Transformer-based unsupervised contrastive learning for histopathological image classification

Xiyue Wang , Sen Yang, Jun Zhang, Minghui Wang
, Jing Zhang, Wei Yang, Junzhou Huang, Xiao Han

College of Biomedical Engineering, Sichuan University, Chengdu, China
College of Computer Science, Sichuan University, Chengdu, China
Tencent AI Lab, Shenzhen, China

Medical Image Analysis 2022

Report: Yu-Chen Lai

Data: 2023.03.23

Outline

- Introduction
- Method
- Experimental results and discussions
- Conclusion

1. Introduction

- The tremendous successes of **self-supervised learning (SSL)** techniques in the computer vision community have promoted the development of SSL in histopathological image analysis.
- There have been some published works, but these approaches process histopathological images by simply **applying existing contrastive learning (CL)-based SSL frameworks** (e.g., SimCLR and MoCo) or tailoring some histopathology-oriented SSL tasks on a **CNN-specific** backbone.

1. Introduction

- Three aspects that could be further improved.
  1. CL assigns two augmented views from the same instance as **one positive pair**, which limits the variability and diversity of positive samples.
  2. The learning of global context features is often **limited by the receptive field of CNN**.
  3. The data currently used for SSL training are relatively **homogeneous** and their **number is rather limited**.
- Contributions
  1. Semantically-relevant contrastive learning (**SRCL**) framework
  2. A hybrid model **CTransPath** as the backbone, which is designed by integrating a **CNN** and a **Swin Transformer** architecture
  3. The used database is the **largest** publicly available in the histopathology scenario(approximately 87T).

2. Methods

- This section presents an overview of our proposed SSL algorithm based on a semantically-relevant contrastive learning (SRCL) and a hybrid backbone (CTransPath).

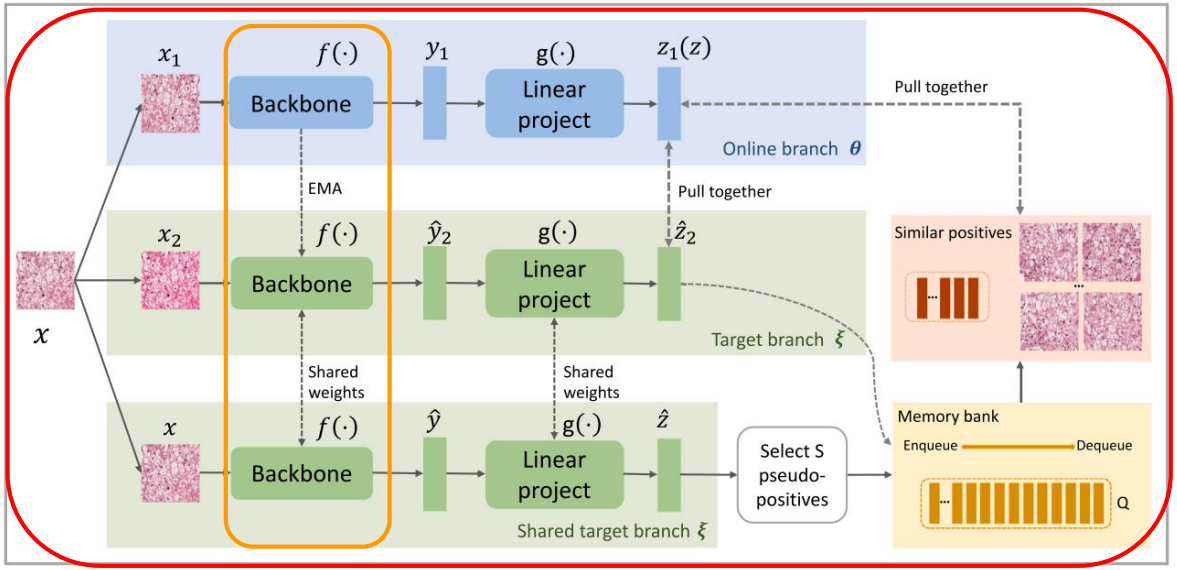
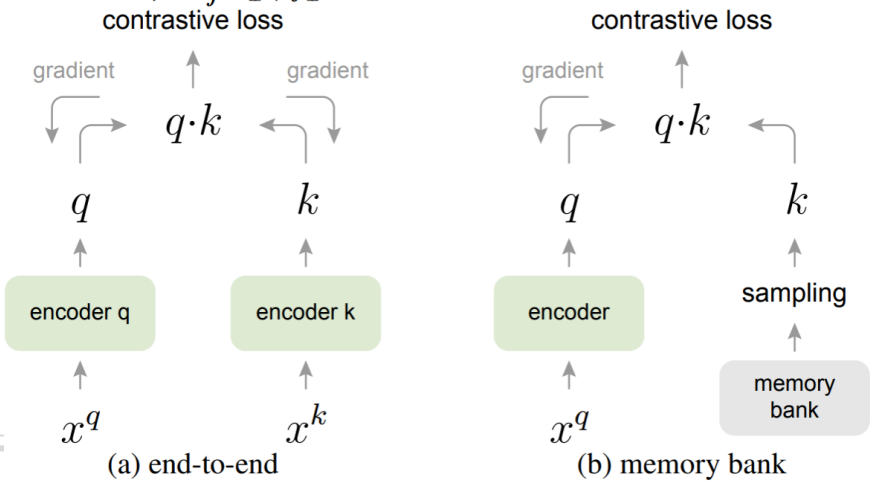


Fig. 1. An overview of our proposed SRCL approach for histopathological image applications. It is an improved framework based on MoCo v3 (Chen et al., 2021). The negative samples are stored in each mini-batch and the positives are from two paths: (i) two data augmentations of the current input image and (ii) top S semantically-relevant images identified by comparing the current input feature with samples in the memory bank. Based on the above design, a semantically-relevant contrastive loss is proposed to guide the network training.

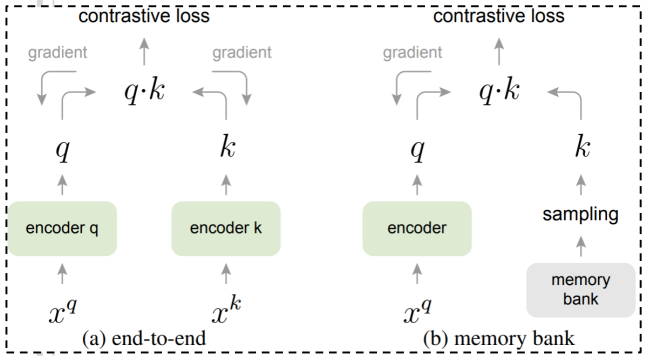
2.1. Problem formulation

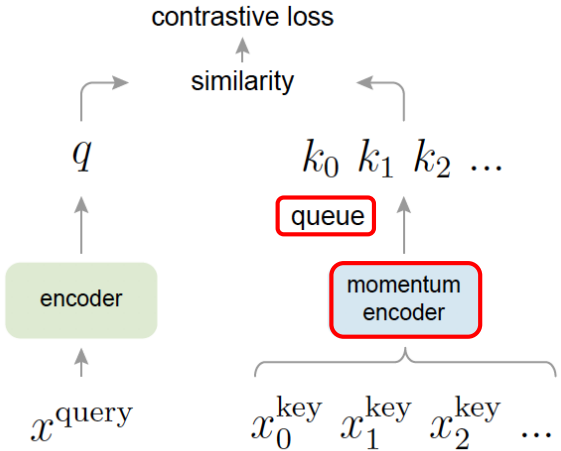
- Let $D^u = \{\mathbf{x}_i^u\}_{i=1}^N$ denote the unlabeled dataset used for SSL pretraining
- The CL-based SSL method performs two data augmentations on two network branches for each sample, generating $D^q = \{\mathbf{x}_i^q\}_{i=1}^N$ and $D^k = \{\mathbf{x}_j^k\}_{j=1}^N$
- The two data augmentations from the same input are regarded as positive pairs
- $D^{pos} = \{\mathbf{x}_i^q, \mathbf{x}_j^k\}_{[\![i=j]\!]}$ while data augmentations from different images are used to form negative pairs $D^{neg} = \{\mathbf{x}_i^q, \mathbf{x}_j^k\}_{[\![i\neq j]\!]}$



2.1. Problem formulation

- Moco V1:
  - **Dictionary as a queue**: The current mini-batch is enqueued to the dictionary, and the oldest mini-batch in the queue is removed.
  - **Momentum update:** Rapidly changing encoder reduces the key representations' consistency. Formally, denoting the parameters of f_k as θ_k and those of f_q as θ_q, we update θ_k by: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$





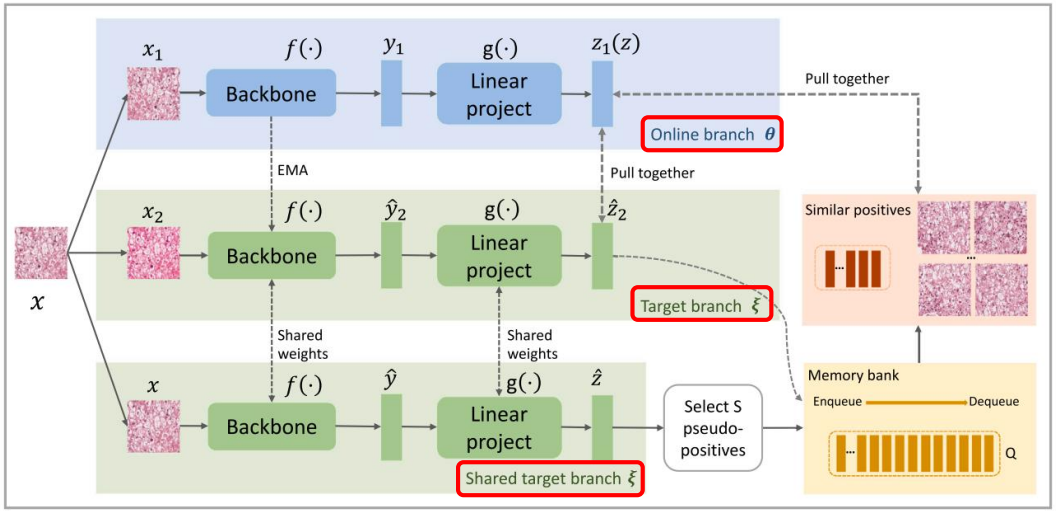
[1] Kaiming He, et al. "Momentum contrast for unsupervised visual representation learning." CVPR. 2020.

2.2. Semantically-relevant contrastive learning

- For **histopathological images**, there are **a large number of similar patches** (i.e., patches with similar cellular and tissue compositions) both within and across WSIs, which are defined as semantically relevant samples.
- Thus, the positive pairs should be counted more instead of fixed one pair in the traditional CL setting.

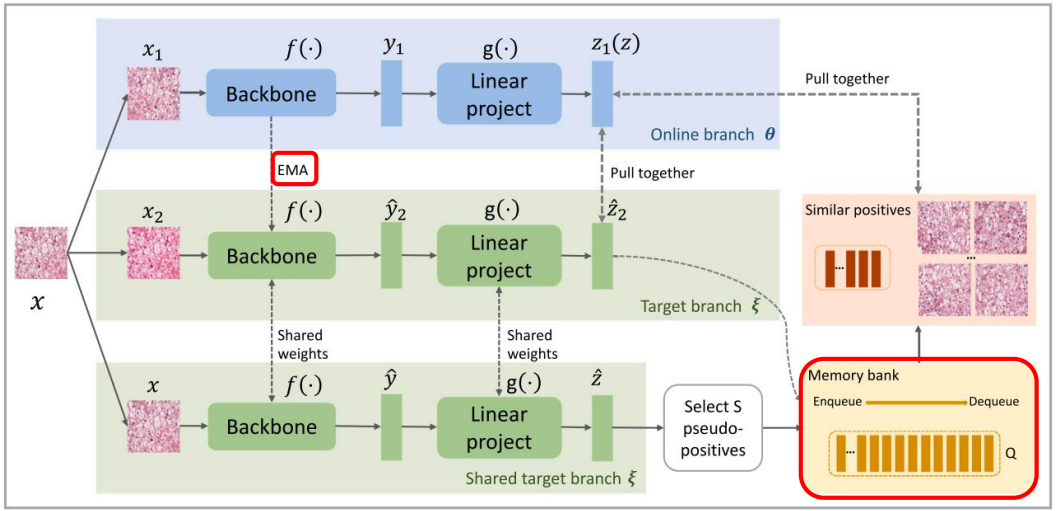
2.2. Semantically-relevant contrastive learning

- There are three parallel paths: online, target, and shared target branches for encoding three different views of the input.
  - **Target branch**: Refresh the memory bank as training proceeds.
  - **Shared target branch**: Generate a query to retrieve semantically-similar samples from the memory bank.



2.2. Semantically-relevant contrastive learning

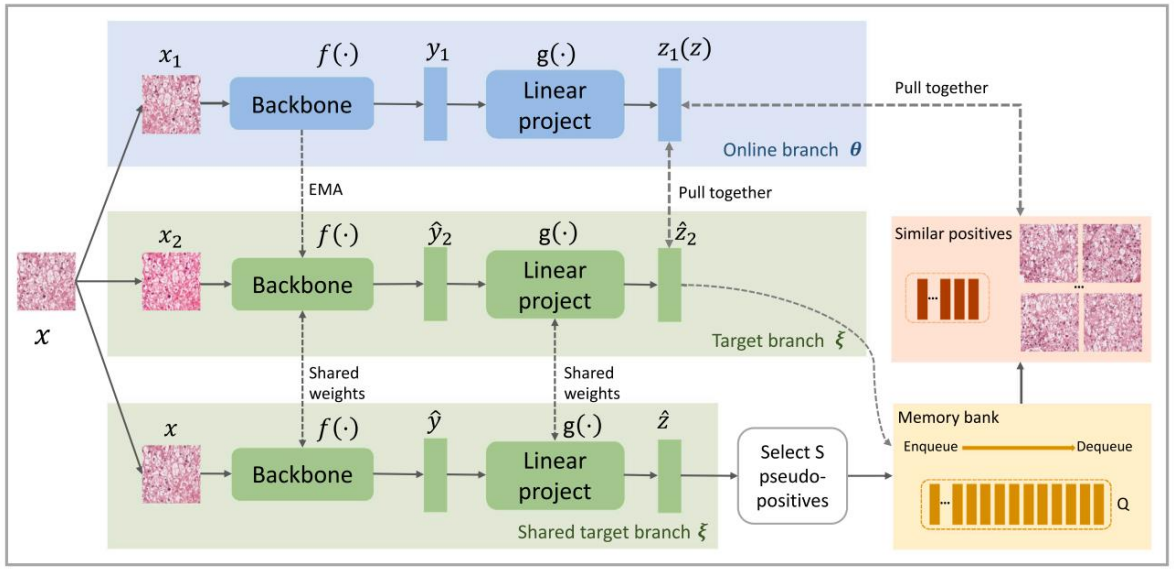
- Similar to Moco v1[**1**]
  - **Dictionary as a queue**: Memory bank that is constructed by enqueueing the features from the target branch during training, which is updated at the end of each iteration.
  - **Momentum update:** Train the online branch with parameter $\boldsymbol{\theta}$, update the target branch with parameter $\boldsymbol{\xi}$ by $\boldsymbol{\xi} \leftarrow m\boldsymbol{\xi} + (1-m)\boldsymbol{\theta}$.



*[**1**] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." CVPR. 2020.*

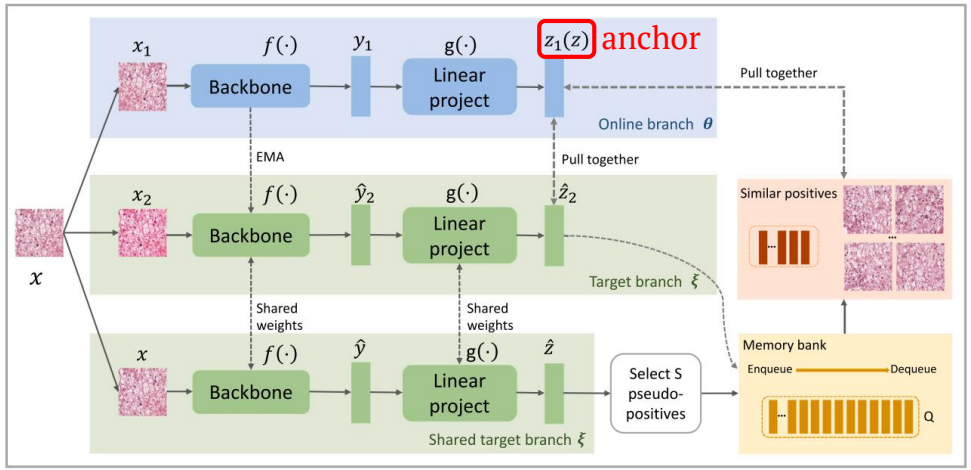
2.2. Semantically-relevant contrastive learning

- Given a histopathological image patch $\mathbf{x}$, it generates two augmentations ($\mathbf{x}_1$ and $\mathbf{x}_2$).
- x1 pass through the online branch: $\mathbf{y}_1 = f^{\theta}(\mathbf{x}_1),\ \mathbf{z}_1 = g^{\theta}(\mathbf{y}_1)$
- x2 pass through the target branch: $\hat{\mathbf{y}}_2 = f^{\xi}(\mathbf{x}_2),\ \hat{\mathbf{z}}_2 = g^{\xi}(\hat{\mathbf{y}}_2)$



3.2. Semantically-relevant contrastive learning

- In conventional CL method, anchor **z** has one positive sample $\hat{\mathbf{z}}_2$.
- To obtain **more positive samples**, we aim to find samples that are visually similar to **z**.
- For this purpose, the top *S* samples with the highest **cosine similarity** are taken as the new positives for anchor **z**.
- Combining the original positive sample $\hat{\mathbf{z}}_2$ in conventional CL, the total number of positive pa



2.2. Semantically-relevant contrastive learning

- Semantically-relevant contrastive loss L_{SRCL}:

$$\mathcal{L}_2(\mathbf{z}, \mathbf{z}^+, \mathbf{z}^-) = -\log \frac{\sum_{i=1}^{S+1} \exp\left(\mathbf{z}_i^+ \cdot \mathbf{z}/\tau\right)}{\sum_{i=1}^{S+1} \exp\left(\mathbf{z}_i^+ \cdot \mathbf{z}/\tau\right) + \sum_{j=1}^{N} \exp\left(\mathbf{z}_j^- \cdot \mathbf{z}/\tau\right)}$$

.

$$\mathcal{L}_{SRCL} = \frac{1}{2}\mathcal{L}_2(\mathbf{z}_1, \hat{\mathbf{z}}_2, \mathbf{z}^-) + \frac{1}{2}\mathcal{L}_2(\mathbf{z}_2, \hat{\mathbf{z}}_1, \mathbf{z}^-)$$

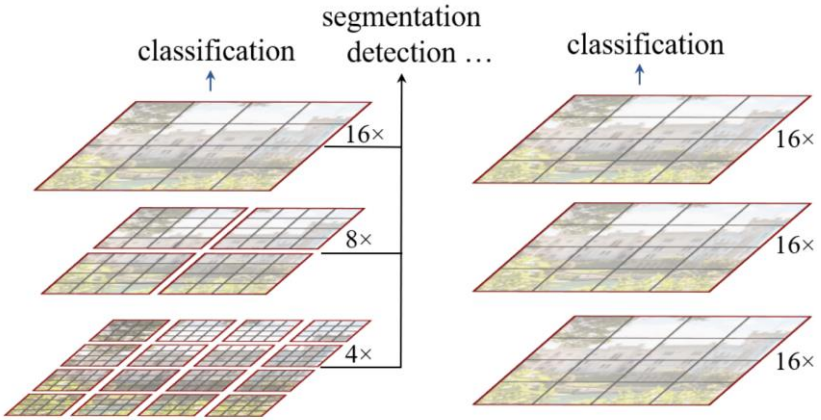
- where $\mathbf{z}$ represents an anchor sample.

  $\mathbf{z}+$ and $\mathbf{z}-$ denote the positive and negative features of the anchor feature.

  S+1 and N represent the number of positive and negative pairs.

2.3. Backbone construction

- Why Swin Transformer[2] better than ViT in computer vision?
  1. Hierarchical feature maps by merging image patches.
  2. **Shifted window** approach for computing self-attention.



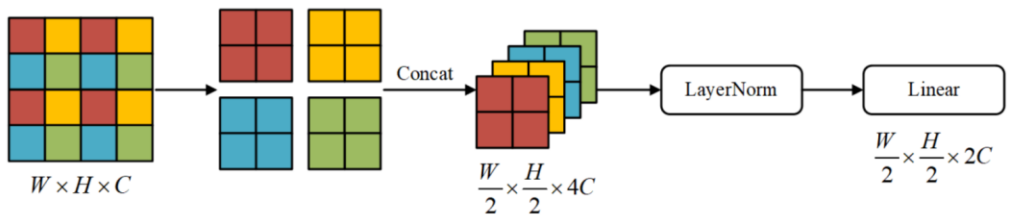
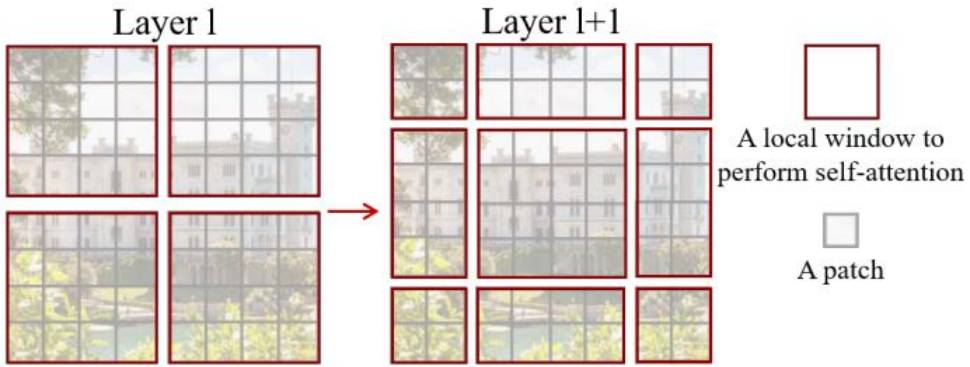


Fig. 4. Patch Merging structure diagram.

[2] Ze Liu, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV. 2021.

2.3. Backbone construction

- Why Swin Transformer[2] better than ViT in computer vision?
  1. Hierarchical feature maps by merging image patches.
  2. **Shifted window** approach for computing self-attention.



[2] Ze Liu, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV. 2021.

2.3. Backbone construction

- Why Swin Transform[2] better than ViT in computer vision?
  1. Hierarchical feature maps by merging image patches.
  2. **Shifted window** approach for computing self-attention.

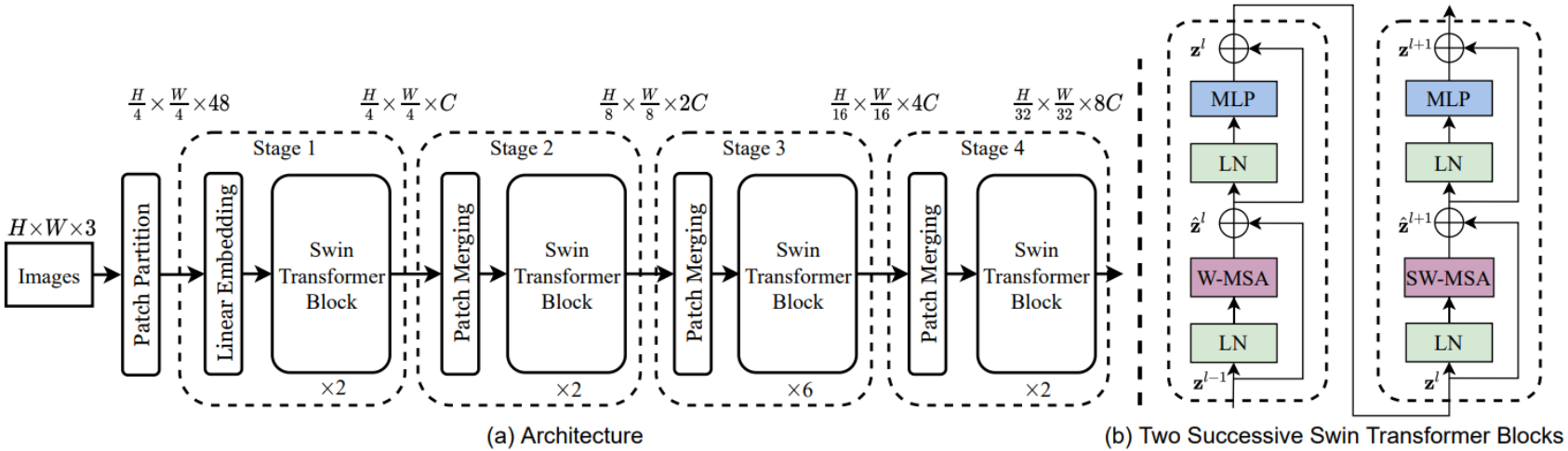
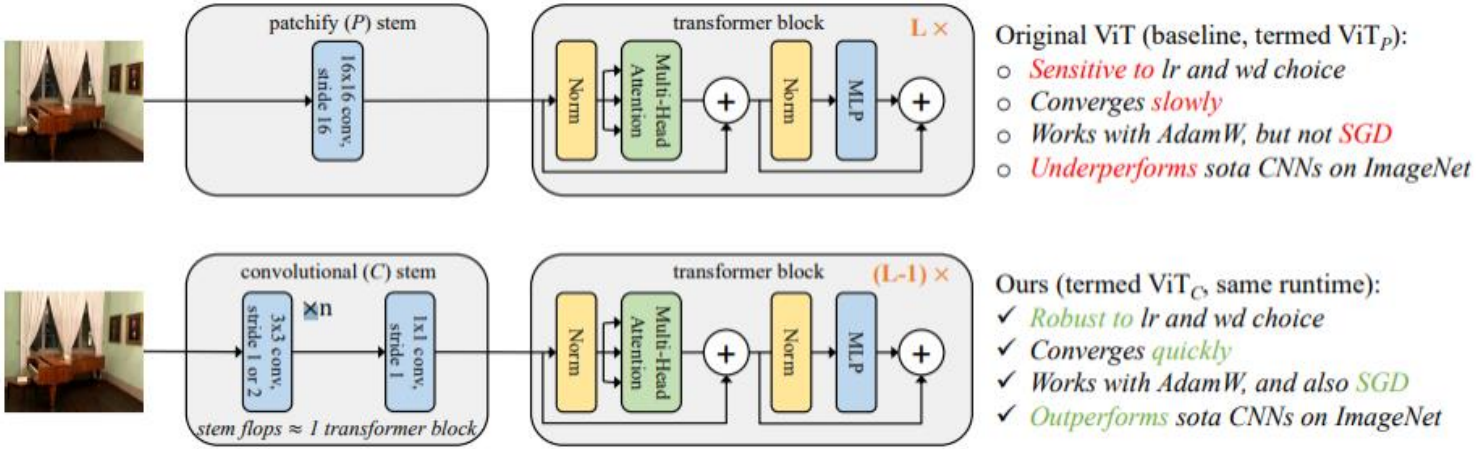


Figure 3. (a) The architecture of a Swin Transformer (Swin-T)

[2] Ze Liu, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV. 2021.

2.3. Backbone construction

- Problem of Transformer architecture
  - [3] indicated that the Transformer architecture is much **harder to optimize** compared with CNNs, mainly due to the **patch projection** implemented through large-kernel large-stride convolution operations.

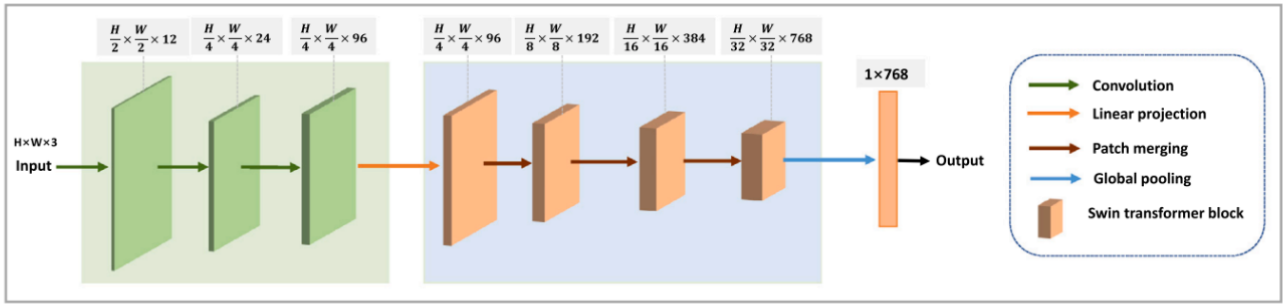


[3] Tete Xiao, et al. "Early convolutions help transformers see better."NIPS. 2021.

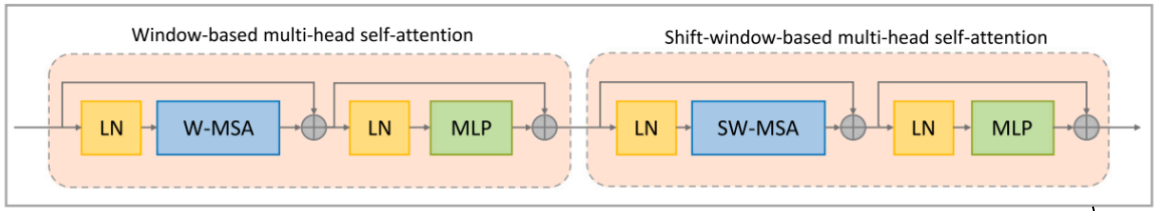
2.3. Backbone construction

- The CNN module is designed with three consecutive convolutional layers with kernel sizes of 3 × 3, 3 × 3, and 1 × 1

A.



B.



$$\hat{\mathbf{z}}^{l} = \text{W–MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}$$
$$\mathbf{z}^{l} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l}$$
$$\hat{\mathbf{z}}^{l+1} = \text{SW–MSA}\left(\text{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l}$$
$$\mathbf{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1}$$

3. Experimental results and discussions

- Five types of downstream experiments
  - patch retrieval
  - patch classification
  - weakly-supervised WSI classification
  - mitosis detection
  - colorectal adenocarcinoma gland segmentation
- These experiments include ablation study, comparisons with state-of-the-art methods on these downstream datasets, and comparisons with different network pretraining methods.

3.1. Datasets

Dataset	# of WSIs	# of types	magnification	patchsize	# of patches
TCGA	29,763	32	20×	1024 × 1024	14,325,848
PAIP	2,457	6	20×	1024 × 1024	1,254,414
UniToPatho	292	4	20×	1812 × 1812	8,699
TissueNet	1,016	4	-	1200 × 1200	5,926
NCT-CRC-HE	86	9	-	224 × 224	107,180
Colorectal cancer (CRC)	-	8	-	150 × 150	5000
Camelyon16	399	2	40×	-	-
TCGA-NSCLC	993	2	-	-	-
TCGA-RCC	884	3	-	-	-
MIDOG	150	-	-	256 × 256	79,399
CRAG	-	-	-	1512 × 1516	213

3.2. Experimental setups in the pretraining stage

- Use SRCL-based framework to train the CTransPath model
  - Mini-batch: 1,024.
  - Histopathology-oriented data augmentation strategies [**4**]
    - random cropping, Gaussian blur, and hue and saturation shifting in the HSV color space.
  - τ: 0.2 (Following MoCo v3 [**5**])
  - Optimizer: AdamW
  - The number of new positive pairs S: 4
  - Epochs: 100

*[**4**] David Tellez, et al. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology." Medical image analysis. 2019.*
*[**5**] Xinlei Chen, Saining Xie, and Kaiming He. "An empirical study of training self-supervised vision transformers." ICCV. 2021.*

3.3. Evaluation metrics

- For the classification task
  - Accuracy (ACC)
  - Area under the curve (AUC) score
  - F1 score
- For image retrieval
  - ***ACC@k:*** *ACC@k* = 1 if any one of the top-k returns has the same label as the query image
  - ***mMV@k:*** *mMV@k* = 1 only if the majority of these retrieved images have the same label as the query image
- For the detection and segmentation tasks
  - F1 scores
  - Dice scores

3.3. Ablation study

- Benefit of in-domain SSL pretraining
- Benefit of hybrid CNN and Transformer encoder
- Benefit of semantically-relevant positives

Table 1
Ablation study. Sup. denotes the supervised pretraining process. ImageTrans and HistoTrans denote ImageNet-pretrained Swin Transformer and histopathology-pretrained Swin Transformer, respectively. HistoTrans+CNN means our *CTransPath* backbone. SN denotes spatial-neighbor-based contrastive learning method.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
ImageTrans (Sup.)	0.5324	0.7892	0.8799	0.5035	0.5334	0.7899	0.8708	0.5463
ImageTrans (SSL)	0.5618	0.8171	0.9047	0.5565	0.5749	0.8103	0.8803	0.5799
HistoTrans+CL	0.6051	0.8395	0.9220	0.5910	0.5935	0.8194	0.8891	0.6011
HistoTrans+CNN+CL	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
HistoTrans+CNN+SN	0.6304	0.8488	0.9158	0.6363	0.6201	0.8305	0.8928	0.6332
HistoTrans+CNN+SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

3.3. Ablation study

- Benefit of in-domain SSL pretraining
- Benefit of hybrid CNN and Transformer encoder
- Benefit of semantically-relevant positives

Table 1
Ablation study. Sup. denotes the supervised pretraining process. ImageTrans and HistoTrans denote ImageNet-pretrained Swin Transformer and histopathology-pretrained Swin Transformer, respectively. HistoTrans+CNN means our *CTransPath* backbone. SN denotes spatial-neighbor-based contrastive learning method.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
ImageTrans (Sup.)	0.5324	0.7892	0.8799	0.5035	0.5334	0.7899	0.8708	0.5463
ImageTrans (SSL)	0.5618	0.8171	0.9047	0.5565	0.5749	0.8103	0.8803	0.5799
HistoTrans+CL	0.6051	0.8395	0.9220	0.5910	0.5935	0.8194	0.8891	0.6011
HistoTrans+CNN+CL	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
HistoTrans+CNN+SN	0.6304	0.8488	0.9158	0.6363	0.6201	0.8305	0.8928	0.6332
HistoTrans+CNN+SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

3.3. Ablation study

- Benefit of in-domain SSL pretraining
- Benefit of hybrid CNN and Transformer encoder
- Benefit of semantically-relevant positives

Table 1
Ablation study. Sup. denotes the supervised pretraining process. ImageTrans and HistoTrans denote ImageNet-pretrained Swin Transformer and histopathology-pretrained Swin Transformer, respectively. HistoTrans+CNN means our *CTransPath* backbone. SN denotes spatial-neighbor-based contrastive learning method.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
ImageTrans (Sup.)	0.5324	0.7892	0.8799	0.5035	0.5334	0.7899	0.8708	0.5463
ImageTrans (SSL)	0.5618	0.8171	0.9047	0.5565	0.5749	0.8103	0.8803	0.5799
HistoTrans+CL	0.6051	0.8395	0.9220	0.5910	0.5935	0.8194	0.8891	0.6011
HistoTrans+CNN+CL	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
HistoTrans+CNN+SN	0.6304	0.8488	0.9158	0.6363	0.6201	0.8305	0.8928	0.6332
HistoTrans+CNN+SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

3.3. Ablation study

Table 2
Effect of different values of S for the number of pseudo-positives on the patch retrieval accuracy.

		$S = 0$	$S = 1$	$S = 2$	$S = 4$	$S = 6$	$S = 8$
TissueNet	ACC@1	0.6239	0.6333	0.6409	**0.6505**	0.6498	0.6417
	mMV@5	0.6247	0.6338	0.6527	**0.6617**	0.6610	0.6537
UniToPatho	ACC@1	0.6183	0.6194	0.6211	**0.6329**	0.6271	0.6223
	mMV@5	0.6294	0.6340	0.6370	**0.6417**	0.6370	0.6311

Table 3
Effect of different number of epochs for warmup on the patch retrieval accuracy.

Epoch		0	2	5	10
TissueNet	ACC@1	0.6304	0.6417	0.6505	0.6493
	mMV@5	0.6358	0.6525	0.6617	0.6606
UniToPatho	ACC@1	0.6209	0.6286	0.6329	0.6309
	mMV@5	0.6332	0.6363	0.6417	0.6407

3.4. Comparison with other SSL methods

Table 4
Results of patch retrieval and comparison with other state-of-the-art SSL frameworks. Note that the implementation of all other SSL methods is based on their publicly available code but the backbone model and the training data are switched to be the same as ours.

Methods	TissueNet				UniToPatho			
	ACC@1	ACC@3	ACC@5	mMV@5	ACC@1	ACC@3	ACC@5	mMV@5
SimCLR (Chen et al., 2020)	0.6019	0.8297	0.9036	0.6021	0.6070	0.8149	0.8819	0.6170
MoBY (Xie et al., 2021)	0.6131	0.8360	0.9077	0.6077	0.6110	0.8197	0.8873	0.6173
DINO (Caron et al., 2021)	0.6169	0.8481	0.9183	0.6183	0.6149	0.8224	0.8909	0.6222
MoCo v3 (Chen et al., 2021)	0.6239	0.8405	0.9109	0.6247	0.6183	0.8234	0.8837	0.6294
SRCL (ours)	**0.6505**	**0.8606**	**0.9261**	**0.6617**	**0.6329**	**0.8370**	**0.8966**	**0.6417**

3.5. Results of patch classification

Table 5
Linear evaluation results on NCT-CRC-HE dataset with different sizes of training data. ImageTrans (Sup.) and ImageTrans (SSL) refer to models pre-trained using the ImageNet data in a supervised and self-supervised manner, respectively. All other compared SSL frameworks are pretrained using our training data. A supervised baseline using 100% of the training data achieves an F1 score of 0.9295 and an ACC of 0.9458.

Methods	Backbone	Percentage of training data									
		0.5%		1%		10%		50%		100%	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
ImageTrans (Sup.)	Swin	0.4770	0.5703	0.5584	0.6157	0.7780	0.8139	0.8206	0.8585	0.8323	0.8705
ImageTrans (SSL)	Transformer	0.6997	0.7348	0.7816	0.8213	0.8715	0.9035	0.8867	0.9171	0.8903	0.9216
SimCLR	ResNet50	0.8062	0.8503	0.8331	0.8649	0.8613	0.9195	0.8971	0.9269	0.9025	0.9315
BYOL		0.8234	0.8636	0.8649	0.8965	0.8876	0.9245	0.9050	0.9324	0.9144	0.9413
SimSiam		0.8348	0.8730	0.8660	0.9054	0.8903	0.9286	0.9085	0.9387	0.9144	0.9457
MoCo v2		0.8435	0.8875	0.8702	0.9130	0.9050	0.9362	0.9156	0.9467	0.9213	0.9514
SimCLR	CTransPath	0.8204	0.8581	0.8439	0.8716	0.8740	0.9297	0.9043	0.9316	0.9081	0.9349
MoBY		0.8317	0.8699	0.8578	0.8965	0.8919	0.9357	0.9144	0.9442	0.9156	0.9465
DINO		0.8355	0.8799	0.8671	0.9128	0.8915	0.9369	0.9135	0.9438	0.9198	0.9502
MoCo v3		0.8682	0.8978	0.8739	0.9228	0.9046	0.9415	0.9208	0.9516	0.9254	0.9548
SRCL (ours)		**0.8988**	**0.9266**	**0.9334**	**0.9539**	**0.9420**	**0.9635**	**0.9474**	**0.9648**	**0.9482**	**0.9652**

3.5. Results of patch classification

Table 6
Results of downstream classification tasks performed on CRC dataset (SVM classification)

Methods	ACC
Combined texture descriptors (Kather et al., 2016)	87.40
Ensemble of DNNs (Ghosh et al., 2021)	92.83
Fine-tuned VGG-19 (Faust et al., 2018)	93.58
KimiaNet (Riasatian et al., 2021)	96.80
Ensemble of CNNs (Nanni et al., 2021)	97.60
Ours	**98.20**

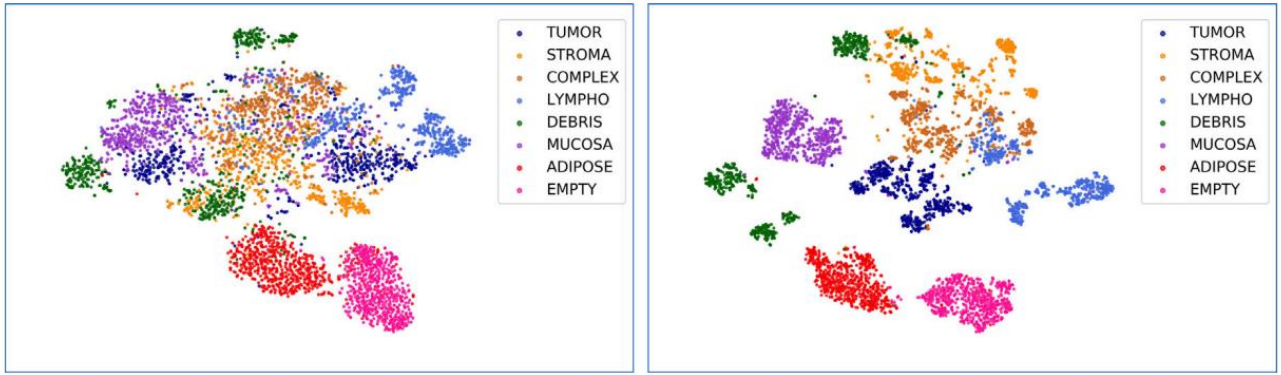


Fig. 3. Visualization (t-SNE) of the classification performance of all images in the CRC dataset based on the features generated from KimiaNet (left) and SRCL-pretrained *CTransPath* (right).

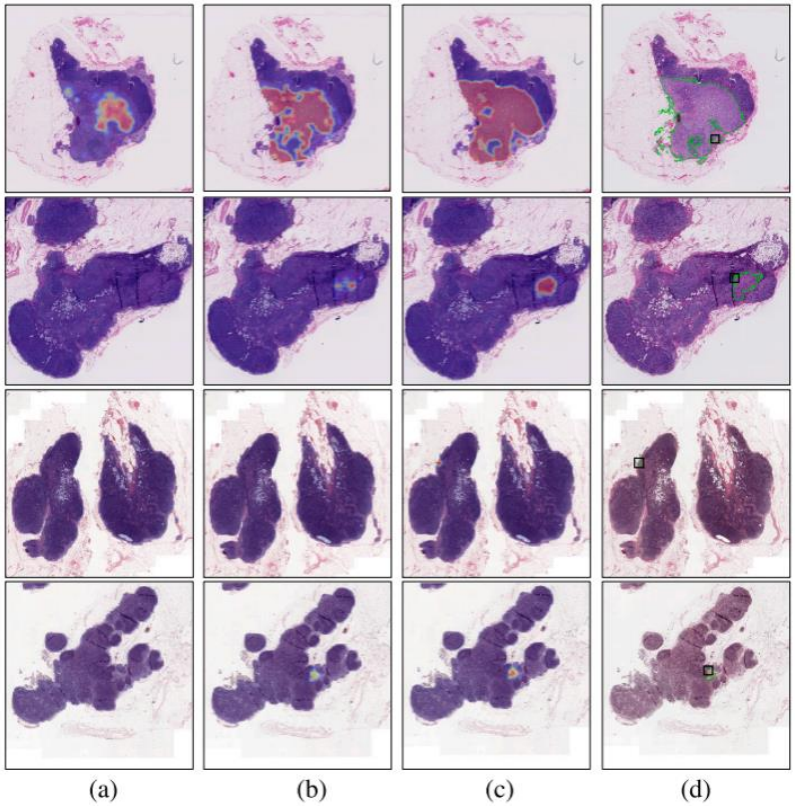
3.6. Results of weakly-supervised WSI classification

- The result are averaged over 5-fold cross-validation
- Note that the results of TransMIL and DSMIL are copied directly from their publications.

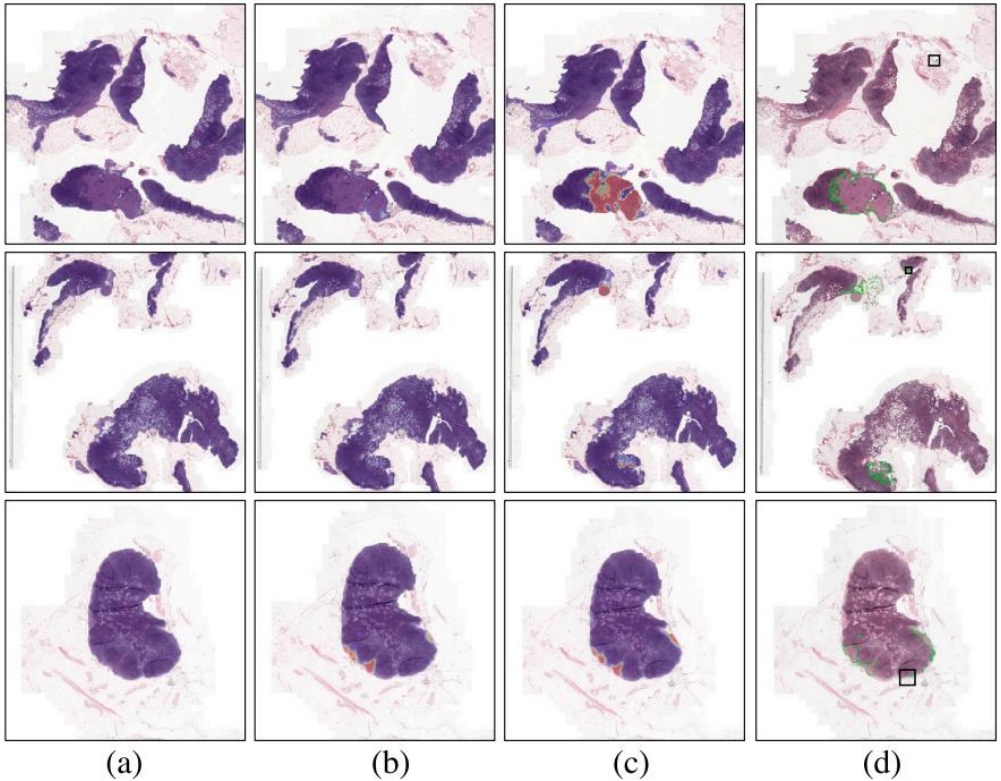
Table 7
Results of weakly-supervised classification on three public datasets.

	CAMELYON16		TCGA-NSCLC		TCGA-RCC	
	ACC	AUC	ACC	AUC	ACC	AUC
Kernel ATT (Rymarczyk et al., 2021)	0.773	0.804	0.841	0.921	0.856	0.945
C2C (Sharma et al., 2021)	0.809	0.841	0.849	0.921	0.909	0.972
MIL-RNN (Campanella et al., 2019)	0.819	0.856	0.856	0.931	0.914	0.974
AbMIL (Ilse et al., 2018)	0.820	0.857	0.838	0.920	0.902	0.980
CLAM-MB (Lu et al., 2021)	0.835	0.854	0.863	0.938	0.925	0.988
CLAM-SB (Lu et al., 2021)	0.837	0.873	0.859	0.938	0.921	0.987
TransMIL (Shao et al., 2021)	0.884	0.931	0.884	0.960	0.947	0.988
DSMIL (Li et al., 2021b)	0.899	0.917	**0.929**	0.958	–	–
CLAM-SB + Ours	**0.922**	**0.942**	0.912	**0.973**	**0.967**	**0.991**

3.7. Visualization for some good cases and bad cases



CLAM-SB DSMIL SRCL method Ground truth



CLAM-SB DSMIL SRCL method Ground truth

3.8. Results of downstream detection and segmentation tasks

- Faster RCNN and U-Net are employed as the detection and segmentation frameworks.

Table 8
Results of downstream mitosis detection and colorectal adenocarcinoma gland segmentation tasks via full network fine-tuning. The ImageTrans adopts Swin Transformer as the encoder and the four compared SSL frameworks employ our *CTransPath* as the encoder.

Model	Mitosis detection (F1)	CRAG segmentation (Dice)
ImageTrans (Sup.)	0.6842	0.8743
ImageTrans (SSL)	0.6958	0.8824
SimCLR	0.7078	0.8962
MoBY	0.7110	0.9010
DINO	0.7083	0.8996
MoCo v3	0.7204	0.9050
Ours	**0.7332**	**0.9156**

3.3. Conclusion

- We propose a customized SSL architecture for various histopathological image analysis, which contains a hybrid CNN-transformer backbone (CTransPath) and a semantically-relevant contrastive learning (SRCL) strategy.
- **CTransPath** makes use of both local and global receptive fields to extract discriminative and rich features.
- **SRCL** aims to select more semantically relevant positives to increase the sample diversity in the instance discrimination process.
- Our SRCL pretrained CTransPath on large-scale histopathological images has the potential to benefit various downstream tasks by transfer learning or direct feature extraction.

Thanks For Listening !