CGAN-Plankton: Towards Large-scale Imbalanced Class Generation and Fine-Grained Classification

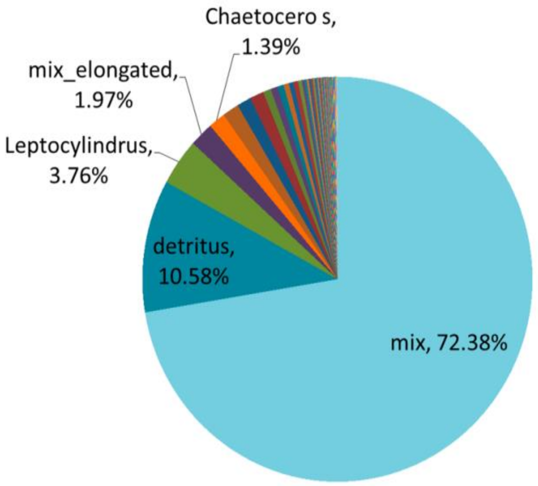
China National Conventional Center

Bejing, China

September 19, 2017

Imbalanced Problem Statement

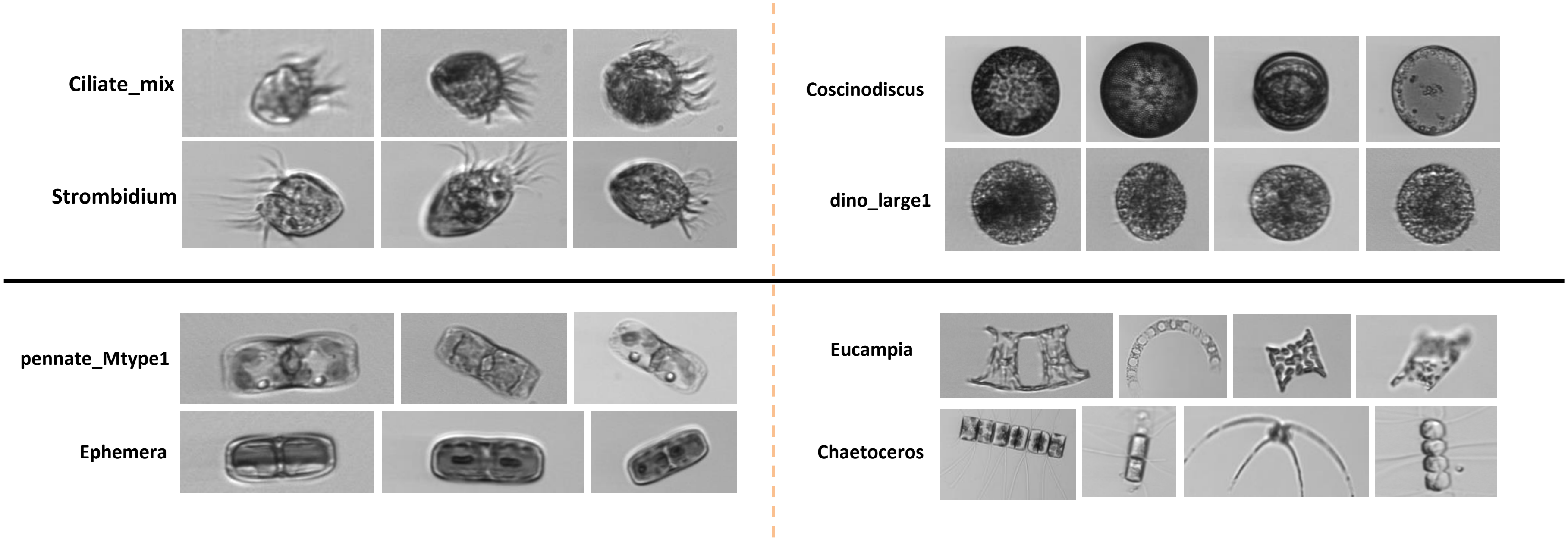
- Data distribution of **WHOI-Plankton**



Class	Total	Training	Testing
Mix	**73.15%**	**72.38%**	**80.69%**
Detritus	10.62%	10.58%	11.02%
Letocylindrus	3.54%	3.75%	1.28%
Mix_elongated	1.86%	2.05%	1.06%
Dino30	1.27%	1.43%	1.17%
Sum	90.60%	90.19%	95.22%

Challenge: Class imbalance

Similarity between class and diversity within class



Challenge: fine-grained

What is a better solution on this problem?

- Average accuracy
- Precision and recall
- F1 score
- Confusion matrix

$$F1-score = 2\frac{precision * recall}{precision + recall}$$

Don’t be fooled by the weighted accuracy!

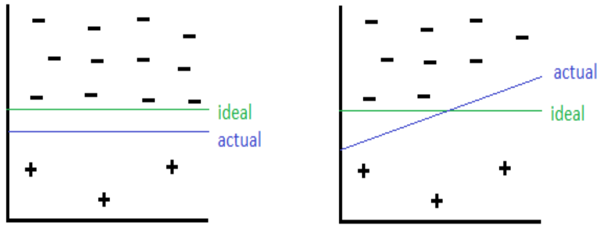
Two ways to solve the problem

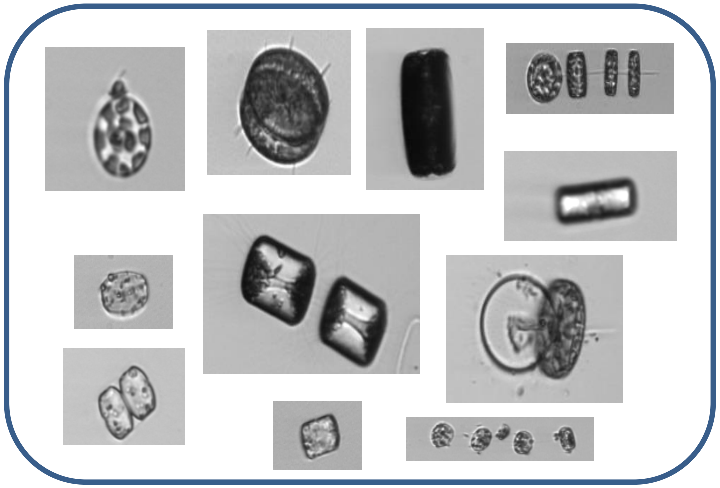
- Common goal: shrink the imbalance

- Approach1: data re-sampling
  - Under sampling
  - Over sampling
  - Mix of over sampling and under sampling
- Approach2: cost-sensitive learning
  - Impose heavy penalty on majority class

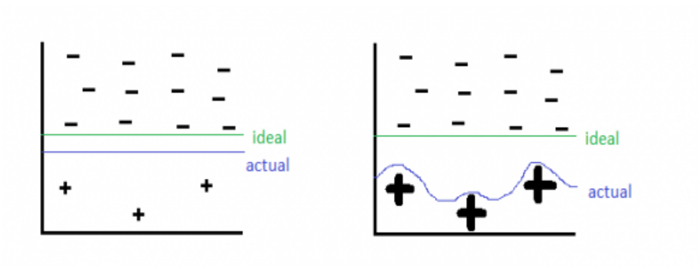
Sampling based approach

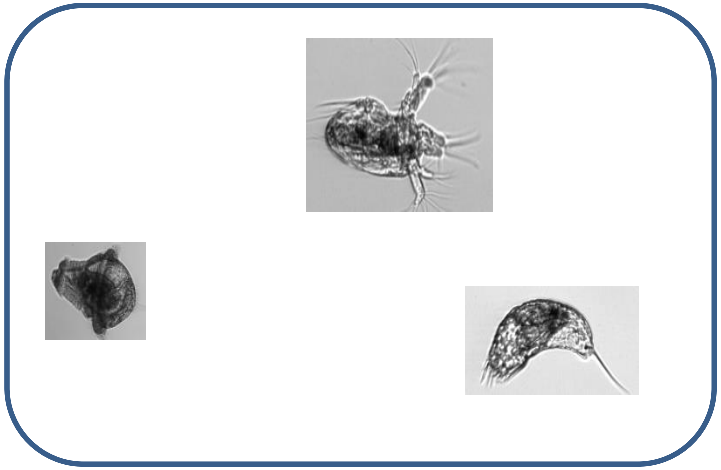
- Under sampling





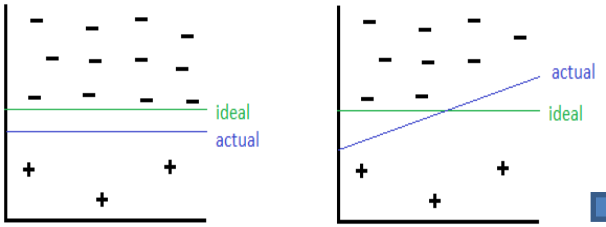
- Over sampling



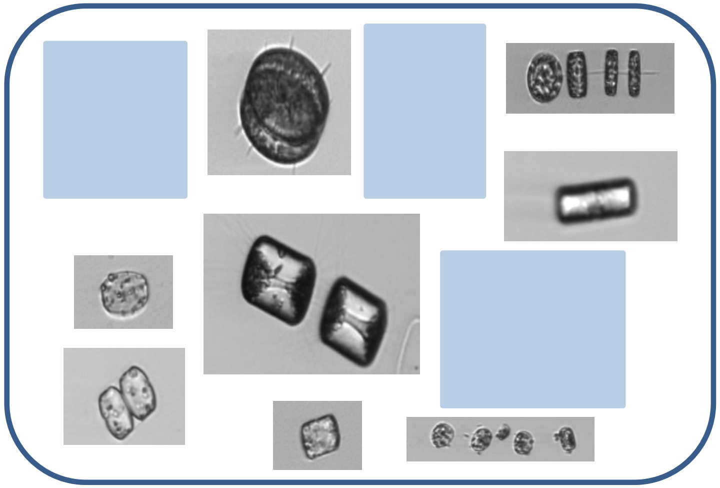


Sampling based approach

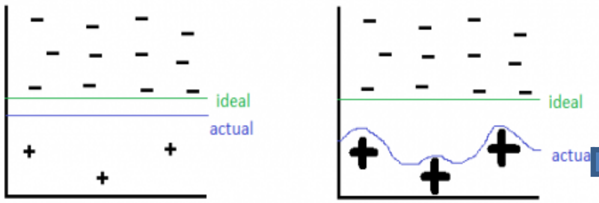
- Under sampling



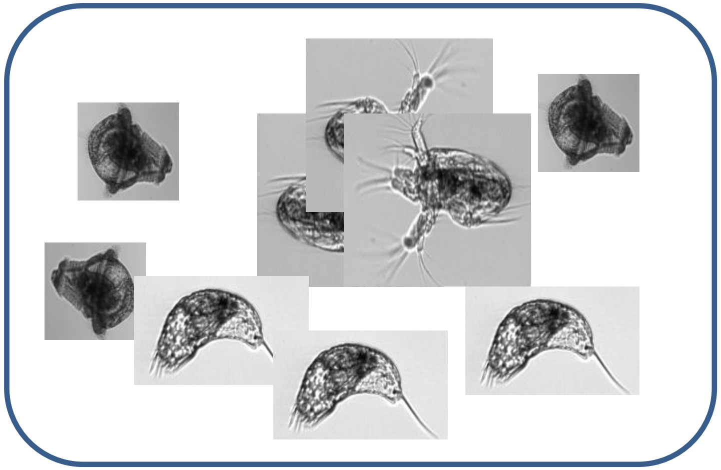
Missing feature



- Over sampling

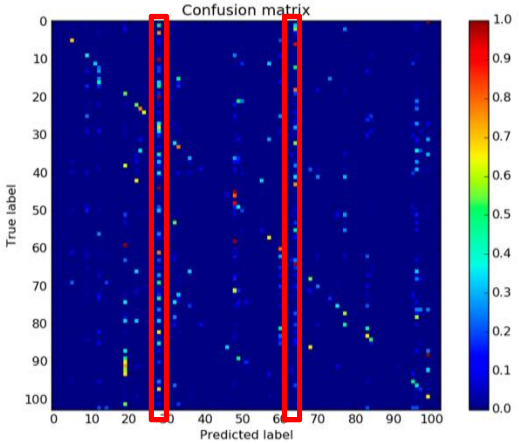


Overfitting

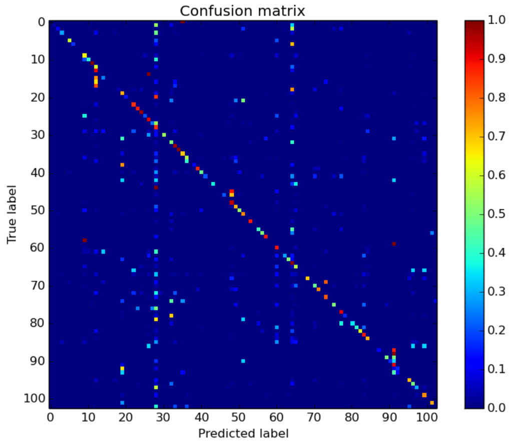


Benchmark

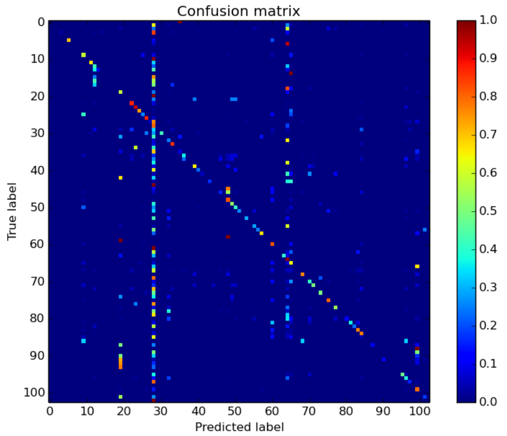
Database	Model	Weighted accuracy	F1 score
WHOI-Plankton	CIFAR10 CNN	0.9297	0.1975
WHOI-Plankton	AlexNet	0.9395	0.3837
WHOI-Plankton	VGG16	0.9505	0.4302



CIFAR10 CNN

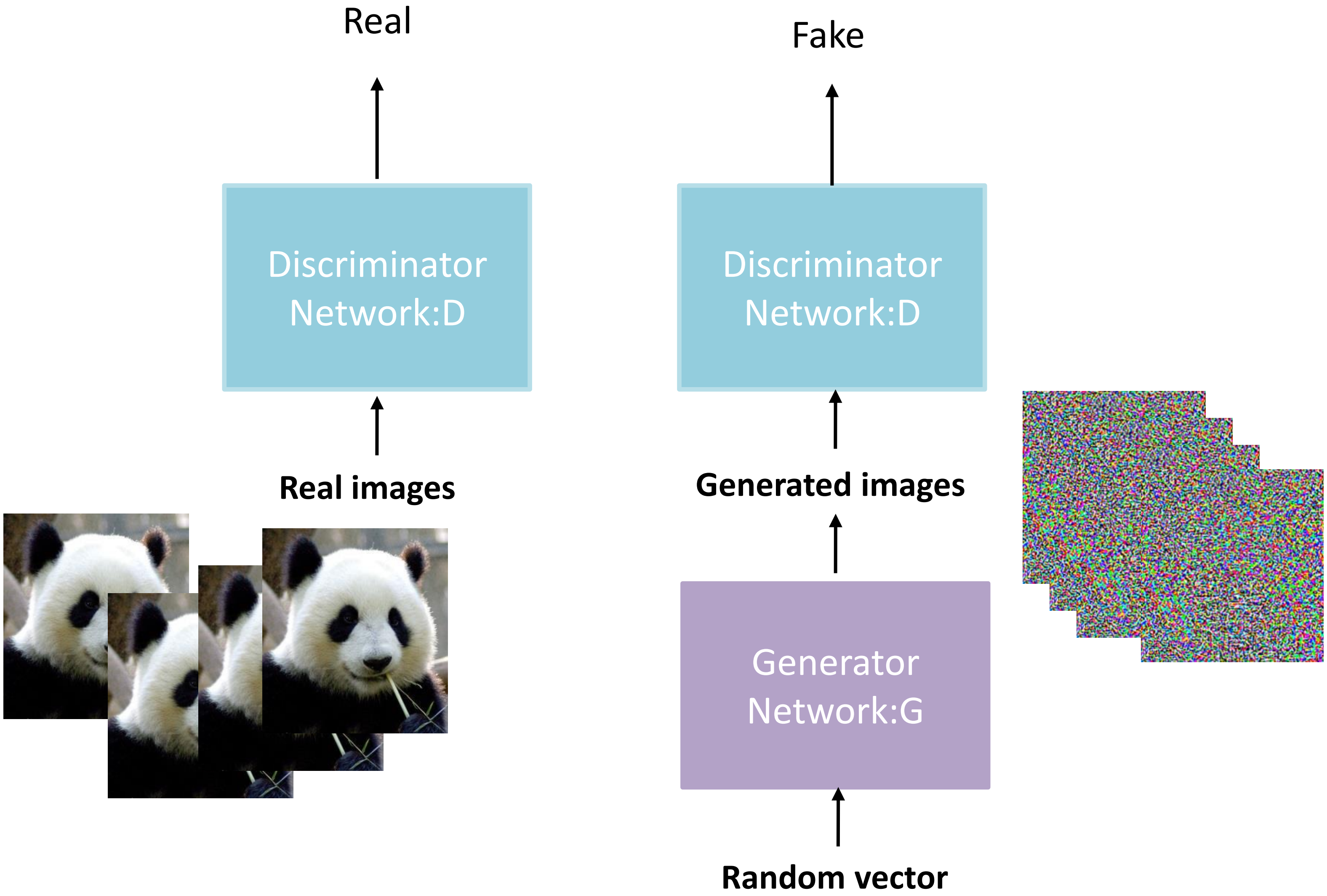


AlexNet



VGG16

Generative Adversarial Networks



Real

Discriminator Network:D

Real images

Fake

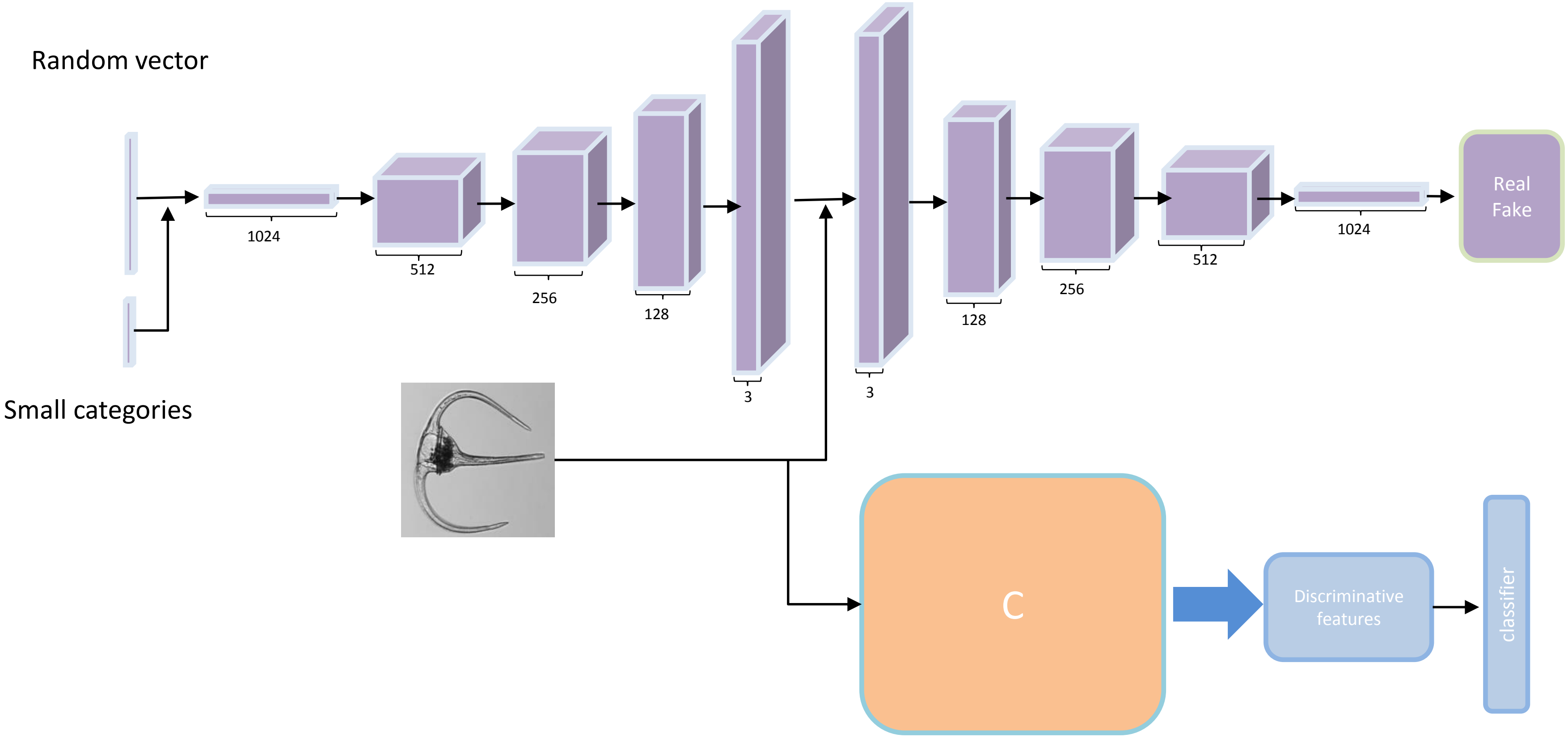
Discriminator Network:D

Generated images

Generator Network:G

Random vector

CGAN-Plankton model



Random vector

Small categories

Real
Fake

C

Discriminative features

classifier

$$y = soft\max(CNN(x=\{full\};\omega^{*}),CNN(x=\{\min ority\};\omega_{0}))$$

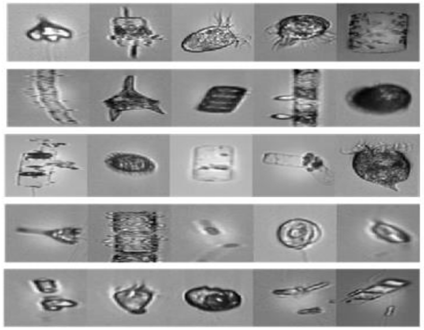
$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_{z}(z)}[\log(1-D(G(z))]$$

Experiments results on WHOI-plank

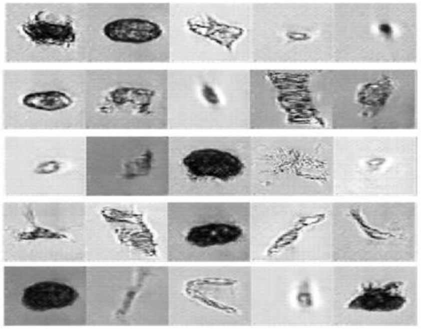
Given a large imbalanced dataset with more than 100 classes

How to generate from the samples with diversity?

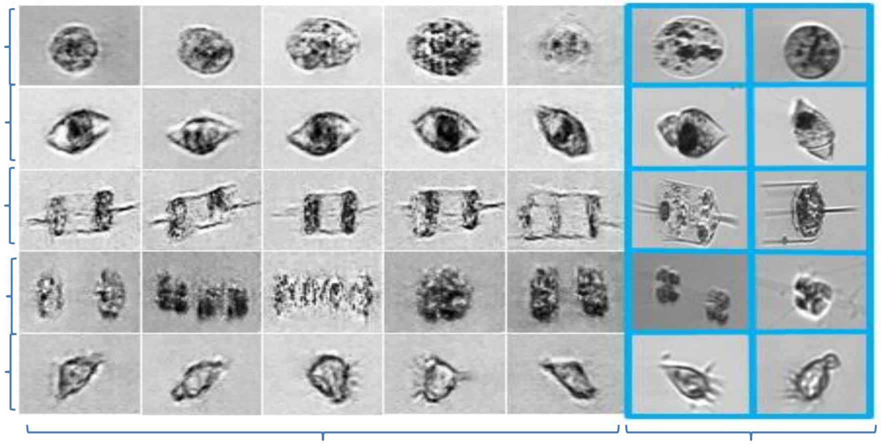
- **Feature transfer form large classes**
- **Conditional generation**
- **Auxiliary classifier**



Real samples from **WHOI**



Generated samples from model

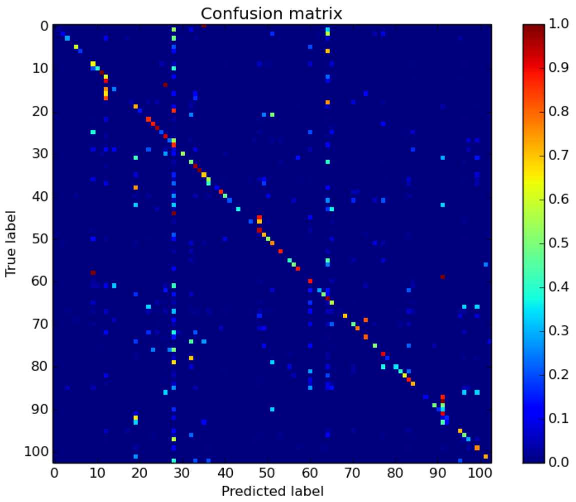


Generated samples with categories

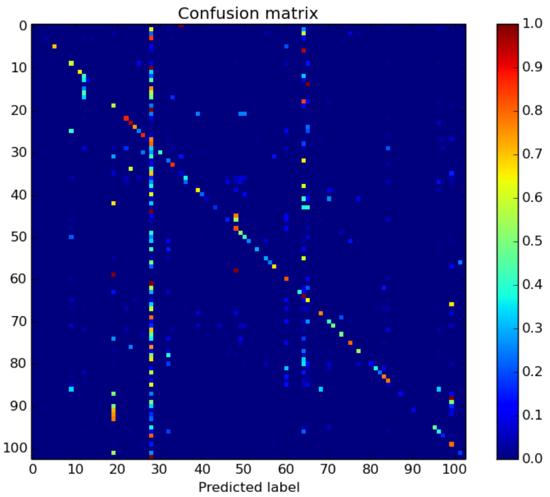
Experiments results

Database	Model	Weighted accuracy	F1 score
WHOI-Plankton	CIFAR10 CNN	0.9297	0.1975
WHOI-Plankton	AlexNet	0.9395	0.3837
WHOI-Plankton	VGG16	0.9475	0.4461
WHOI-Plankton + sampling	Transfer learning	0.9280	0.3339
WHOI-Plankton	CGAN-plankton	0.9425	0.4777
WHOI-Plankton +generated samples	CGAN-plankton	**0.9443**	**0.4992**

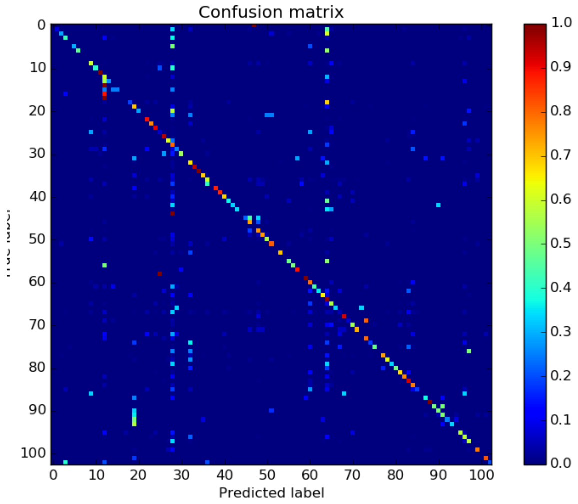
Visualization of confusion matrix



Transfer learning



VGG16



CGAN-plankton

Conclusions & discussion

- **Use GAN to solve the imbalance problem(data driven)**
- **Feature transfer form large classes**
- **Conditional generation and auxiliary classifier**

Database	Model	Accuracy	F1 score
original	CIFAR10 CNN	0.7109	0.6744
generated	CIFAR10 CNN	0.6017	0.4877
generated + original	CIFAR10 CNN	**0.7374**	**0.7259**

Q&A