Automatic Microarray Gridding using Transfer Learning

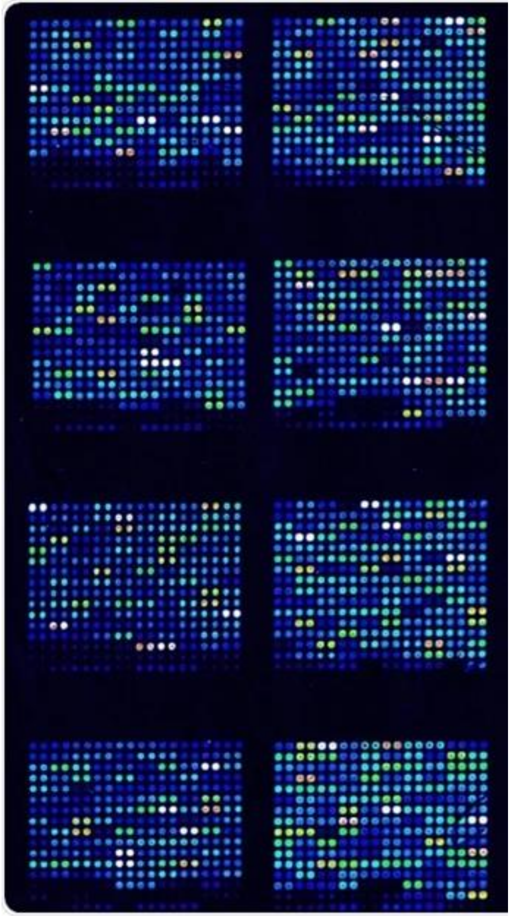
Advisor： Tien-Hao Chang
Chien-Sheng Chen
Guan-Da Syu
Student： Li-Fan Hsieh

Department of Electrical Engineering
National Cheng Kung University, Taiwan

Introduction

Microarray (微陣列晶片、生物晶片)

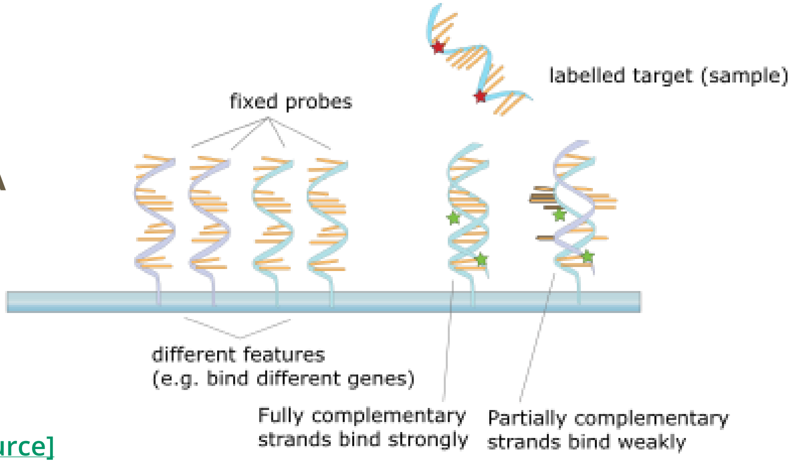
- A chip with only a few cm^2 that is used to biological features (or **biomarkers**, such as **protein, DNA segment**)
- The signal intensity of a **biosensor** reflects the existence of a specific biomarker. Cover the whole chip with test subject's serum, then thousands of biomarkers are simuteniously measured.
- Applications of microarrays are to diagnose diseases, or to discover new biomarkers to diagnose diseases



[source]

DNA Microarray

- Also called gene chip
- Use complementary DNA strands of the target DNA we want to detect are as probes
- Measure signal that depends on the binding conditions, such as temperature or fluorescence labeling on the target DNA

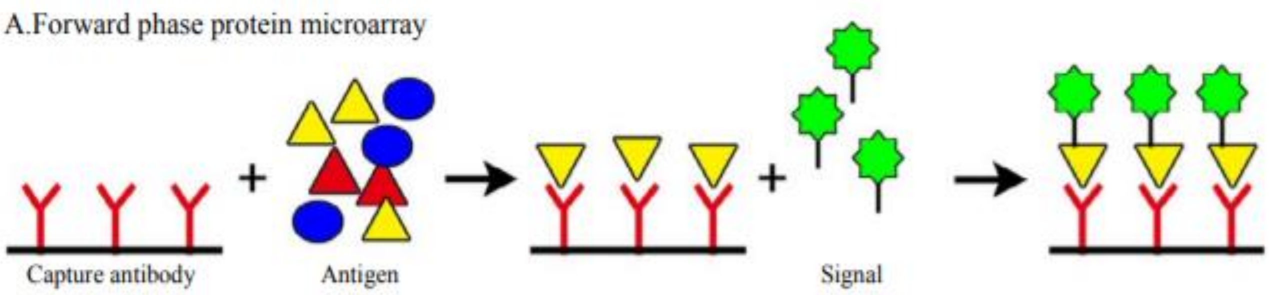


labelled target (sample)
fixed probes
different features
(e.g. bind different genes)
Fully complementary strands bind strongly
Partially complementary strands bind weakly

[source]

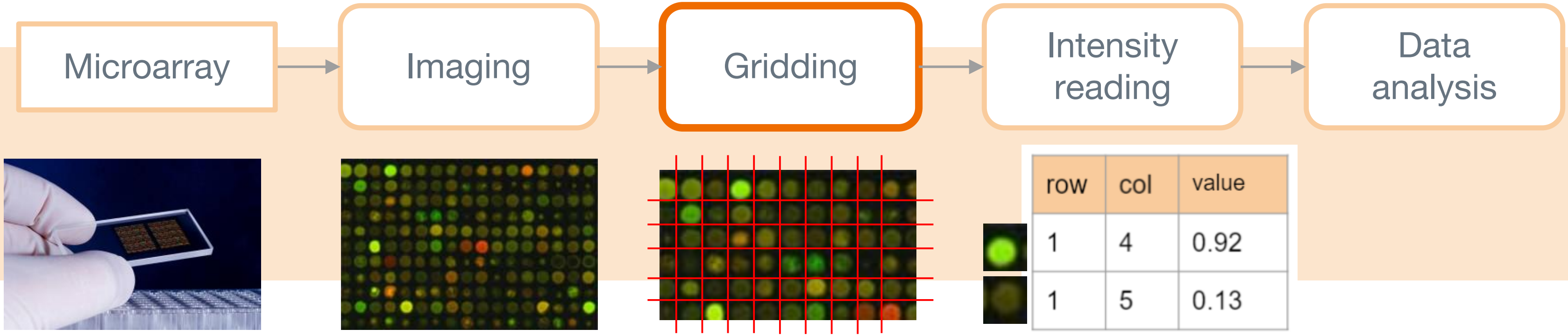
Proteome Microarray

- Utilize antibody as probes to capture specific antigen, or reversed
- Measure the fluorescence labeling on the paired proteins
- It is usually protein, rather than gene, that has the functional role in cell response.



[source]

Microarray Analysis

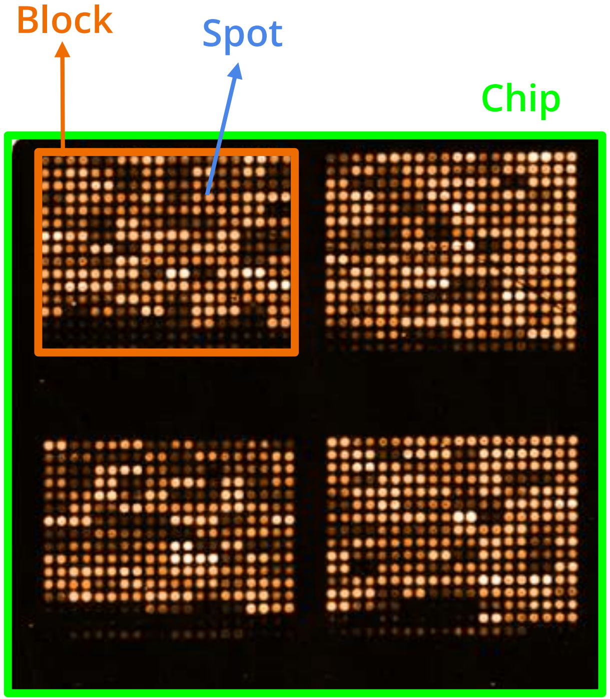


Microarray → Imaging → Gridding → Intensity reading → Data analysis

row	col	value
1	4	0.92
1	5	0.13

Gridding

- A microarray **chip** consists of multiple **blocks**, each block is a 2D array of circular **spots** with fixed number of row and columns, a spot can be indexed by <block, row, column>
- Gridding is the process of finding the location of each spot on the image
- Gridding can be done manually, but it is a very time-consuming process to label every single spot (thousands of spots per chip). Many efforts have been made to automate this process.

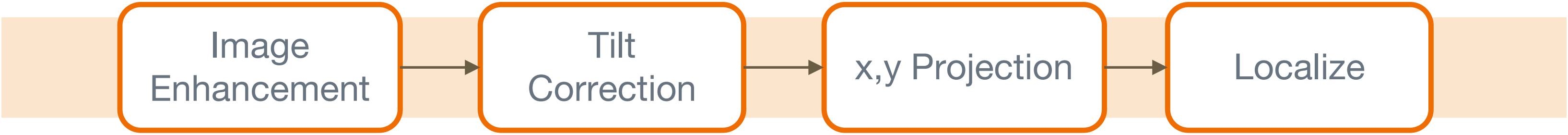


Related Work

Related Work

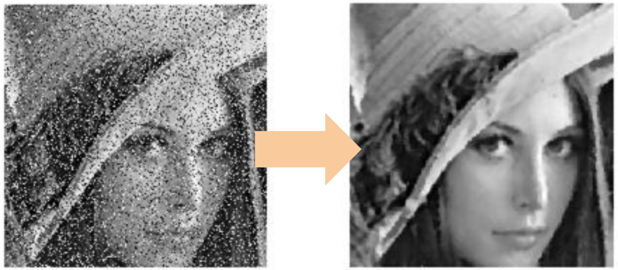
- **Genepix Pro 6.0 (Genepix)**
  - A broadly used commercial software for microarray analysis, provides tools for both manual and automatic gridding.
- **A Fully Automated Gridding Technique for Real Composite cDNA Microarray Images, 2020 (Joseph)**
  - IEEE Access, 2020
  - Uses various image processing algorithm, newest SOTA paper we found

Many previous works ([5] [6] [7]) are based on the following pipeline including Joseph



Previous Methods

Image enhancement: A series of denoising or contrast enhancement algorithms. Joseph uses median filter and dehazing algorithm.

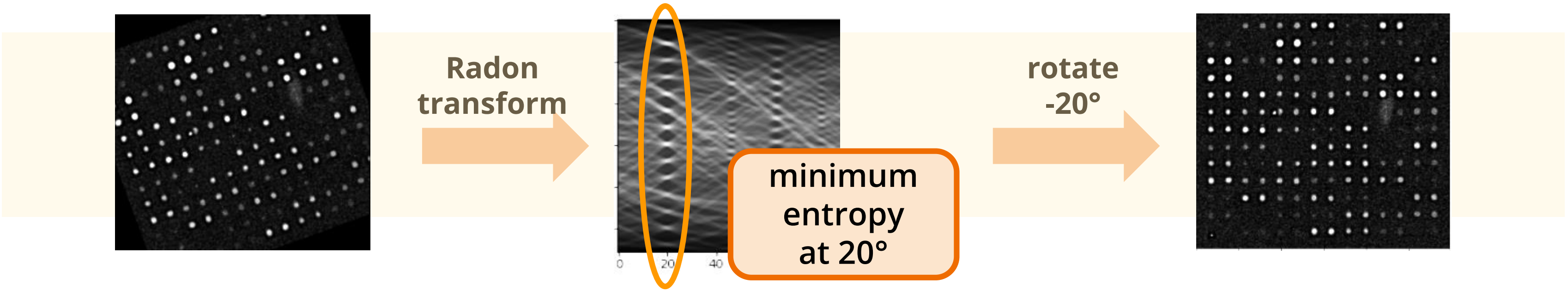


median filter

image dehaze

Previous Methods

Tilt correction: Rotate the image so that the spots are aligned with the x, y axis of the image. Usually done by **Radon transform**. (appendix)



Radon transform

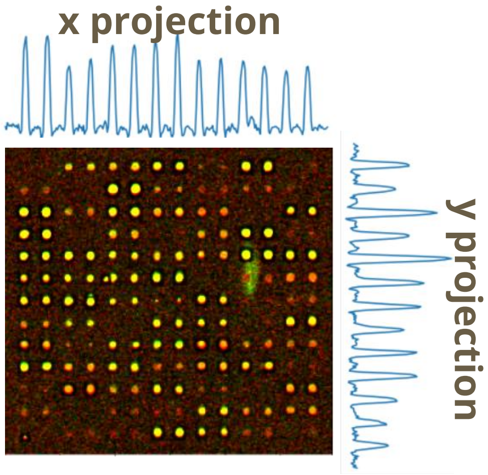
minimum entropy at 20°

rotate -20°

Previous Methods

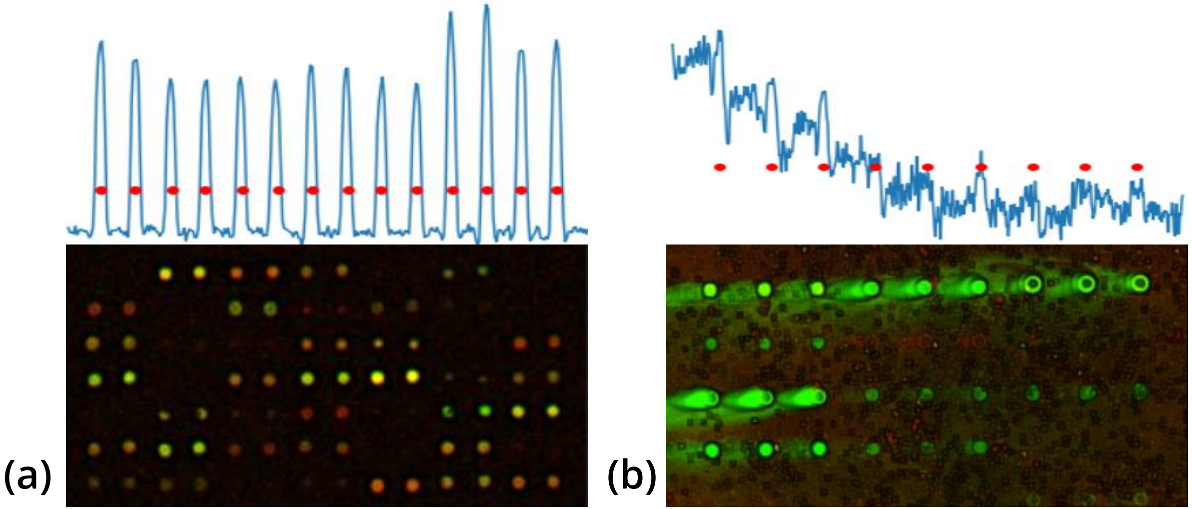
x, y projections: The 1D projection of the image on x axis and y axis.

Localize: Get the coordinates of the seperating lines from the x, y projections. Joseph's method simply binarize the profiles by Otsu's method to get the center positions of valleys.



Problems of Previous Methods

Previous algorithms rely heavily on rule-based image processing, but some unexpected disruption may exist in some datasets, resulting in noisy projections.



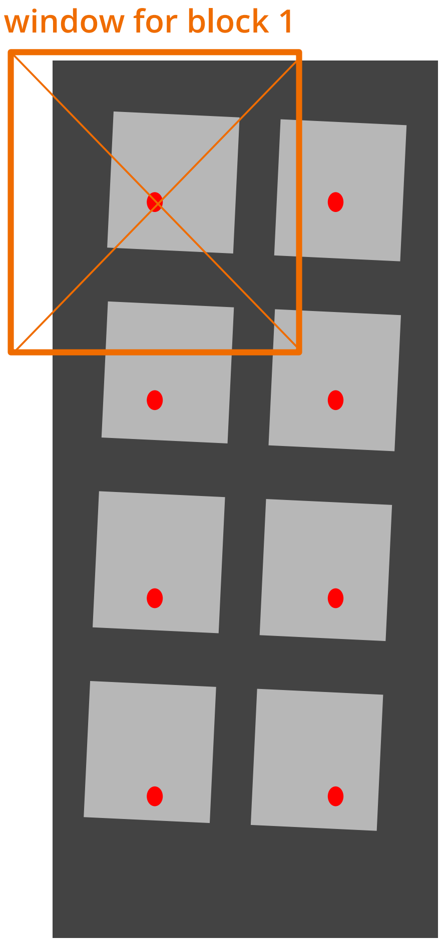
(a) Image from PS dataset
(b) Image from COVID dataset

Both images **are enhanced**, the red dots show the ground truth spot position. Ideally, the peaks of the histograms should match the red spots.

Method

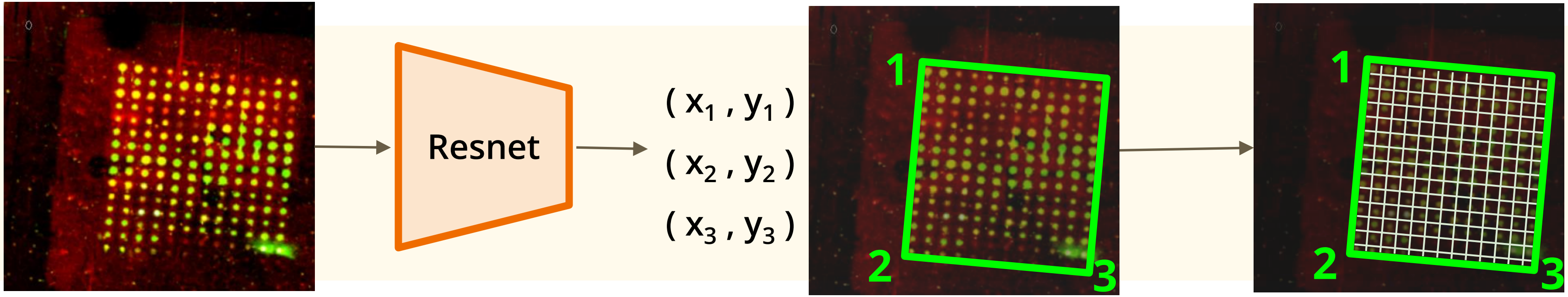
Input

- We want to predict each block individually, so we pre-defined a **window** for each block on a chip. The position and size of the windows are fixed among all chips in a dataset.
- The position of the windows are known information once the microarray is designed, no labeling is needed.
- Windows may overlap, or contain a portion of neighboring blocks, but guaranteed to contain a whole block.
- Microarray images usually contains 2 channels, scanned by different wavelengths of light. The spot coordinates of the 2 channels should be identical.



Proposed Method

Input image



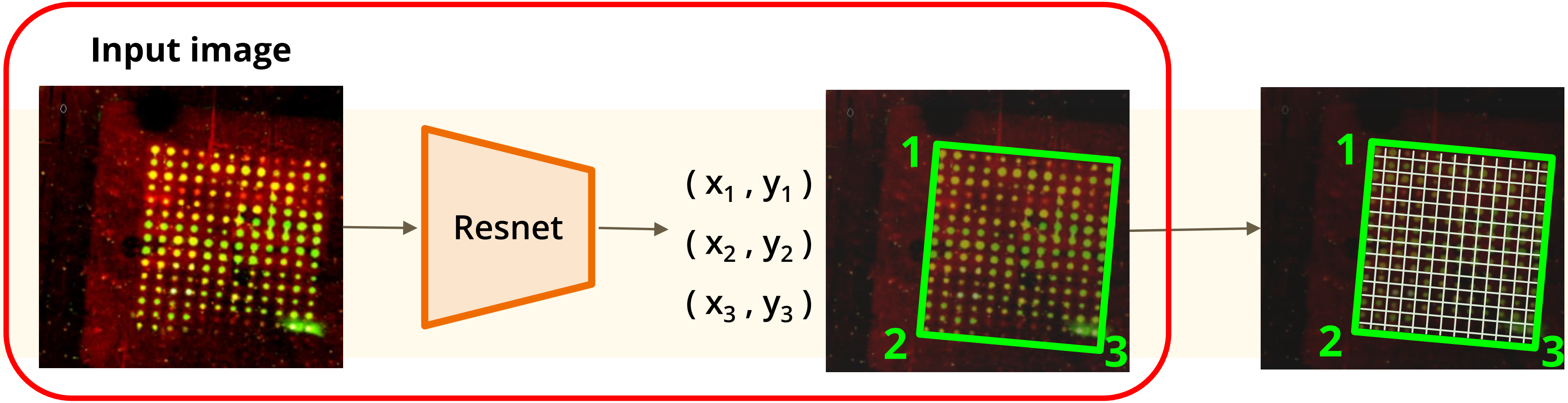
Resnet

(x_1 , y_1)

(x_2 , y_2)

(x_3 , y_3)

Proposed Method: Block Localization



Input image

Resnet

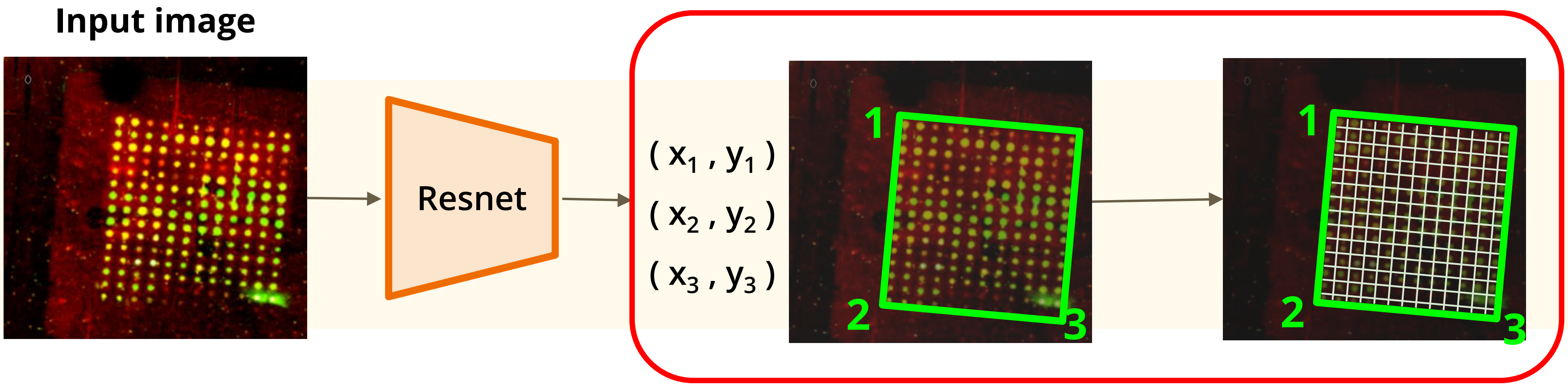
(x_1, y_1)

(x_2, y_2)

(x_3, y_3)

- **Block localization** is done by a convolutional neural network, which outputs 3 pairs of x, y coordinates representing the positions of the 3 spots at the corners of a block, forming a bounding box.

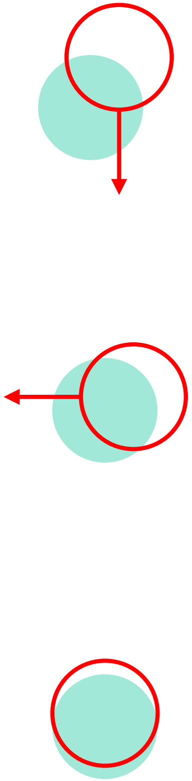
Proposed Method: Spot Localization



- **Spot localization** is done by interpolating the corner coordinate values according to the rows / columns of spots in a block.

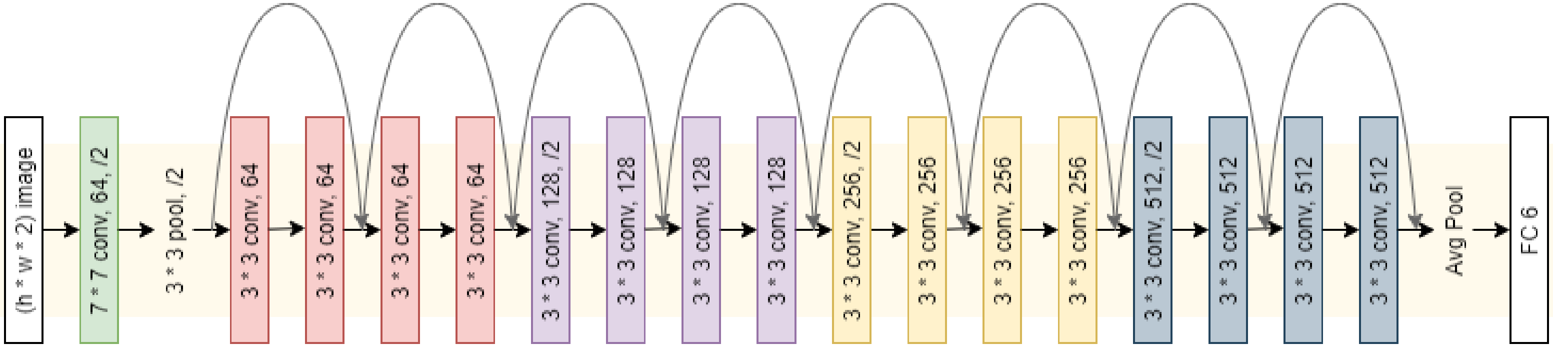
Position Finetuning

- Pre-define a circle with the interpolated spot coordinates as center. Move the circle in 4 directions and choose the one with the highest mean brightness within the circle.
- Terminates if none of the 4 directions' mean brightness is higher than the current one.



Our Model

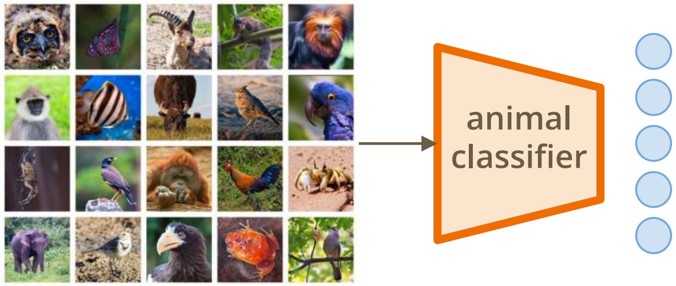
- Inspired by **Region Proposal Network (RPN)** from Faster RCNN using **ResNet** [4] as backbone.
- **Input:** (height × width × channel) microarray image with single block
- **Output:** 1D vector [x_1, y_1, x_2, y_2, x_3, y_3], representing the top left, bottom left, and bottom right spots of a block



Model Pretraining

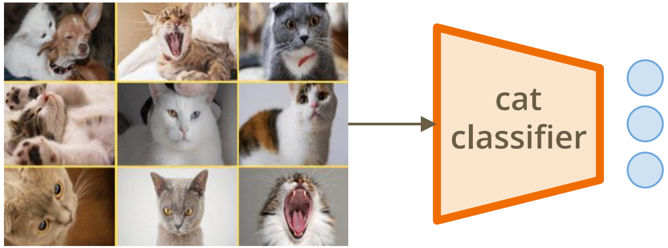
- A type of transfer learning. Transfer knowledge to a different task.
- Let the model learn some “prior knowledge” by **training on similar tasks** (pretrain), in order to more quickly adapt to the target task (fine-tune).
- Pretraining is especially helpful when there is little training data for your target task.

Pretrain



animal classifier

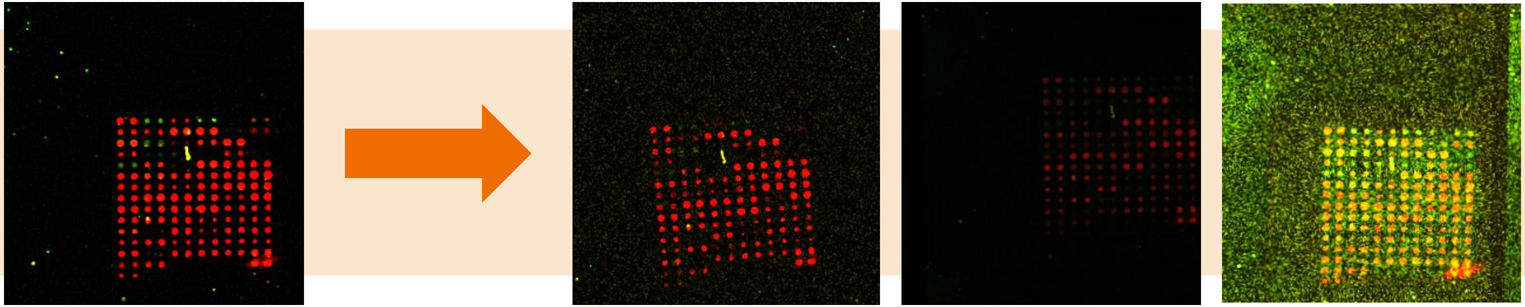
Fine-tune



cat classifier

Data Augmentation

- Get new data for free by modifying the original data, making the model more robust. Has a big impact if there is little original data.
- In this work, we adjusted the offset, rotation, blur, brightness and contrast of the input image in random order. The coordinates are shifted as well.



Original image

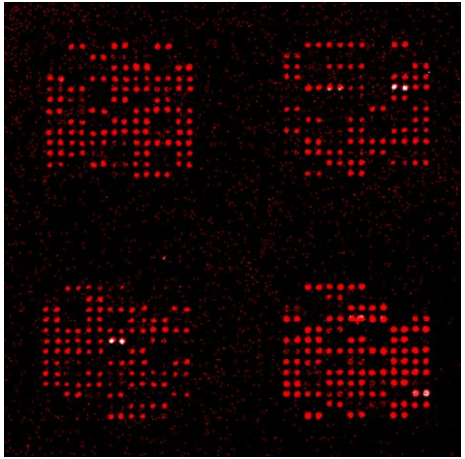
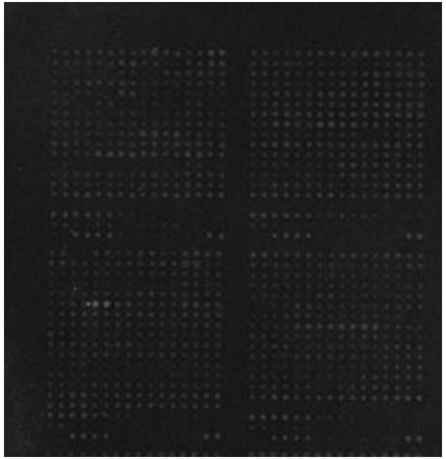
Augmented images

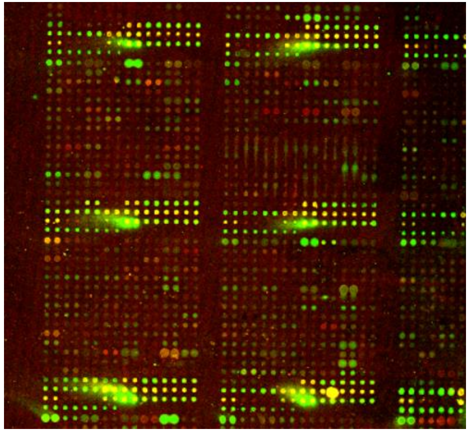
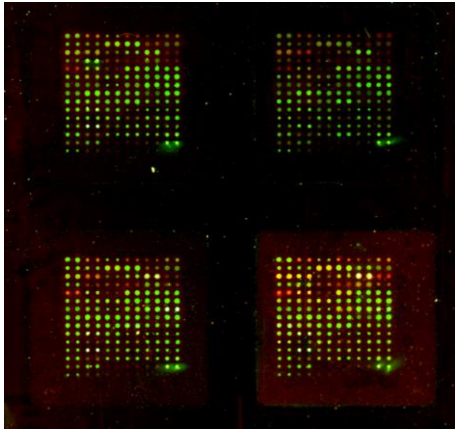
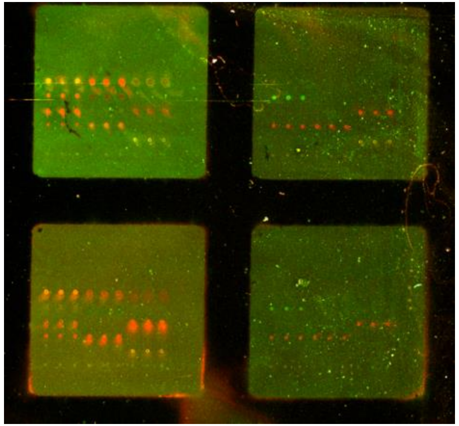
Datasets

Datasets

⅔ chips as train+val

⅓ chips as test

Dataset	P. Salmonis **(PS)**	Bipolar Disorder **(BD)**
type	gene chip (mRNA)	protein chip
chips	4	2
blocks per chip	48	48
spots per block	13×14	19×16
sample (after equalization)		

Dataset	Kawasaki Disease **(KD)**	**KD focus**	covid-19 **(COVID)**
type	protein chip	protein chip	protein chip
chips	80	18	7
blocks per chip	48	14	14
spots per block	19×16	15×12	6×9
sample (after equalization)			

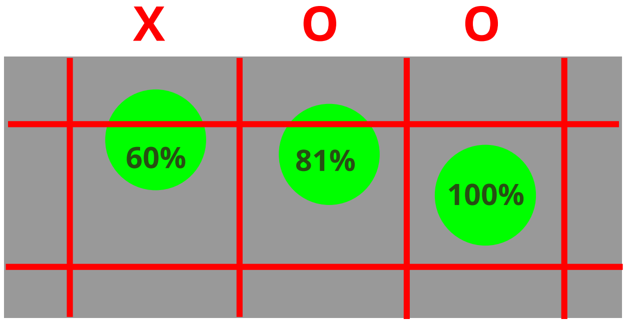
Experiments

Evaluation Metric

Joseph used lines to seperate row / columns of spots, and they use accuracy (acc) as their metric.

*"If more than **80%** of spot pixels reside in the equivalent compartment, the spot is correctly gridded."*

$$acc = \frac{number\ of\ correct\ spots}{number\ of\ all\ spots}$$



Our work predicts the center coordinates of spots instead of seperating lines, but we transformed this metric into an approximately equivalent metric that is compatible to our gridding. (appendix)

Comparison to Previous Works

	KD (2150)	KD focus (134)	PS (76)	COVID (44)	BD (38)
GenePix auto	99.0	**100**	97.8	49.7	**99.5**
Joseph (paper)	-	-	97.9	-	-
Joseph (reproduce)	49.3	72.2	94.8	20.6	80.0
Ours (IN+MA, ×30)	**99.1**	**100**	**100**	**99.3**	97.6

Training settings:

2000 epochs
test using the epoch with the lowest validation loss
batch size 64
Adam optimizer
triangular cyclic learning rate [0.01, 0.001]

Comparison to Previous Works

	KD (2150)	KD focus (134)	PS (76)	COVID (44)	BD (38)
GenePix auto	99.0	**100**	97.8	49.7	**99.5**
Joseph (paper)	-	-	97.9	-	-
Joseph (reproduce)	49.3	72.2	94.8	20.6	80.0
Ours (IN+MA, ×30)	**99.1**	**100**	**100**	**99.3**	97.6

Training settings:

2000 epochs
test using the epoch with the lowest validation loss
batch size 64
Adam optimizer
triangular cyclic learning rate [0.01, 0.001]

Comparison to Previous Works

	KD (2150)	KD focus (134)	PS (76)	COVID (44)	BD (38)
GenePix auto	99.0	**100**	97.8	49.7	**99.5**
Joseph (paper)	-	-	97.9	-	-
Joseph (reproduce)	49.3	72.2	94.8	20.6	80.0
Ours (IN+MA, ×30)	**99.1**	**100**	**100**	**99.3**	97.6

Training settings:

2000 epochs
test using the epoch with the lowest validation loss
batch size 64
Adam optimizer
triangular cyclic learning rate [0.01, 0.001]

Effect of Real Data Amount

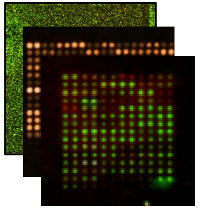
Tested using IN+MA pretrained, aug×30 model

real samples	KD (2150)	KD focus (134)	PS (76)	COVID (44)	BD (38)
2150	**99.2**	-	-	-	-
134	99.2	**100**	-	-	-
100	99.1	**100**	-	-	-
76	97.9	98.8	**100**	-	-
44	95.0	97.7	94.9	**99.3**	-
38	-	-	-	**-**	**97.6**
20	59.6	85.0	74.8	64.5	88.0

Effect of Pretraining: Datasets

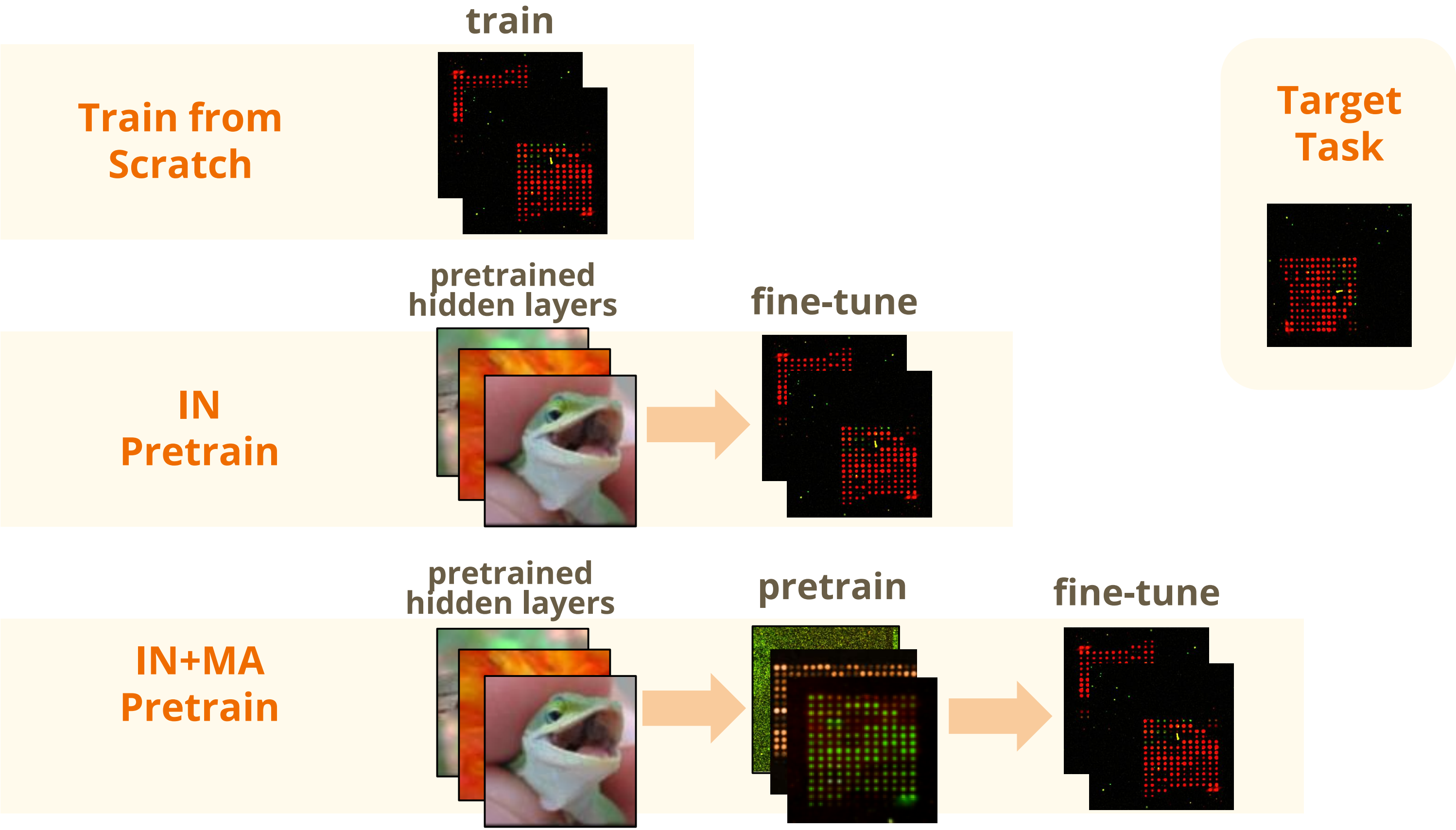
	ImageNet (IN)	Microarray (MA)
amount	14M	~5000
description	Natural image classification task of 1000 classes	The other 4 microarray datasets' training data. Uses 5 times augmentation.

large but not similar



similar but not large

Effect of Pretraining: Settings



train

Train from Scratch

pretrained hidden layers

fine-tune

IN Pretrain

pretrained hidden layers

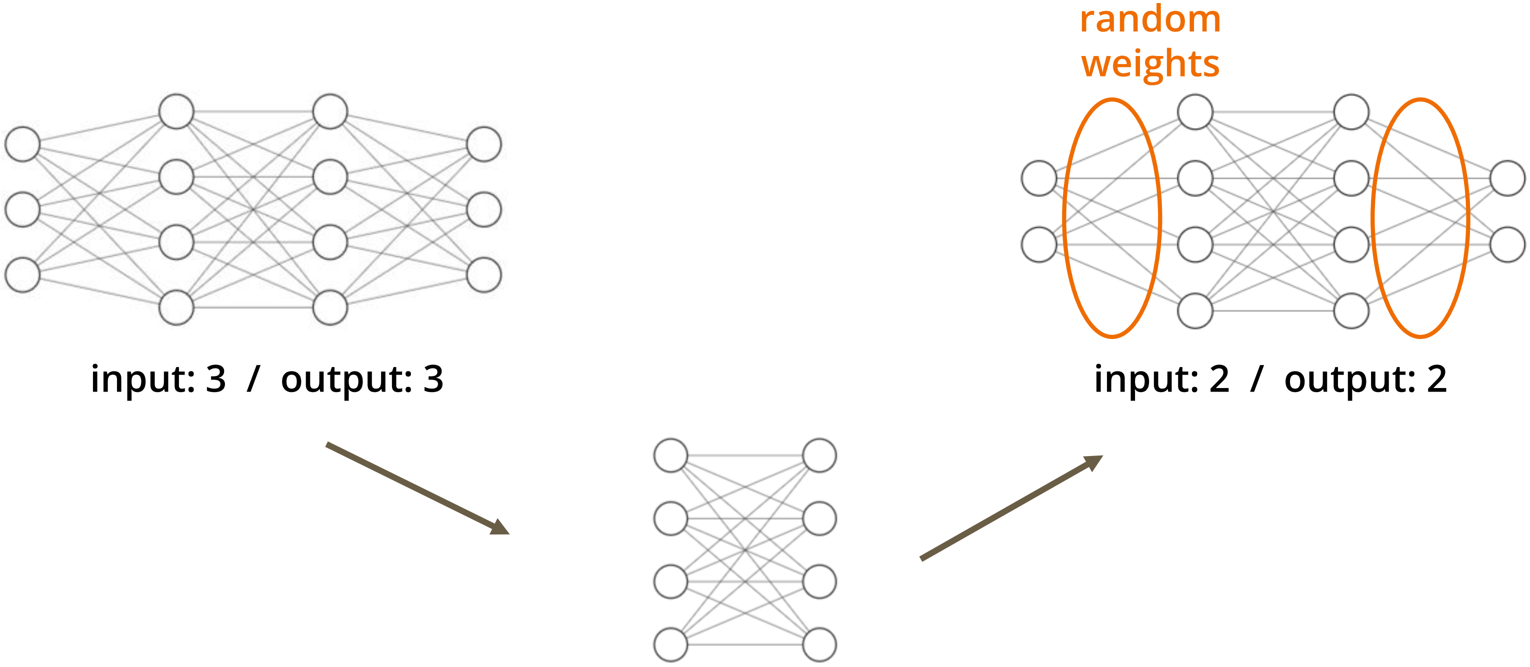
pretrain

fine-tune

IN+MA Pretrain

Target Task

Effect of Pretraining: Settings



random weights

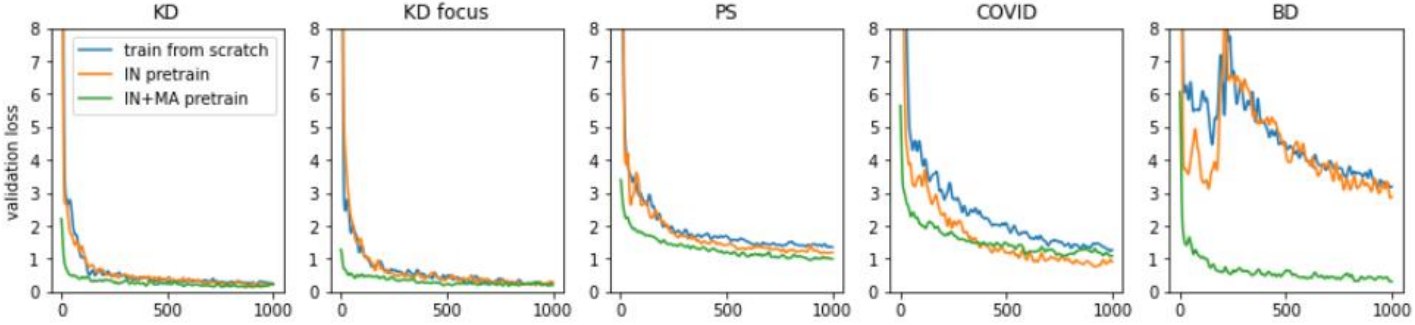
input: 3 / output: 3

input: 2 / output: 2

Effect of Pretraining

Tested using aug×30 model

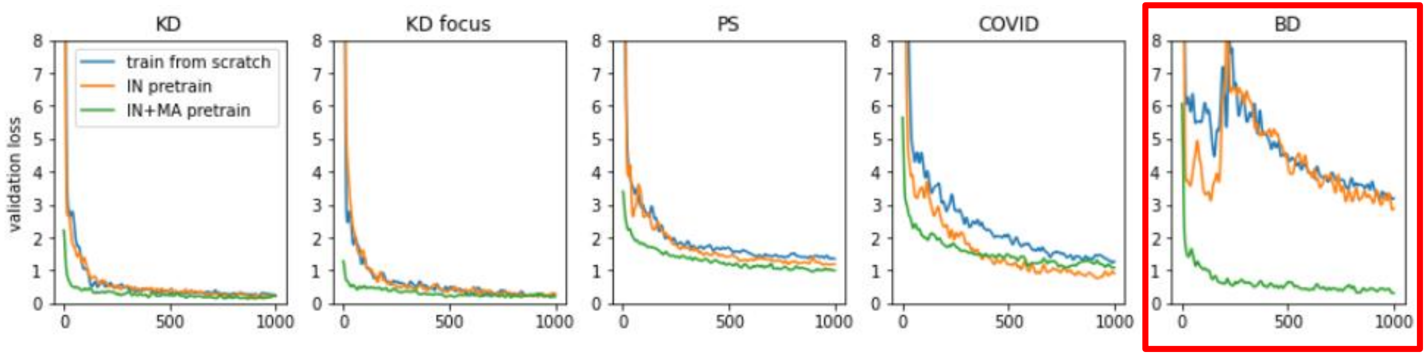
dataset (#blocks)	KD (100)	KD focus (100)	PS (76)	COVID (44)	BD (38)
train from scratch	95.2	90.8	93.1	95.0	44.3
IN pretrain	**99.1**	99.1	96.9	97.2	38.2
IN+MA pretrain	**99.1**	**100**	**100**	**99.3**	**97.6**



Effect of Pretraining

Tested using aug×30 model

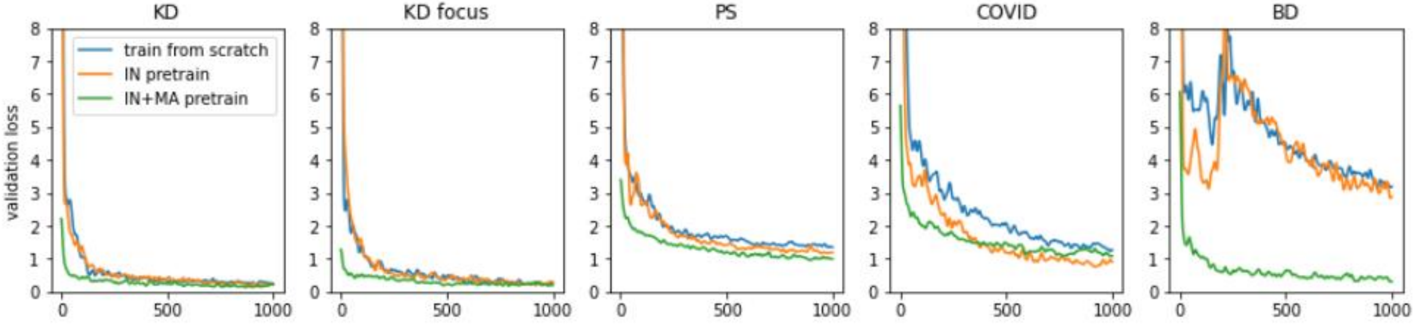
dataset (#blocks)	**KD (100)**	**KD focus (100)**	**PS (76)**	**COVID (44)**	**BD (38)**
train from scratch	95.2	90.8	93.1	95.0	44.3
IN pretrain	**99.1**	99.1	96.9	97.2	38.2
IN+MA pretrain	**99.1**	**100**	**100**	**99.3**	**97.6**



Effect of Pretraining

Tested using aug×30 model

dataset (#blocks)	KD (100)	KD focus (100)	PS (76)	COVID (44)	BD (38)
train from scratch	95.2	90.8	93.1	95.0	44.3
IN pretrain	**99.1**	99.1	96.9	97.2	38.2
IN+MA pretrain	**99.1**	**100**	**100**	**99.3**	**97.6**



Effect of Data Augmentation

		KD (100)	KD focus (100)	PS (76)	COVID (44)	BD (38)
IN	**aug 0**	87.5	95.1	93.8	56.4	42.9
	aug ×20	96.8	96.0	96.0	97.1	68.7
	aug ×30	**99.1**	**99.8**	**96.9**	**97.2**	**38.2**
IN + MA	**aug 0**	85.2	73.5	85.5	63.7	92.6
	aug ×20	98.7	99.0	98.9	90.9	96.6
	aug ×30	**99.1**	**100**	**100**	**99.3**	**97.6**

Effect of Data Augmentation

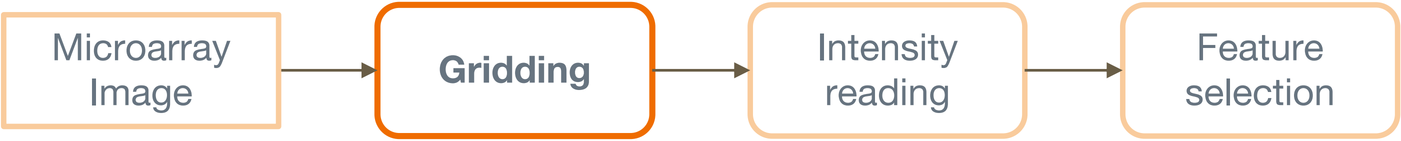
		KD (100)	KD focus (100)	PS (76)	COVID (44)	BD (38)
IN	**aug 0**	87.5	95.1	93.8	56.4	42.9
	aug ×20	96.8	96.0	96.0	97.1	38.7
	aug ×30	**99.1**	**99.8**	**96.9**	**97.2**	**68.2**
IN + MA	**aug 0**	85.2	73.5	85.5	63.7	92.6
	aug ×20	98.7	99.0	98.9	90.9	96.6
	aug ×30	**99.1**	**100**	**100**	**99.3**	**97.6**

Effect of Data Augmentation

		KD (100)	KD focus (100)	PS (76)	COVID (44)	BD (38)
IN	**aug 0**	87.5	95.1	93.8	56.4	42.9
	aug ×20	96.8	96.0	96.0	97.1	38.7
	aug ×30	**99.1**	**99.8**	**96.9**	**97.2**	**68.2**
IN + MA	**aug 0**	85.2	73.5	85.5	63.7	92.6
	aug ×20	98.7	99.0	98.9	90.9	96.6
	aug ×30	**99.1**	**100**	**100**	**99.3**	**97.6**

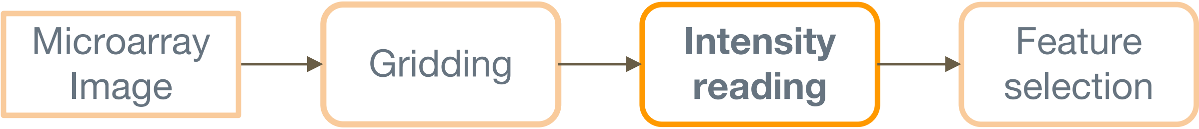
Aggregate into an Existing Pipeline

To prove the practicality of or gridding method, we replace the gridding step of an existing microarray analysis pipeline to our method. **The final objective is to select features for a binary classification problem** on diagnosing covid-19.



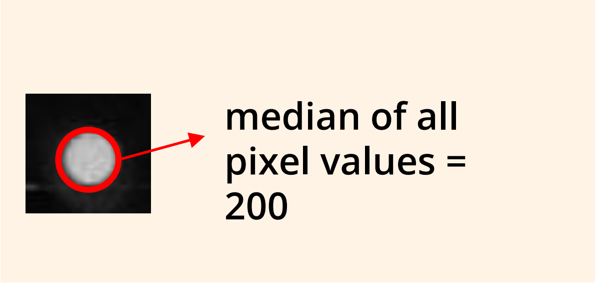
Originally the gridding step is done by GenePix manual gridding. This is the correrct answer.

Intensity Reading



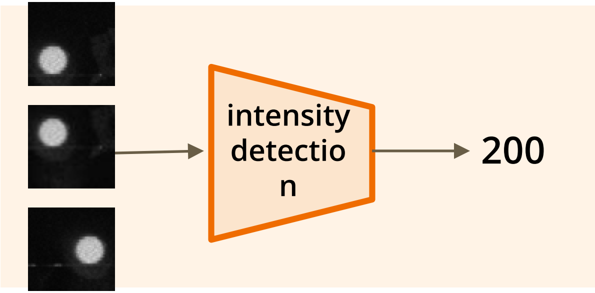
GenePix:

GenePix has built-in intensity reading tools, but require **percise coordinate** and **radius** for each spot.



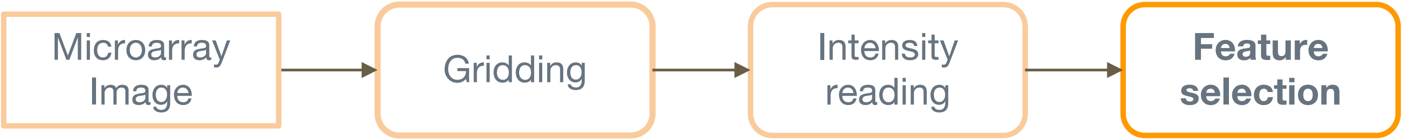
Intensity Model:

Require just rough spot coordinates. Given an image containing a spot, we use an intensity detection model[4] to predict the spot's intensity.



Using our gridding, the MAPE of the output intensity is **4.89%**.

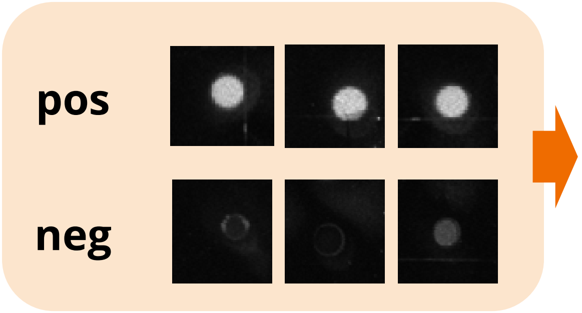
Feature Selection



Microarray Image → Gridding → Intensity reading → **Feature selection**

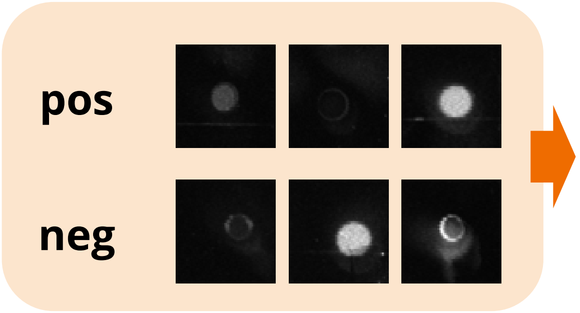
We wish to select 1 or 2 features out of 18 features that best classifies positive and negative samples. We run every combinations using logistic regression and select the features with top 3 training auc.

biomarker A



auc
0.98

biomarker B



auc
0.66

Aggregate into Existing Pipeline

Manual gridding (ground truth)

	rank	biomarker	test auc
1 feature	1	COVID19-S1+S2.550	98.7
	2	COVID19-N.550	93.9
	3	SARS-N.550	93.3
2 features	1	COVID19-S1+S2.550 H3N2-HA.650	97.4
	2	COVID19-S1+S2.550 SARS-S1.650	97.2
	3	COVID19-S1+S2.550 COVID19-S1.650	98.2

Our gridding

	rank	biomarker	test auc
1 feature	1	COVID19-S1+S2.550	98.7
	2	SARS-N.550	93.2
	3	COVID19-N.550	93.5
2 features	1	COVID19-S1+S2.550 H3N2-HA.650	97.8
	2	COVID19-S1+S2.550 COVID19-S1.650	98.5
	3	COVID19-S1+S2.550 COVID19-N.650	98.2

Conclusion

Conclusion

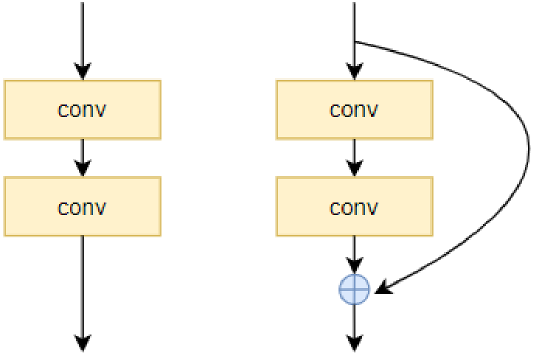
- The first to propose a deep learning method on microarray gridding, offers an option that can decrease the cost of labeling.
- Outperforms traditional methods especially on irregular datasets if sufficient data is provided.
- Discover that pretraining on both ImageNet or microarray datasets increases accuracy.

Thank you for listening

Appendix

ResNet

- From *Deep Residual Learning for Image Recognition, 2015*
- As neural networks become deeper, degration problem emerges and stop the model from performing better.
- ResNet adds shortcuts across multiple layers, called residual connections. The forwarded layer is added to the output of the skipped layers.
- This simple architecture is broadly used in various image-related tasks.



without residual connection

with residual connection

Data Augmentation Settings

- Guassian blur (20% chance)
- Contrast adjust 0.8 ~ 1.2
- Brightness adjust (multiply) 0.8 ~ 1.0
- Offset -0.2 ~ 0.2
- Rotate -5º ~ 5º

Above operations are executed in random order

If any of the coordinates of an augmented sample are out of bounds, delete the sample and generate a new one

Our Output Format Compared to Faster RCNN

We compare our output format to faster RCNN, a famous object detection model

Faster RCNN	Ours	Resultant Changes
$\mathrm{H} \times \mathrm{W} \times \mathrm{A} \times 4 \; (x\, y\, h\, w)$	$1 \times 1 \times 6 \; (x_1\, y_1\, x_2\, y_2\, x_3\, y_3)$	-
Multiple bounding boxes per image	Only one bounding box per image	Only outputs 1 set of coordinate ($\mathrm{H}{\times}\mathrm{W} \rightarrow 1{\times}1$)
Uses anchor boxes	No need for anchor boxes	No presence of A
Bounding box cannot rotate	Bounding box can rotate	3 points defines a rotated box ($x\, y\, h\, w \rightarrow x_1\, y_1\, x_2\, y_2\, x_3\, y_3$)

Anchor boxes are frequently use in object detection models to better handle variable **h×w** ratios, but in this task, the shape of bounding boxes are usually fixed

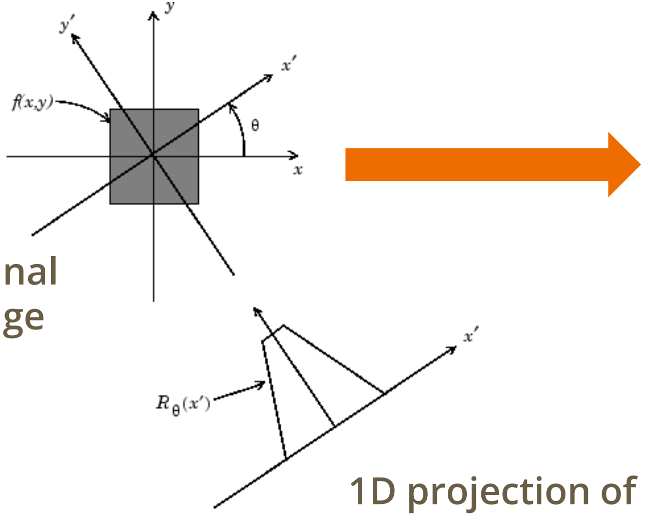
Radon Transform

$$R_\theta(x') = \int_{-\infty}^{\infty} f(x' \cos\theta - y' \sin\theta, x' \sin\theta + y' \cos\theta) dy'$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

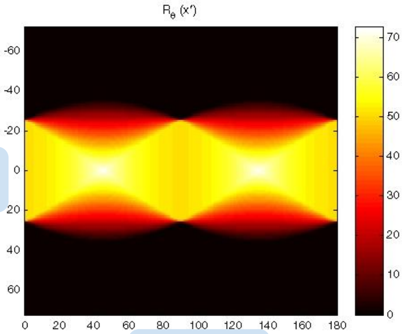
$R_\theta(x')$ is the 1D projection of an 2D image $f(x, y)$ on the rotated x axis x'. For every angle θ we can get a new projection, forming a transformed 2D image. Radon transform is usually used in finding the direction of line patterns in the image.



original image

1D projection of θ

$R_\theta(x')$



θ

transformed image

[source]

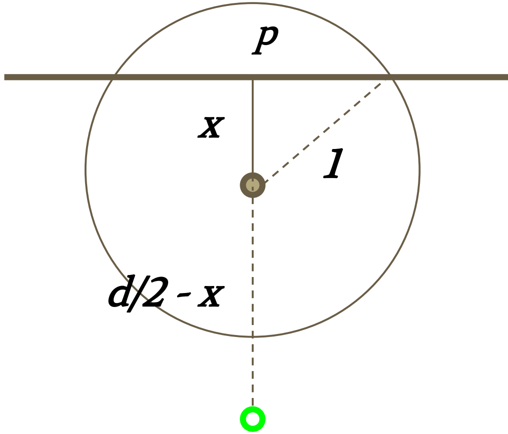
Gridding Correctness

Assume spot radius = 1, and ***d*** = the average spacing between spot centers, ***d/2*** = the distance of correct spot center to a seperation line

Consider the boundary condition shown on the right, let ***p*** be the proportion of top area exceeding the line to the whole spot's area, then we can get the equation:

$$p = \frac{1}{\pi}\cos^{-1}(x) - \frac{x}{\pi}\sqrt{1-x^2}$$

where ***x*** is the distance of the center of spot to the seperation line. When ***p*** = 0.2, ***x*** ~= 0.49, therefore if the distance of a predicted spot to the correct spot is more than ***d/2 - 0.49*** , meaning ***p*** > 0.2, then this prediction is incorrect.



Position Finetuning

accuracy (%)

	KD	KD focus	PS	COVID	BD
without position fintuning	99.1	**100**	**100**	**99.3**	97.6
with position fintuning	**99.2**	99.9	99.7	98.8	**99.8**

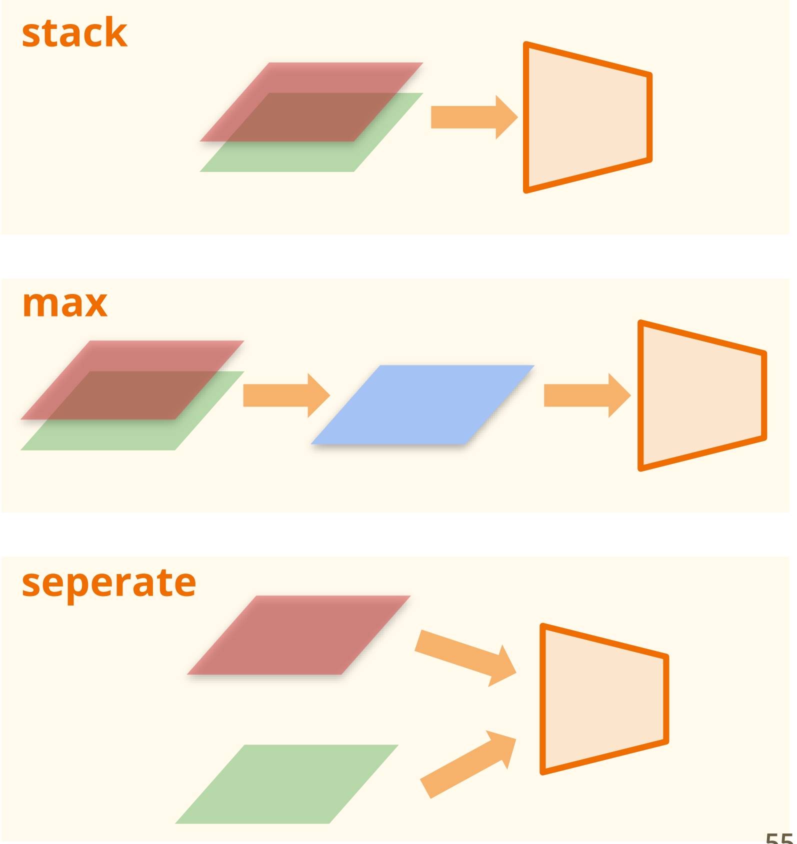
time (sec)

	KD	KD focus	PS	COVID	BD
without position fintuning	**103.0**	**14.7**	**7.7**	**7.6**	**4.1**
with position fintuning	24898.9	3552.9	915.5	473.7	1693.5

Channel Mode

mode	covid
stack	**97.1**
max	81.1
seperate	28.9

The ‘separate’ mode splits one sample into two, thus doubles the amount of data. But experiments show that inputting both channels is necessary.



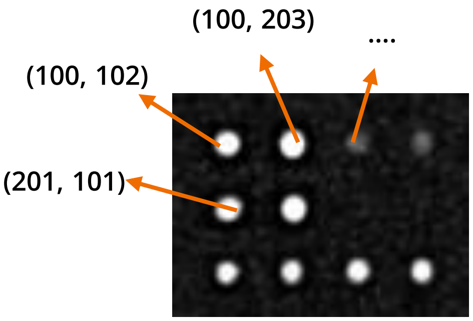
Types of Gridding

There are 2 output formats:

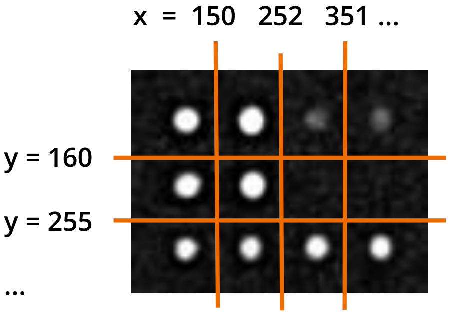
1. **Finding the center coordinates of all spots**
- GenePix, this work

1. **Finding the seperating lines for rows / columns**
- Joseph, most previous works

- It is possible to additionally label the radius of each spot using GenePix, but in this work, we consider only the center of circle.
- To evaluate accuracy, we convert our result to type 2



Type 1



References

[3] A new microarray analyzer based on deep learning, 2020 (**Hsieh**)

[4] Deep Residual Learning for Image Recognition, 2015

[5] Gridline: Automatic Grid Alignment in DNA Microarray Scans, 2004

[6] M^3G: Maximum Margin Microarray Gridding, 2010

[7] Multilevel Segmentation Optimized by Physical Information for Gridding of Microarray Images, 2019