mathbf{X}_j, \mathbf{X}_k)/\tau)}{\sum_{\mathbf{Y}_j \neq Y_k} exp(cos(\mathbf{X}_j, \mathbf{X}_k)/\tau)}\right)\right], \tag{8}$$

where $\mathbf{X}_j$ ($\mathbf{X}_k$) represents the j (k)-th cell, and τ is a temperature parameter.

By integrating the above four aspects: gene sub-vector completion, cell type prediction, mutual learning and contrastive regularization, the overall optimization of the proposed SCTrans is expressed as follows:

$$\begin{aligned}
&\min_{E}\ \mathcal{L}_{comp} + \mathcal{L}_{eval} + \mathcal{L}_{ctrst},\\
&\min_{C}\ \mathcal{L}_{eval} + \mathcal{L}_{ctrst},\\
&\min_{D}\ \mathcal{L}_{comp},\\
&\min_{E_S}\ \mathcal{L}_{kl},
\end{aligned} \tag{9}$$

As shown in the above equation, the multi-scale encoding transformer is jointly optimized with other constituent networks to learn cell-type-aware representation for cell type annotation, and the discovered salient gene sub-vectors lead to interpretable results.

4 Experiment

In this work, we conduct experimental studies to evaluate our method on seven representative benchmark datasets, the tissues of datasets including Peripheral Blood Mononuclear Cell (PBMC), Pancreas, Liver and Lung. Table 1 provides concise descriptions of these datasets.We carefully selected these tissues due to their critical roles in health and disease. Our objective in utilizing these diverse datasets is to demonstrate the flexibility and effectiveness of our method.

Dataset	Tissue	Protocol	Cell #	Population #
Zheng68K	PBMC	10X CHROMIUM	68450	11
Baron	Pancreas	inDrop	8569	14
Xin	Pancreas	SMARTer	1449	4
Segerstolpe	Pancreas	SMART-Seq2	2133	13
Muraro	Pancreas	CEL-Seq2	2122	9
MacParland	Liver	10X CHROMIUM	8444	13
Lung	Lung	10X Genomics sequencing	39778	9

Table 1: Dataset descriptions.

4.1 Comparison Results

In this section, we compare our model with several representative cell type annotation methods, including scBERT [Yang *et al.*, 2022], Seurat [Hao *et al.*, 2021], SingleR [Aran *et al.*, 2019], CellID [Cortal *et al.*, 2021], scmap [Kiselev *et al.*, 2018], scNym [Kimmel and Kelley, 2021], and Scibet [Li *et al.*, 2020a]. It is worth noting that both CellID and scmap include two variants with different strategies: CellID_cell and CellID_group for CellID, and scmap_cell and scmap_cluster for scmap. We evaluate the effectiveness of these cell type annotation methods using classification accuracy and F1 Score, where higher values indicate better performance. For each dataset, we randomly select 80% of the total data for model training and use the remaining 20% as test data. Table 2 provides a quantitative comparison between our model and other competing methods across seven datasets. Based on these results, we make the following observations: 1) PBMC (Zheng68k) presents a challenge in

Methods	Zheng68k		Baron		Muraro		Xin		Segerstolpe		MacParland		Lung	
	Accuracy ↑	F1↑	Accuracy ↑	F1↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1↑	Accuracy ↑	F1↑	Accuracy ↑	F1↑	Accuracy ↑	F1↑
scNym	0.700	0.627	0.983	0.805	0.960	0.801	0.906	0.550	0.833	0.648	0.974	0.957	0.925	0.875
SciBet	0.679	0.667	0.975	0.864	0.964	0.819	0.984	0.795	0.789	0.728	0.971	0.957	0.909	0.847
Seurat	0.686	0.581	0.980	0.833	0.964	0.858	0.959	0.594	0.846	0.652	0.979	0.965	0.901	0.833
SingleR	0.326	0.550	0.969	0.878	0.965	0.877	0.977	0.809	0.672	0.692	0.951	0.933	0.870	0.781
CellID_cell	0.566	0.509	0.957	0.857	0.958	0.904	0.901	0.520	0.829	0.724	0.946	0.924	0.950	0.911
CellID_group	0.539	0.567	0.936	0.815	0.943	0.816	0.870	0.528	0.743	0.672	0.882	0.908	0.879	0.837
scmap_cell	0.291	0.212	0.844	0.398	0.801	0.382	0.822	0.281	0.499	0.266	0.826	0.421	0.574	0.287
scmap_cluster	0.463	0.482	0.912	0.826	0.901	0.758	0.956	0.780	0.600	0.619	0.931	0.899	0.893	0.717
scBERT	0.759	0.691	0.977	0.849	0.976	0.932	0.980	0.793	0.892	0.759	0.976	0.959	0.957	0.914
SCTrans	**0.817**	**0.717**	**0.980**	**0.881**	**0.986**	**0.984**	**0.995**	**0.946**	**0.977**	**0.826**	**0.981**	**0.966**	**0.965**	**0.921**

Table 2: Quantitative comparison with competing methods.

scRNA-seq data, as indicated by the unsatisfactory results obtained by competing methods. In contrast, our method consistently achieves the best results, with an accuracy value exceeding 0.8, significantly outperforming other competing methods. 2) While scBERT achieves the second-best results, our proposed method significantly outperforms all competing methods across all seven datasets in terms of accuracy and F1 score. It consistently attains an accuracy exceeding 0.95 for all datasets except Zheng68k.

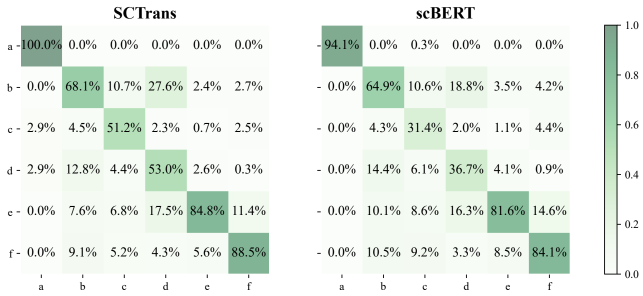

Figure 3: Heatmaps for the confusion matrices of the cross validation results on the Zheng68K dataset for SCTrans and scBERT. (a) to (f) represemts 'CD34+', 'CD4+/CD25 T Reg', 'CD4+/CD45RA+/CD25- Naive T', 'CD4+/CD45RO+ Memory', 'CD8+ Cytotoxic T', 'CD8+/CD45RA+ Naive Cytotoxic'.

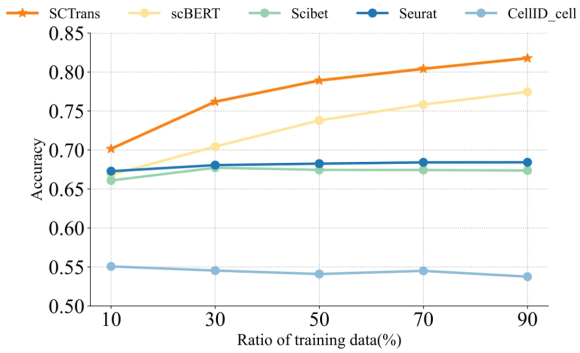

Figure 4: The impact of varying the ratio of training data on model performance on the Zheng68K dataset.

Moreover, we conduct a detailed comparison of the performance of our method and scBERT for each cell type. Figure 3 displays the confusion matrix to illustrate the classification results of our model and scBERT across all cell types on the Zheng68k dataset. Our approach outperforms scBERT in classifying every cell type, with particularly notable improvements in classification performance for the'CD4+/CD45RA+/CD25- Naive T' and 'CD4+/CD45RO+ Memory' cell types.

In addition, we explore the impact of varying training sample sizes on model performance, comparing our method with scBERT, Scibet, Seurat and CellID_cell as depicted in Figure 4. Specifically, we randomly select 10%, 30%, 50%, 70%, and 90% of PBMC cells from the Zheng68K dataset as training data, while the remaining samples serve as testing data. As observed, our model consistently outperforms other methods in terms of accuracy across all settings.

4.2 Model Analysis

Robustness Analysis on Imbalanced Data

In scRNA-seq data, the imbalanced distribution of categories among different cell types poses a challenge. To assess the efficacy of the proposed method in handling imbalanced scRNA-seq data, we specifically chose four distinct PBMC cell populations ('CD19+ B', 'CD8+ cytotoxic T', 'CD34+', and 'CD8+/CD45RA naive cytotoxic cells') from the Zheng68K dataset. In detail, we randomly sampled cells from these types, obtaining counts of 100, 10,000, 100, and 10,000, respectively, to construct the training dataset. For model testing, 100 cells were randomly selected as the testing dataset. As depicted in Figure 5, in the same experimental setup, scBERT exhibits low annotation accuracy for recognizing cell types, whereas our model consistently achieves satisfactory results across all cell types. This underscores the robustness of our method in classifying imbalanced data.

Robustness Analysis on Batch Effects

In the collection of scRNA-seq data, it is common to integrate multiple samples from different sequencing platforms, leading to batch effects. These batch effects can significantly impact the accuracy of experimental results. To assess our model's capability to overcome batch effects, we conduct experiments across datasets, integrating human pancreas datasets generated from diverse sequencing technologies, including Baron, Muraro, Segerstolpe, and Xin. These datasets were harmonized into a unified dataset by aligning common genes and preserving four shared pancreas cell types ('alpha', 'beta', 'delta', and 'gamma').

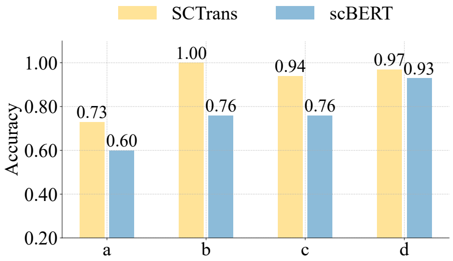

Figure 5: Performance of SCTrans and scBERT on the imbalanced dataset reconstructed from Zheng68K dataset. (a) to (d) represents ‘CD19+ B’, ‘CD8+ cytotoxic T’, ‘CD34+’, and ‘CD8+/CD45RA naive cytotoxic cells’.

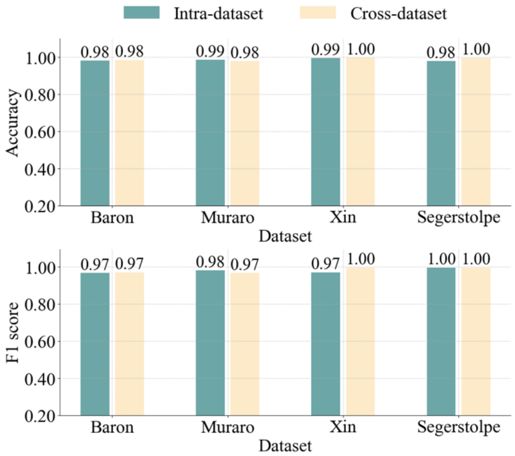

Figure 6: Performance of intra-dataset and cross-dataset classification on human pancreas dataset.

In this setting, three of the datasets were utilized for training, while the fourth served as the test set. We apply our SCTrans method to this constructed dataset (cross-dataset), and for comparative analysis, we also perform tests on single source data (intra-dataset), where both the training and test sets originated from the same human pancreas dataset, such as Baron or Muraro. As illustrated in Figure 6, our model demonstrates high accuracy and F1 scores on the unified constructed dataset. Notably, it achieves accuracy results that are comparable to, or even better than, those obtained on single source data. These findings indicate that the representation learning from scRNA-seq data facilitated by our model is robust against batch effects.

Biological Meaningful Gene Sub-vector Selection

To assess whether our model can concentrate on the most influential gene segments for cell type recognition task, we conduct experiments involving the selection of a specific ratio of relevant gene sub-vectors selected by SCTrans shown in Figure 7. As scBERT lacks the capability for gene segments selective, we randomly remove a certain ratio of gene sub-vectors. Subsequently, we compare the resulting accuracy and F1 scores under different settings. Notably, our model demonstrates robust performance, maintaining high accuracy and F1 scores even after the removal of a ratio of gene sub-vector through selection. In contrast, scBERT experiences a notable decline in classification performance as the ratio of removed gene sub-vectors increases.

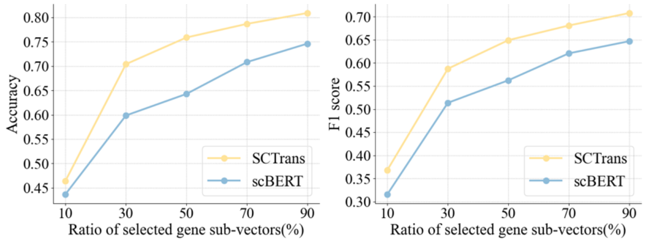

Figure 7: Accuracy and F1 score of selecting different ratios of gene sub-vectors on the Zheng68K dataset.

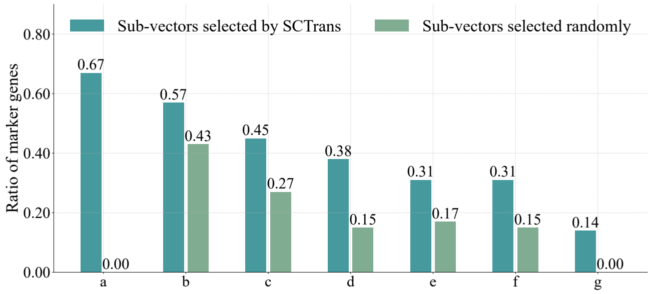

Figure 8: Meaningful gene sub-vector selection ratio. The attention-score based selection mechanism can make SCTrans focus the salient tokens.(a) to (g) represents ‘CD14+ Monocyte’,‘CD4+/CD25 T Reg’,‘CD4+ T Helper2’,‘CD34+’,‘Dendritic’,‘CD4+/CD45RO+ Memory’,‘CD56+ NK’.

Moreover, we anticipate that our SCTrans model can identify meaningful and interpretable gene sub-vectors, which would be highly valuable for downstream biological analyses. To validate this, we undertake a meticulous selection of salient gene sub-vectors and evaluate their biological significance. Specifically, if a marker gene is present within the selected gene sub-vector, this sub-vector is considered biologically meaningful. As depicted in Figure 8, we compare the gene sub-vectors selected by SCTrans against those selected randomly. The results from this comparison indicate that SCTrans is adept at selecting biologically relevant gene sub-vectors. Additionally, SCTrans’s selection varies from 0.67 to 0.14, we think this may due to the meaningful tokens’s mapping is not average, and the meaningful tokens have weak relation with gene expression in some specific situation. SCTrans’s selection may provide insights from other views.

Combining the outcomes from Figures 7 and 8, we can conclude that SCTrans is not only effective in gene selection, which contributes to accurate cell annotation, but also excels in providing biologically meaningful and interpretable results, demonstrating its multifaceted utility in the realm of biological research.

Representation Learning

Figure 9 (right) depicts the UMAP visualization of gene data representations learned by our model, while the left panel illustrates the UMAP visualization based on the original gene

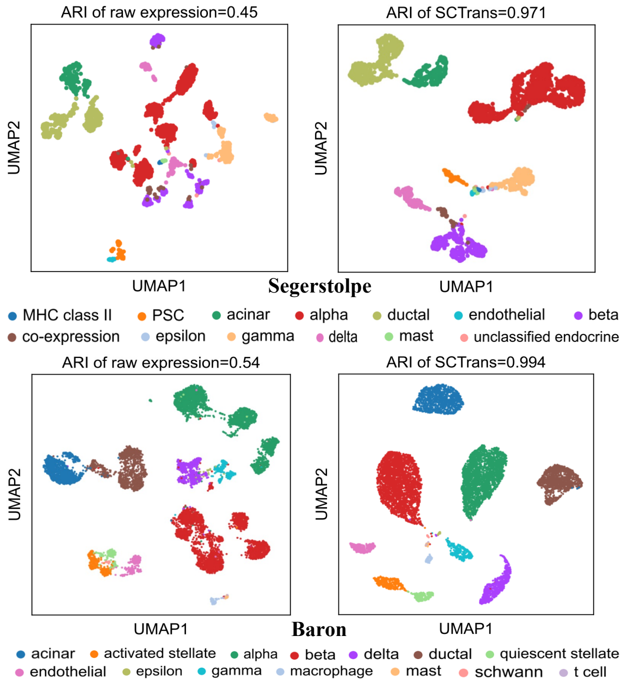

Figure 9: UMAP representation of cells from the Segerstolpe dataset coloured by cell types, based on the raw expression (left) and the SCTrans embedding (right) of each cell. The adjusted Rand index (ARI) score is calculated and shown in the plot.

Variants	Zheng68k		MacParland		Lung	
	Accuracy ↑	F1↑	Accuracy ↑	F1↑	Accuracy ↑	F1↑
SCTrans w/o MA	0.772	0.668	0.926	0.851	0.943	0.907
SCTrans w/o CCR	0.801	0.708	0.971	0.961	0.959	0.914
SCTrans	**0.817**	**0.717**	**0.981**	**0.966**	**0.965**	**0.921**

Table 3: Quantitative results of ablative models.

data. The visualizations clearly indicate that the representations obtained by our model are more distinguishable than those derived from the original gene data, leading to superior results in terms of ARI scores.

Ablation Study

To further validate the significance of multi-scale attention and contrastive learning in the proposed model, we compare the proposed SCTrans with two variants:

- Without the Multi-scale Attention strategy (SCTrans w/o MA): In this setting, only the single-scale attention will be considered.
- Without Conditional Contrastive Regularization (SCTrans w/o CCR): In this setting, we disable the conditional contrastive regularition term $\mathcal{L}_{ctrst}$ in Eq.(8).

In detail, we conduct an ablation study on the Zheng68k, MacParland, and Lung datasets to assess the impact of these two important components. As shown in Table 3, the proposed model with multi-scale attention and contrastive learning outperforms the two related variants, validating the effectiveness of these two components in the proposed method.

Additionally, we conduct an evaluation to assess the impact of the multi-scale attention strategy in terms of the reconstruction task. As depicted in Figure 10, the results clearly demonstrate that the implementation of the multi-scale attention strategy significantly enhances the reconstruction process, both in terms of efficiency and accuracy.

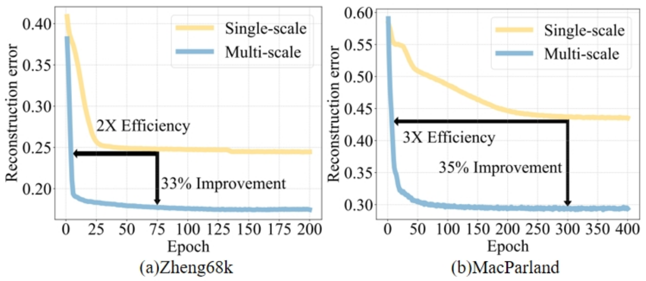

Figure 10: The impact of multi-scale attention and single scale attention about reconstruction tasks on the Zheng68k and MacParland datasets.

5 Conclusion

In this study, we present SCTrans, a novel multi-scale scRNA-seq Sub-vector Completion Transformer designed to overcome the limitations of current cell type annotation methods in scRNA-seq data analysis. SCTrans capitalizes on the richer expressiveness of gene sub-vectors and incorporates self-attention and contrastive regularization techniques. This approach allows SCTrans to excel in identifying crucial genes associated with specific cell types, leading to more accurate cell type annotation compared to current state-of-the-art methods. Furthermore, SCTrans performs attention computation among genes, potentially identifying salient ones, and provides interpretable results for downstream biological research and analysis, enhancing its utility in advancing our understanding of cellular behaviors and functions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189, 62106136), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), in part by the GuangDong Basic and Applied Basic Research Foundation (Project No. 2020A1515010484, 2022A1515011160, 2022A1515010434, 2023A1515030154), and in part by TCL Science and Technology Innovation Fund (Project No. 20231752).

Contribution Statement

L.L. and W.X. are joint first authors, conceived of and designed the computational method; C.L. and S.W are joint corresponding authors. L.L. implemented the main algorithm; L.L. and W.X. did experiments and interpreted the results; X.D.W. and W.J.S. provided an analysis of the biological insights of the experiments; W.X., C.L., S.W. and L.L. wrote the manuscript; W.X., C.L., S.W. and H.S.W revised the manuscript.

References

[Abdelaal *et al.*, 2019] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20:1–19, 2019.

[Abnar and Zuidema, 2020] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[Alsaigh *et al.*, 2022] Tom Alsaigh, Doug Evans, David Frankel, and Ali Torkamani. Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. *Communications biology*, 5(1):1084, 2022.

[Aran *et al.*, 2019] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

[Brbić *et al.*, 2020] Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, 17(12):1200–1206, 2020.

[Chen *et al.*, 2023] Jiawei Chen, Hao Xu, Wanyu Tao, Zhaoxiong Chen, Yuxuan Zhao, and Jing-Dong J Han. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, 2023.

[Connell *et al.*, 2022] William Connell, Umair Khan, and Michael J Keiser. A single-cell gene expression language model. *arXiv preprint arXiv:2210.14330*, 2022.

[Cortal *et al.*, 2021] Akira Cortal, Loredana Martignetti, Emmanuelle Six, and Antonio Rausell. Gene signature extraction and cell identity recognition at the single-cell level with cell-id. *Nature biotechnology*, 39(9):1095–1102, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2021] Alexy Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Hao *et al.*, 2021] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[Hao *et al.*, 2023] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Le Song, and Xuegong Zhang. Large scale foundation model on single-cell transcriptomics. *bioRxiv*, pages 2023–05, 2023.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Hu *et al.*, 2022] Hang Hu, Zhong Li, Xiangjie Li, Minzhe Yu, and Xiutao Pan. Sccaes: deep clustering of single-cell rna-seq via convolutional autoencoder embedding and soft k-means. *Briefings in Bioinformatics*, 23(1):bbab321, 2022.

[Ianevski *et al.*, 2022] Aleksandr Ianevski, Anil K Giri, and Tero Aittokallio. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature communications*, 13(1):1246, 2022.

[Kimmel and Kelley, 2021] Jacob C Kimmel and David R Kelley. Semisupervised adversarial neural networks for single-cell classification. *Genome research*, 31(10):1781–1793, 2021.

[Kiselev *et al.*, 2017] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[Kiselev *et al.*, 2018] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

[Kiselev *et al.*, 2019] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[Li *et al.*, 2020a] Chenwei Li, Baolin Liu, Boxi Kang, Zedao Liu, Yedan Liu, Changya Chen, Xianwen Ren, and Zemin Zhang. Scibet as a portable and fast single cell type identifier. *Nature communications*, 11(1):1818, 2020.

[Li *et al.*, 2020b] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, 11(1):2338, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Ma *et al.*, 2023] Anjun Ma, Xiaoying Wang, Jingxian Li, Cankun Wang, Tong Xiao, Yuntao Liu, Hao Cheng, Juexin

Wang, Yang Li, Yuzhou Chang, et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14(1):964, 2023.

[Meng *et al.*, 2023] Rui Meng, Shuaidong Yin, Jianqiang Sun, Huan Hu, and Qi Zhao. scaaga: Single cell data analysis framework using asymmetric autoencoder with gene attention. *Computers in Biology and Medicine*, 165:107414, 2023.

[Shao *et al.*, 2021] Xin Shao, Haihong Yang, Xiang Zhuang, Jie Liao, Penghui Yang, Junyun Cheng, Xiaoyan Lu, Huajun Chen, and Xiaohui Fan. scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic acids research*, 49(21):e122–e122, 2021.

[Tian *et al.*, 2021] Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Model-based deep embedding for constrained clustering analysis of single cell rna-seq data. *Nature communications*, 12(1):1873, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Neural Information Processing Systems*, 2017.

[Wan *et al.*, 2022] Hui Wan, Liang Chen, and Minghua Deng. scname: neighborhood contrastive clustering with ancillary mask estimation for scrna-seq data. *Bioinformatics*, 38(6):1575–1583, 2022.

[Wang *et al.*, 2017] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414–416, 2017.

[Wang *et al.*, 2022] Jiacheng Wang, Quan Zou, and Chen Lin. A comparison of deep learning-based pre-processing and clustering approaches for single-cell rna sequencing data. *Briefings in Bioinformatics*, 23(1):bbab345, 2022.

[Wu *et al.*, 2021] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvR: introducing convolutions to vision transformers. *arXiv:2103.15808*, 2021.

[Xu *et al.*, 2019] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. June 2019.

[Yang *et al.*, 2022] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[Zhai *et al.*, 2023] Yuyao Zhai, Liang Chen, and Minghua Deng. Realistic cell type annotation and discovery for single-cell rna-seq data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4967–4974, 2023.

[Ziegenhain *et al.*, 2017] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.