Computer Vision 1
Bag-of-Words based Image Classification
Final Project Part 1

Gongze Cao, Kai Liang, Xiaoxiao Wen and Zhenyu Gao
Faculty of Science, University of Amsterdam

1 Introduction

In the first part of the final project, we mainly focus on implementing a system for a five-class image classification task based on the Bag-of-Words approach. Firstly, the detailed steps of this pipeline are stated. Then, the experiment results are presented. Finally, we will draw conclusions from the obtained experimental results and propose possible ways to address the problems.

Throughout the assignment, we distribute our work fairly in terms of coding and writing the report.

2 Description

The experiment can be divided to five steps:

- Dataset preparation.
- Building visual vocabulary.
- Building representations.
- Training SVM.
- Evaluation of SVM.

2.1 Dataset preparation

As required by the task, a 5-class image classification is done, namely for airplanes, birds, ships, horses and cars. For training and evaluation, the STL-10 dataset is used which originally contains 10 classes of objects. Thus, firstly, the required classes are extracted both from the training set and the test set with their corresponding labels. Furthermore, as originally the pixels of the images are laid out in column-major order and per R, G and B channels, the images are reshaped into 96x96x3 for convenience.

2.2 Building visual vocabulary

We begin with using SIFT feature extractor to collect the features of the grayscale images using both the keypoint and dense sampling methods. Besides, two other variants of the grayscale-SIFT are also implemented, namely the RGB-SIFT and the opponent-SIFT. For the RGB-SIFT and opponent-SIFT methods with keypoint sampling, the grayscale-SIFT is first executed to obtain the features, after which the corresponding frames are used to extract features from the color channels respectively. Subsequently, the features extracted from the three channels are concatenated to the shape of 128×3 and used as the feature as a whole. In this step, we empirically set the peak threshold of keypoint sampling to 1.5 and the step size of dense sampling to 5, by which we hope to on the one hand, extract more meaningful features, while on the other hand, speed up the entire process.

Secondly, a subset of the training images (i.e. 20%) are randomly selected per class as the vocabulary-building images and their SIFT features are extracted. Then, the K-means clustering algorithm is performed on all the features collected from all the vocabulary building images in order to find a specified number of clusters. The clusters obtained through this procedure are the visual bag-of-words that form the visual vocabulary.

2.3 Building representations

In order to build the representation for the images, features are encoded by the visual vocabulary. For each image, its SIFT features are first obtained. Then, each feature is compared with the visual bag-of-words in the vocabulary and the feature is counted towards its nearest visual word in the feature space in terms of Euclidean distance. At the end, a L2-normalized histogram is obtained, which becomes the representation of the image in the form of a vector with specified length (i.e. number of clusters).

For the test set, all images are encoded as described above, while for the training set, only images which are not used for building the vocabulary are encoded.

2.4 Training SVM

We use the feature histograms extracted in the previous step to train a SVM classifier, which is achieved by the built-in `fitcsvm` function. We also do random shuffling on all pairs of the features and labels. In total, we train five classifiers and each is responsible for separating one class apart from the others so as to solve the multi-class classification. To resolve the class imbalance issue, we set the cost matrix to be: $\begin{bmatrix} 0 & 1 \\ 4 & 0 \end{bmatrix}$. For other parameters we mostly keep the default values, except that we choose the RBF kernel rather than the linear one.

2.5 Evaluation of SVM

After training the SVM, we evaluate all five classifiers on the test set composed by 4000 instances of 5 classes. We evaluate the accuracy of each classifier by regarding the 800 instances with the same label of the classifier as the positive class, and the other instances as the negative class. We then report the average precision of each classifier and average them to get the mean average precision (mAP).

3 Results

The experiment results are presented jointly in Appendix A. For each experimental setting, we specify the following information:

1. The sampling method.
2. The vocabulary size.
3. The type of SIFT descriptor.
4. The training paradigm.
5. The AP for each classifier.
6. The mAP of all five classifiers.

Additionally, we plot the top-5 and bottom-5 images for each classifier based on the confidence score.

4 Analysis

4.1 Keypoint verses Dense Sampling

Based on the results, using dense sampling as features generally behaves significantly better than using solely keypoints under the current settings, which is due to the fact that keypoint sampling does not capture enough information about the image compared with dense sampling.

4.2 The effect of vocabulary size

For keypoint sampling, we observe that the performance does not differ significantly between different vocabulary sizes. However, for dense sampling, the larger vocabulary size, the worse performance we get. This is possibly because we use a fixed step number in the K-means clustering, which in the dense sampling case where we have a lot of instances might not handicap the training process. In fact, we do observe that some indications that the K-means does not converge properly during the experiment of dense sampling. Moreover, this might reflect that 400 is already enough as the vocabulary size and adding size upon it only introduce noises.

4.3 The effect of different types of SIFT descriptor

Surprisingly, we find that the greyscale-SIFT works best in general, while the opponent-SIFT performs the worst. This is probably due to the dimension of the embedding space we are concerned with, which is 128 for greyscale-SIFT, and 384 for RGB-SIFT and opponent-SIFT. It infers that we can already build effective embeddings using only shape and lighting information, and using color information in extra will introduce redundant information and is generally not good for the training of cluster algorithm.

5 Conclusion

In the previous sections, we discuss our system for five-class image classification based on the Bag-of-Words approach. We first use SIFT in different color space with different sampling methods to extract features and descriptors. Then, we build the visual vocabulary by running the K-means algorithm with different cluster sizes. Afterwards, we use the vocabulary to encode the features and represent images by frequencies of visual words. As for classification, we train a SVM classifier for each class of objects. Finally, we evaluate the performance of the system under different conditions both quantitatively and qualitatively. We conclude that systems using greyscale-SIFT with dense sampling generally work the best, and that the size of vocabulary has no prominent influence on the performance.

A Qualitative Evaluation

(from next page)

Gongze Cao, Zhenyu Gao, Kai Liang, Vincent Wen

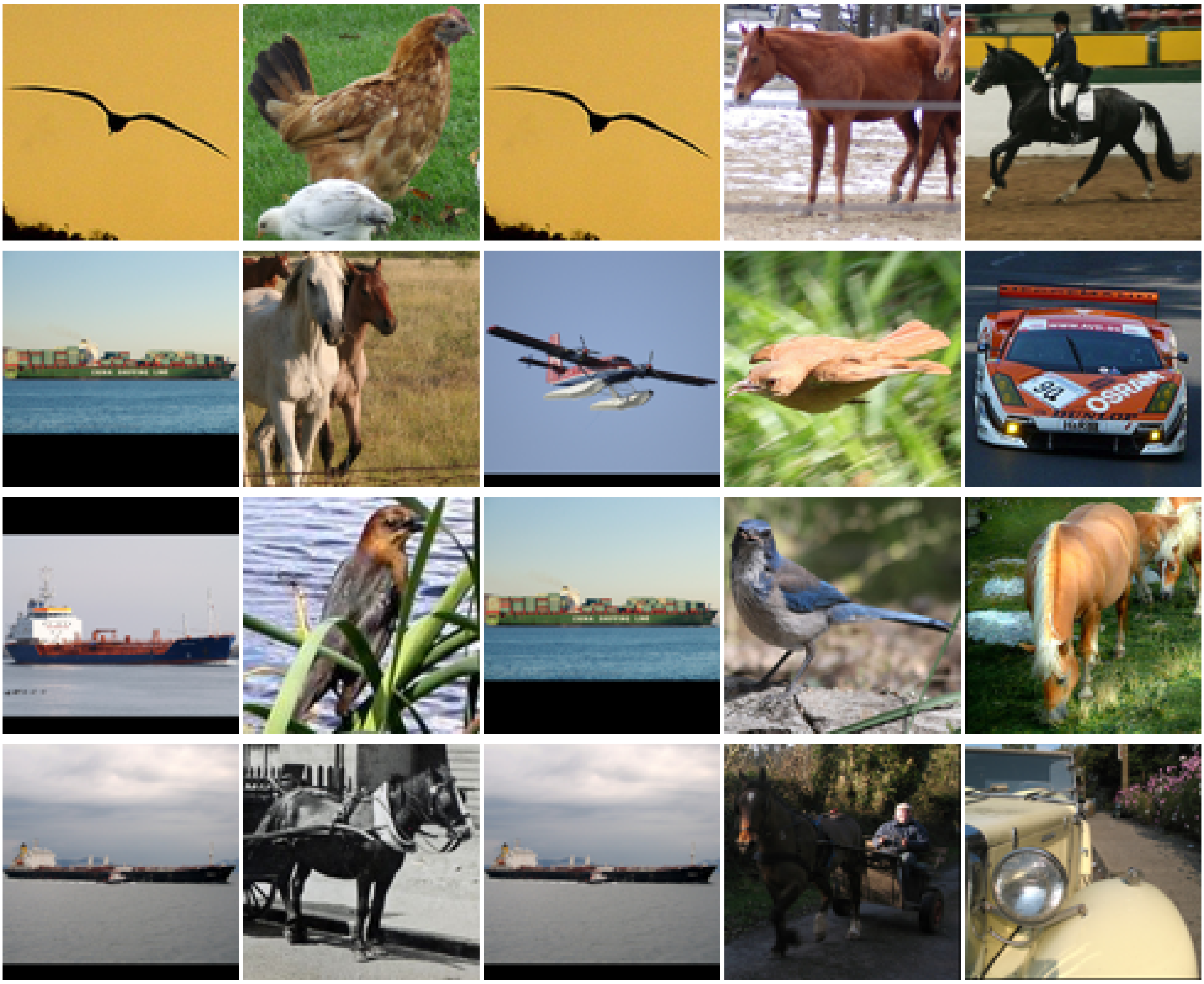
Settings

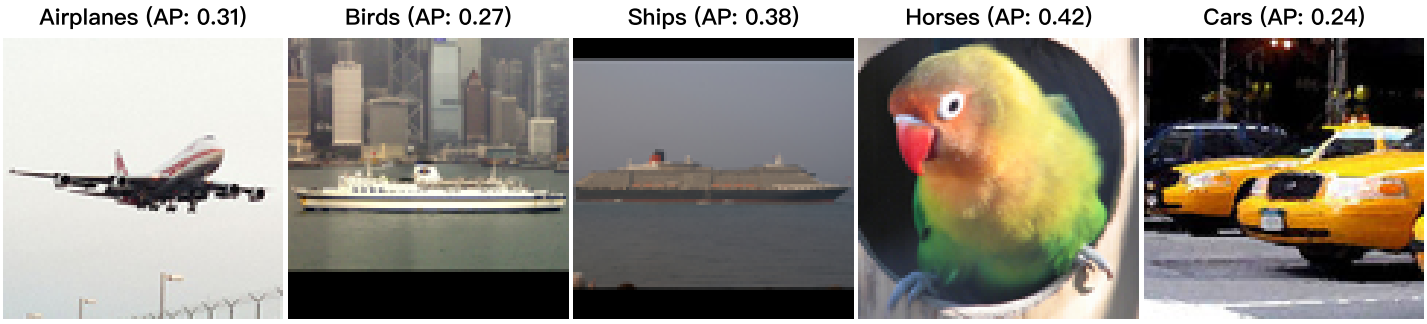
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	400 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.32)

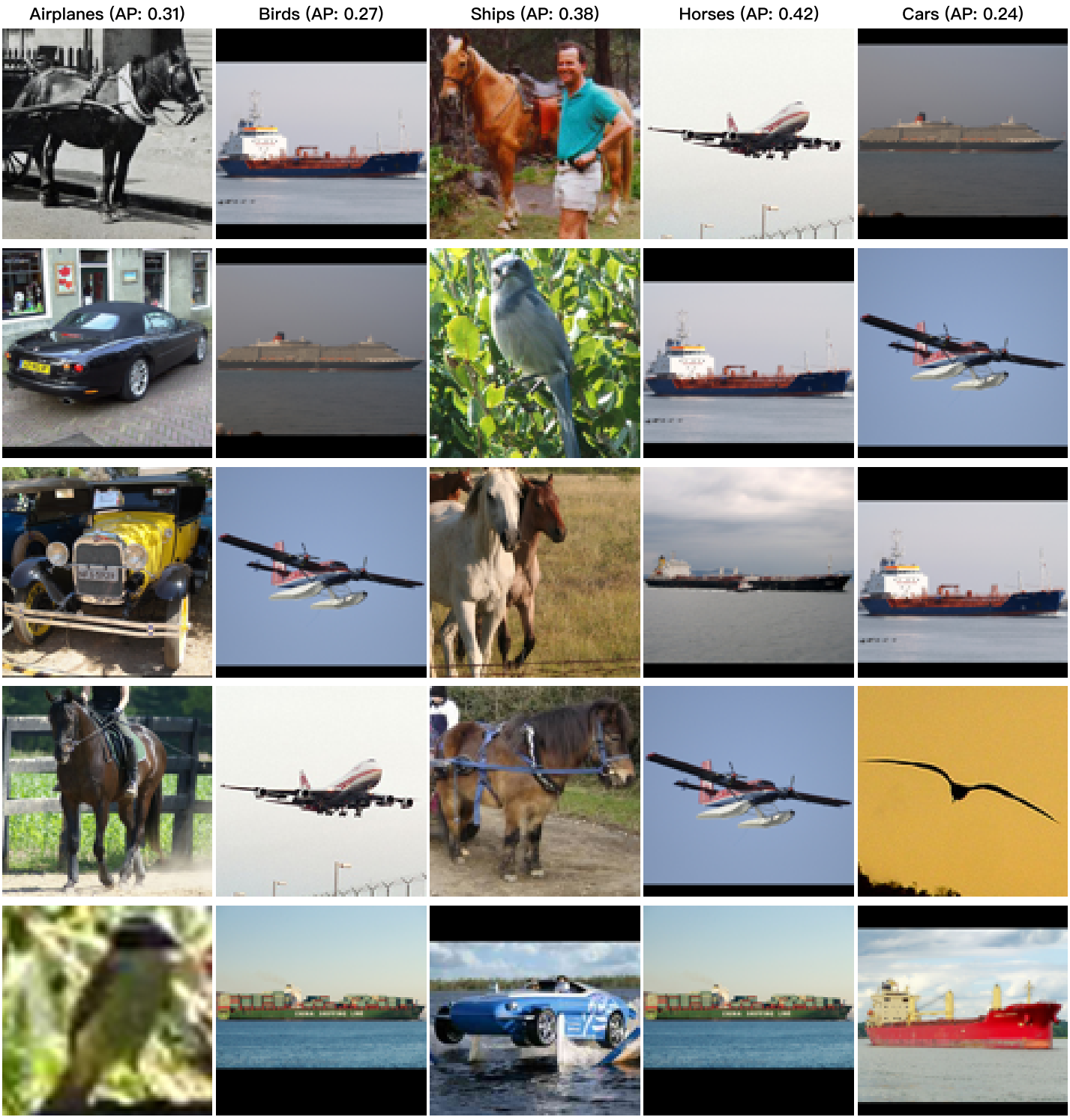
Positive Order

Airplanes (AP: 0.31) **Birds (AP: 0.27)** **Ships (AP: 0.38)** **Horses (AP: 0.42)** **Cars (AP: 0.24)**





Reverse order



Settings

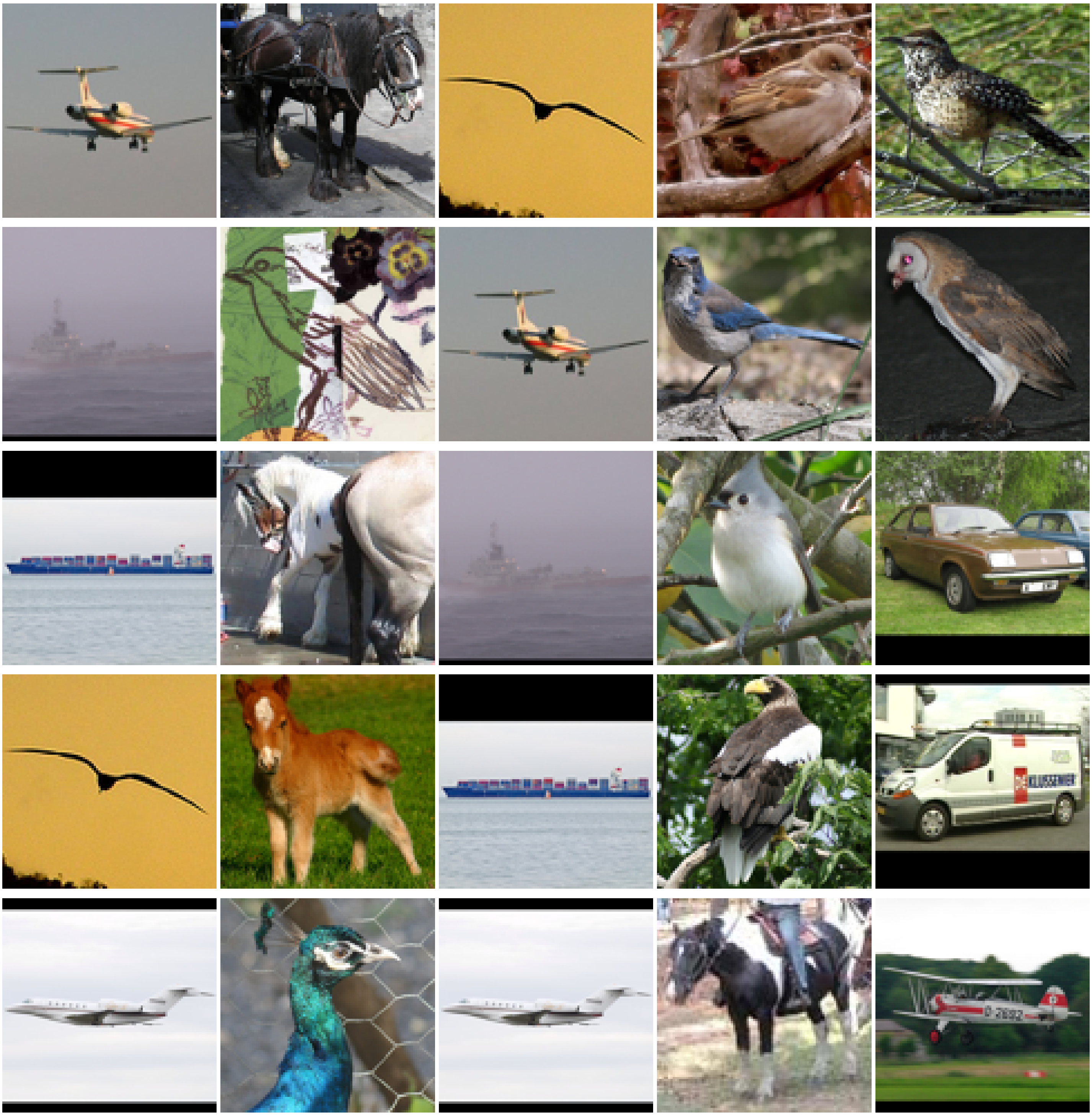
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	400 words

color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.33)

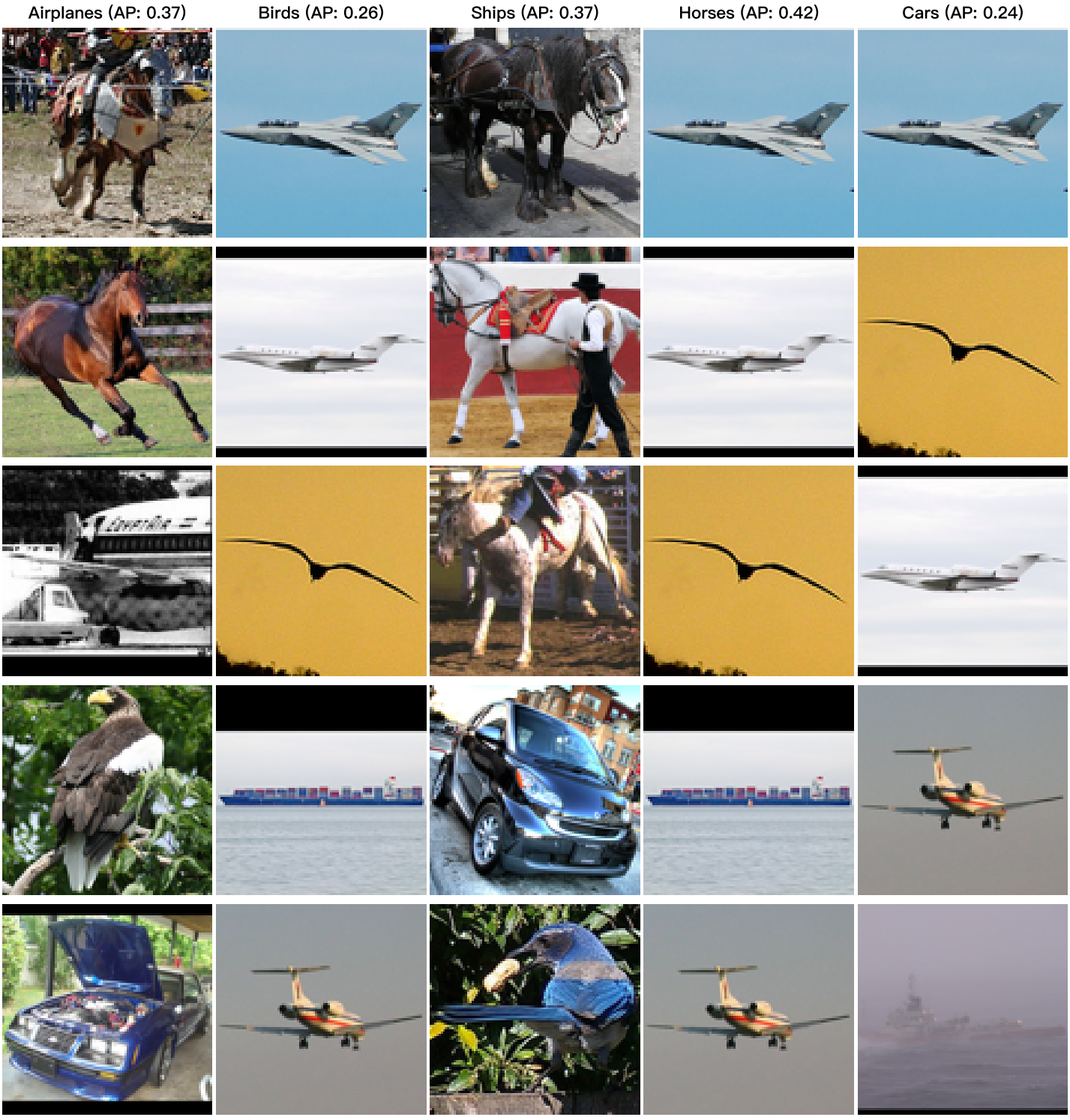
Positive Order

Airplanes (AP: 0.37)	Birds (AP: 0.26)	Ships (AP: 0.37)	Horses (AP: 0.42)	Cars (AP: 0.24)



Reverse order

Airplanes (AP: 0.37)	Birds (AP: 0.26)	Ships (AP: 0.37)	Horses (AP: 0.42)	Cars (AP: 0.24)

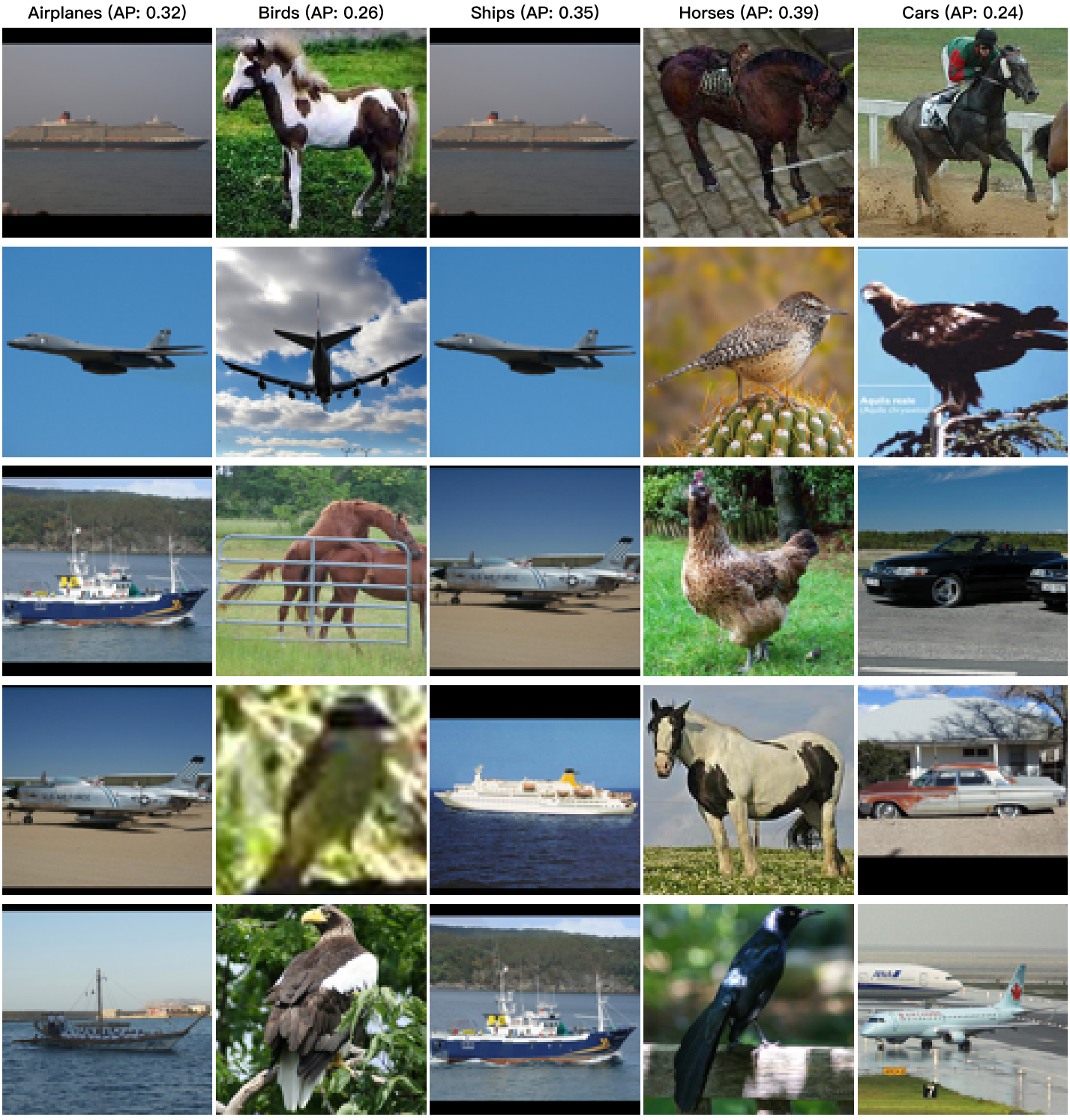


Settings

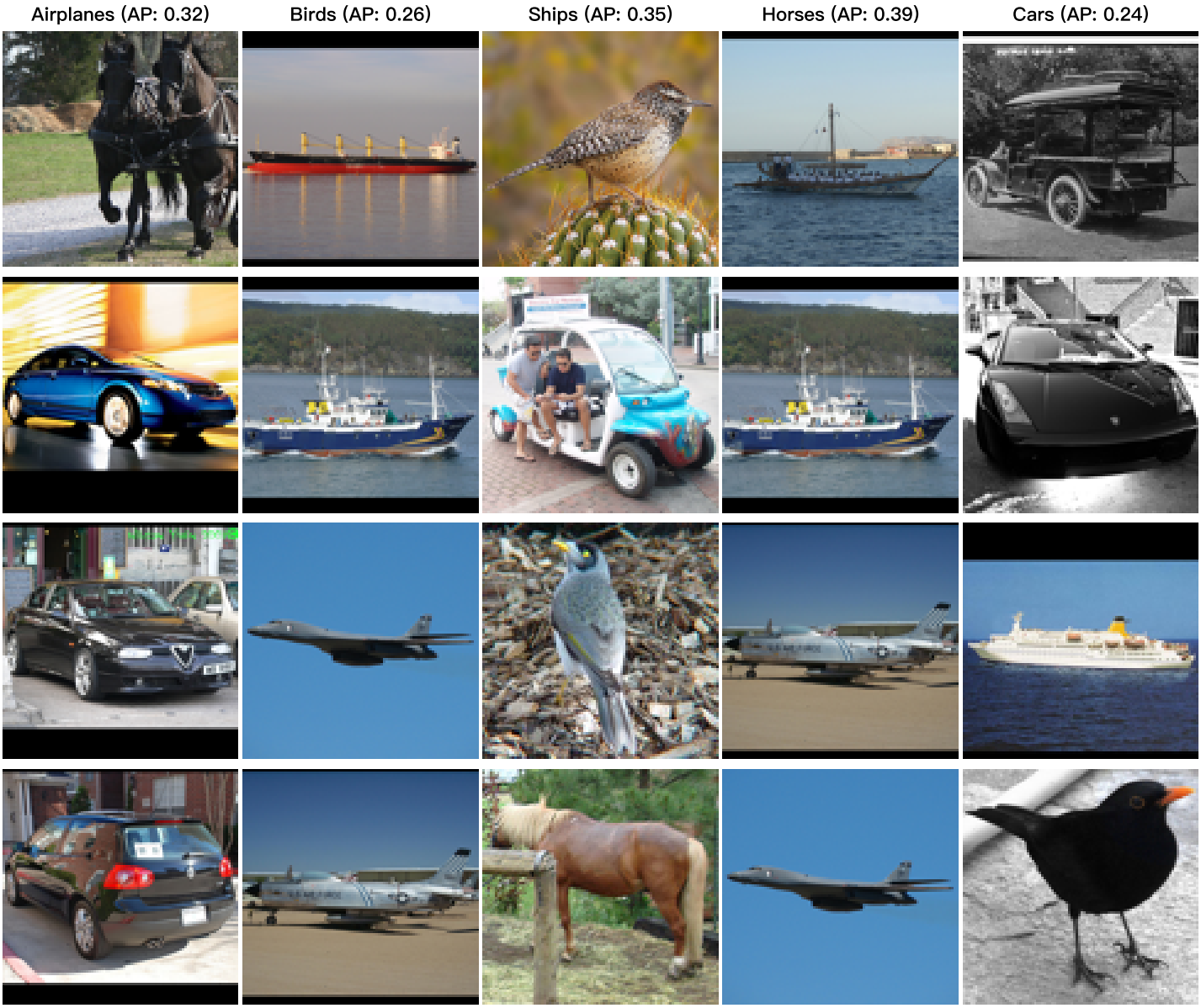
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	400 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.31)

Positive Order



Reverse order

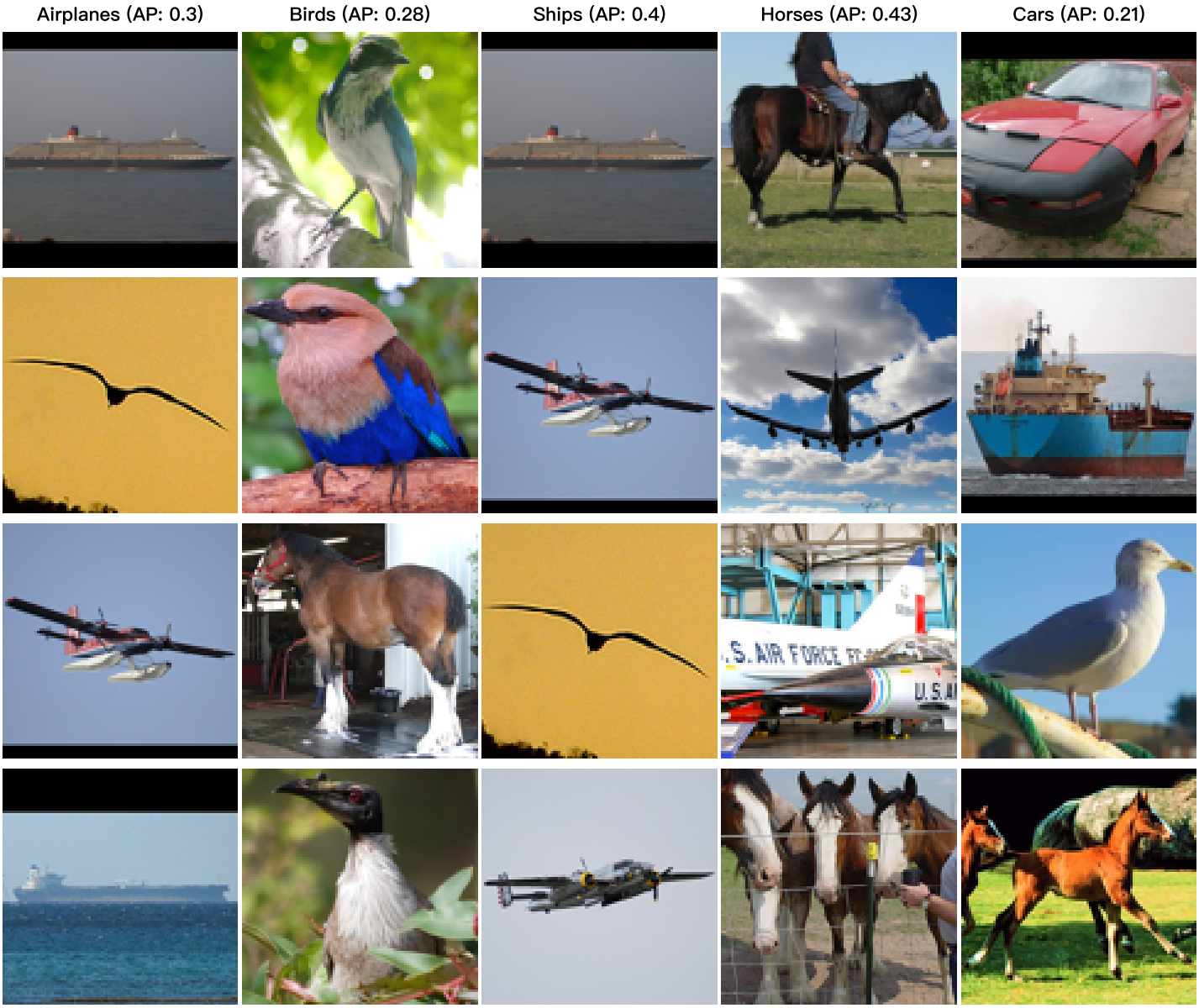


Settings

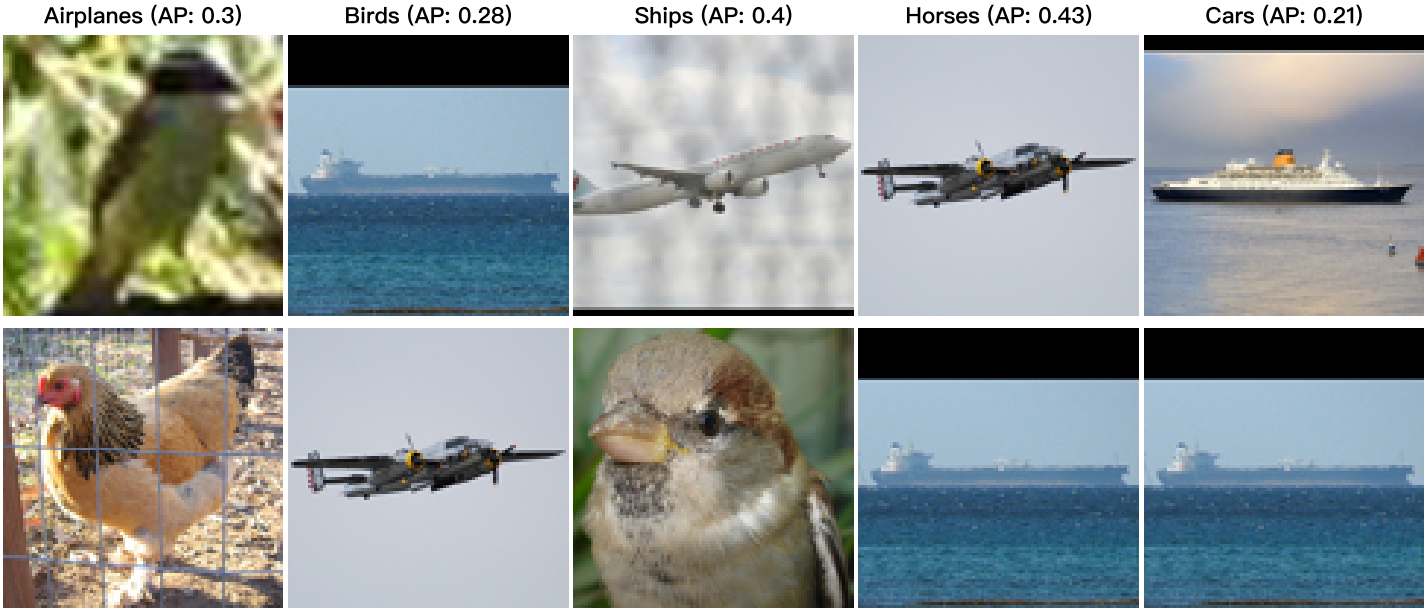
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	1000 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

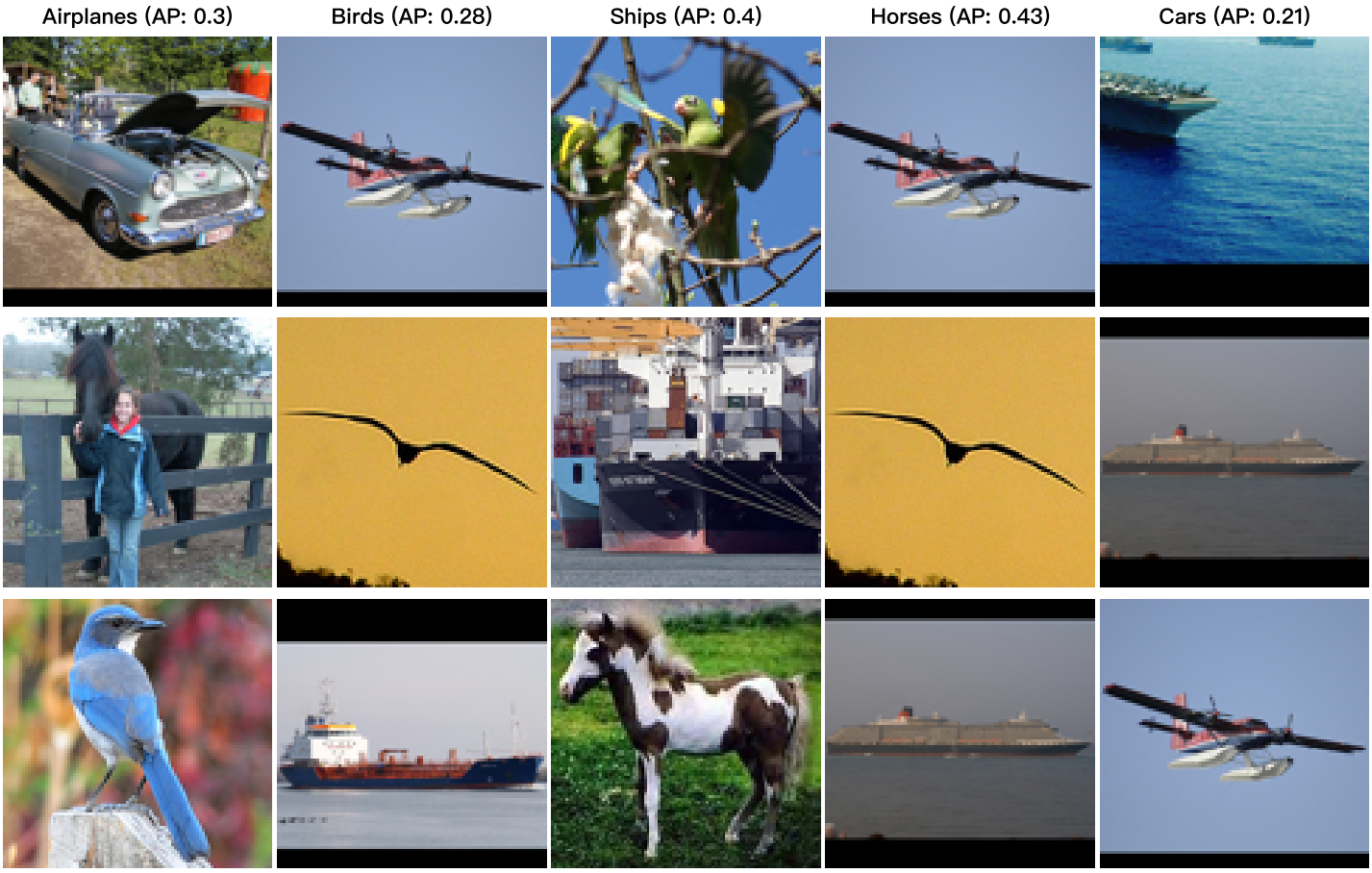
Prediction lists (MAP: 0.32)

Positive Order



Reverse order



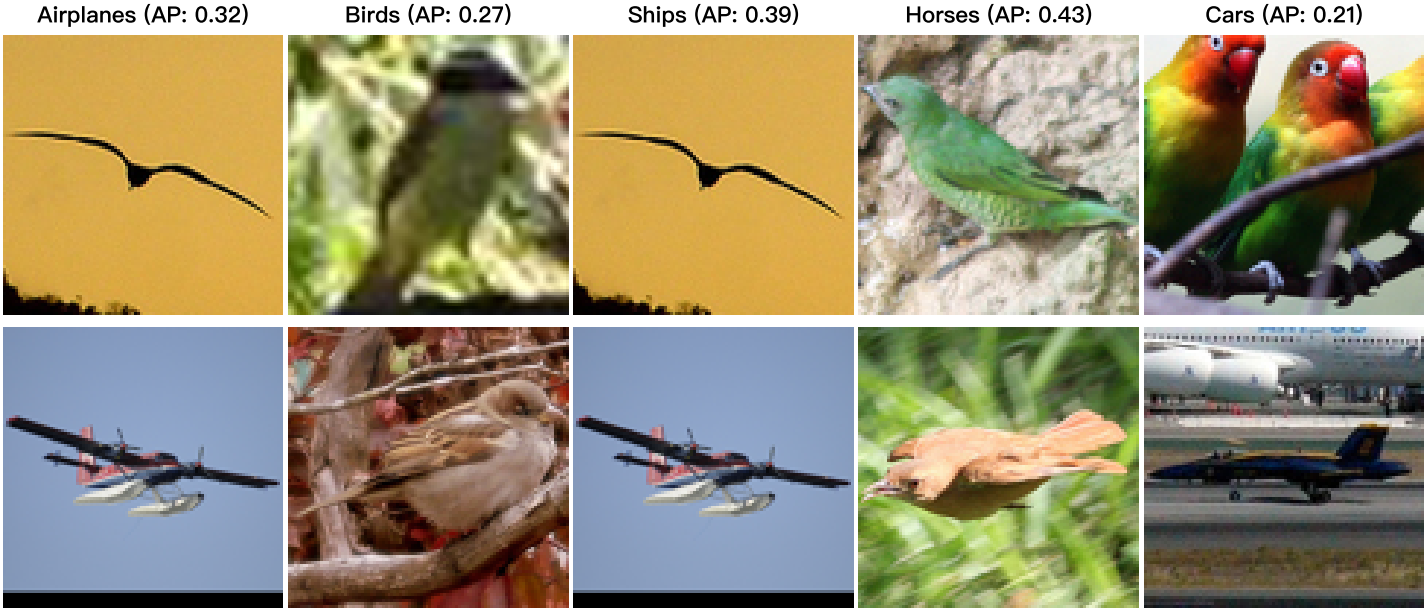


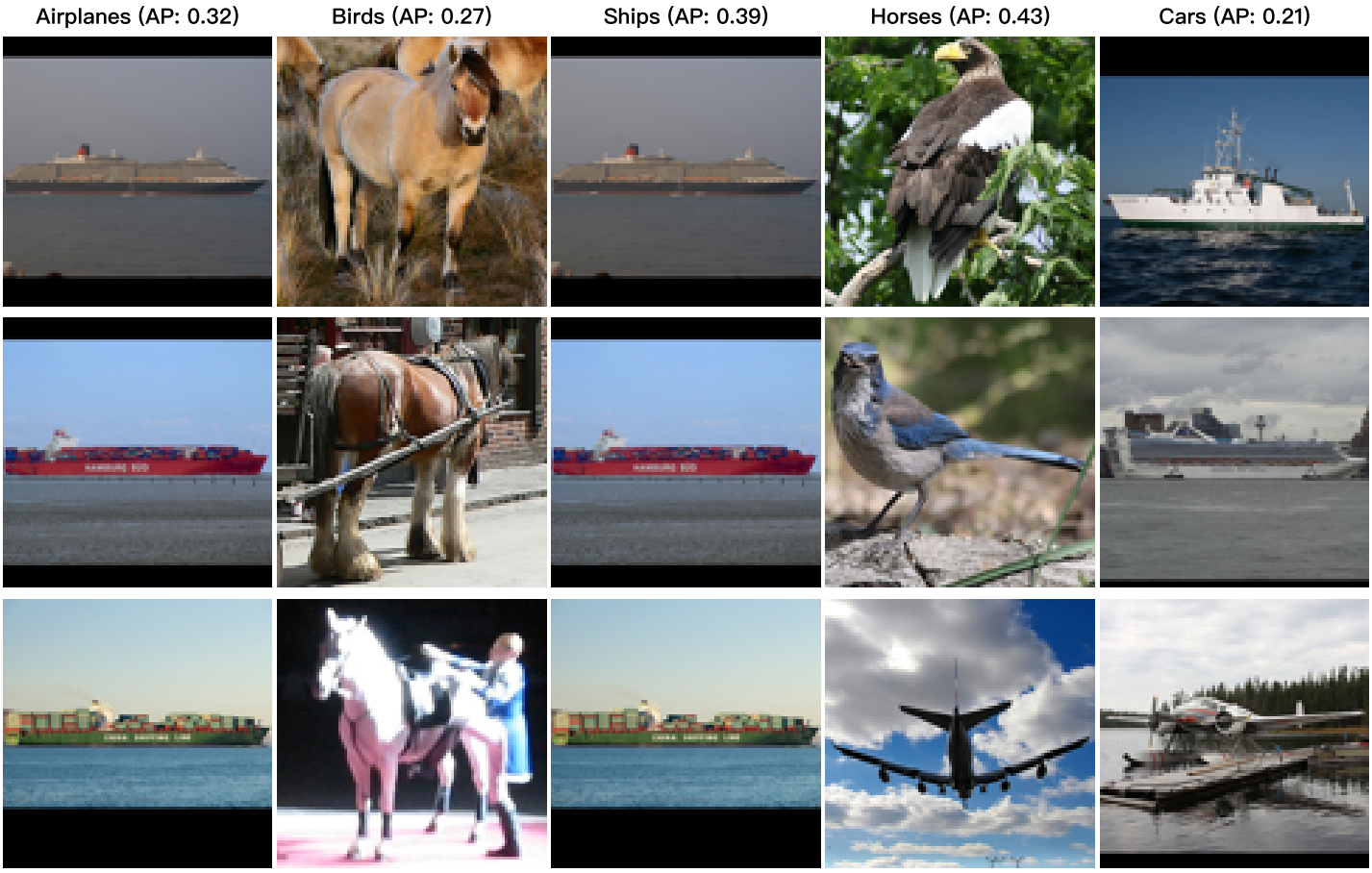
Settings

SIFT step size	5 px
SIFT sample method	key
Vocabulary size	1000 words
color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

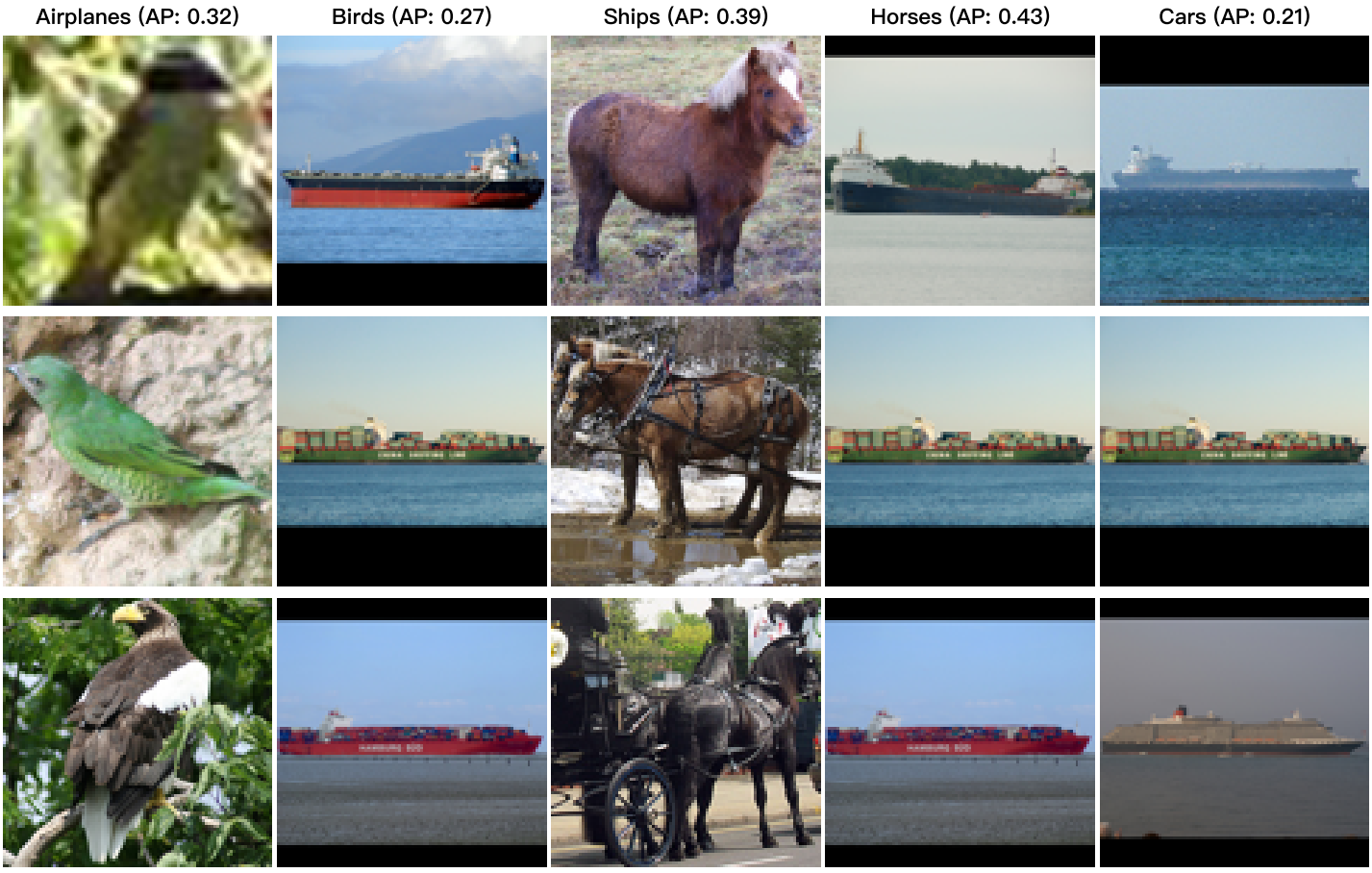
Prediction lists (MAP: 0.32)

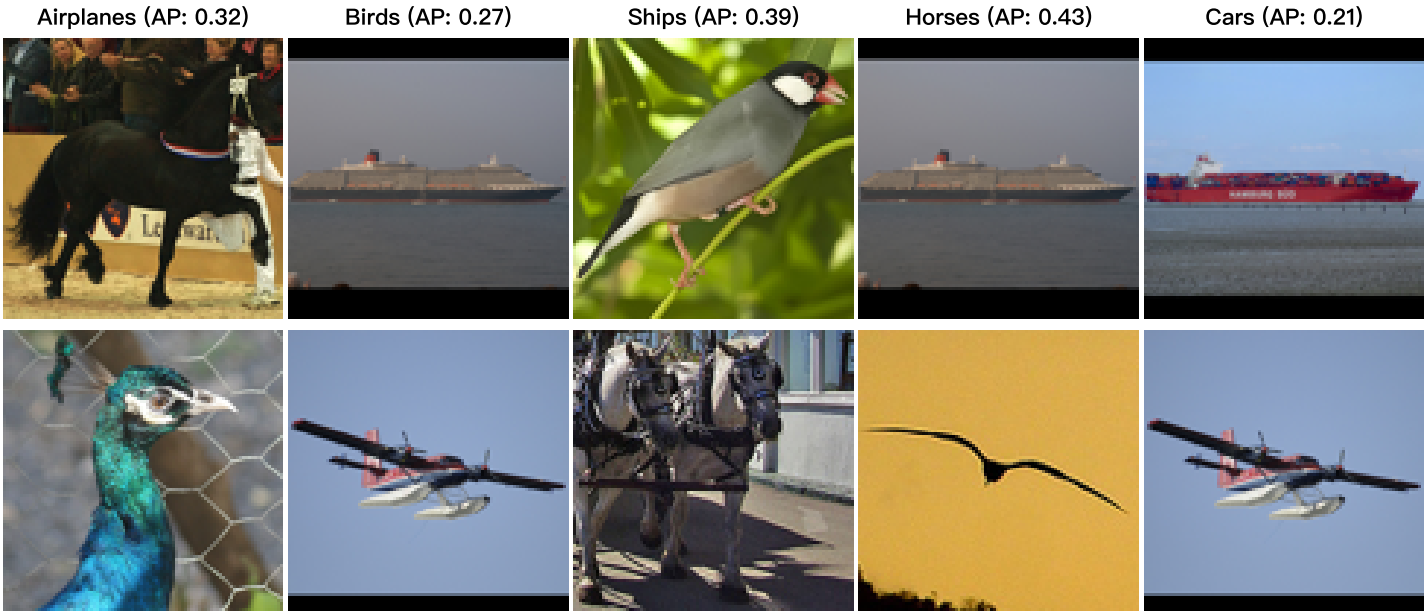
Positive Order





Reverse order



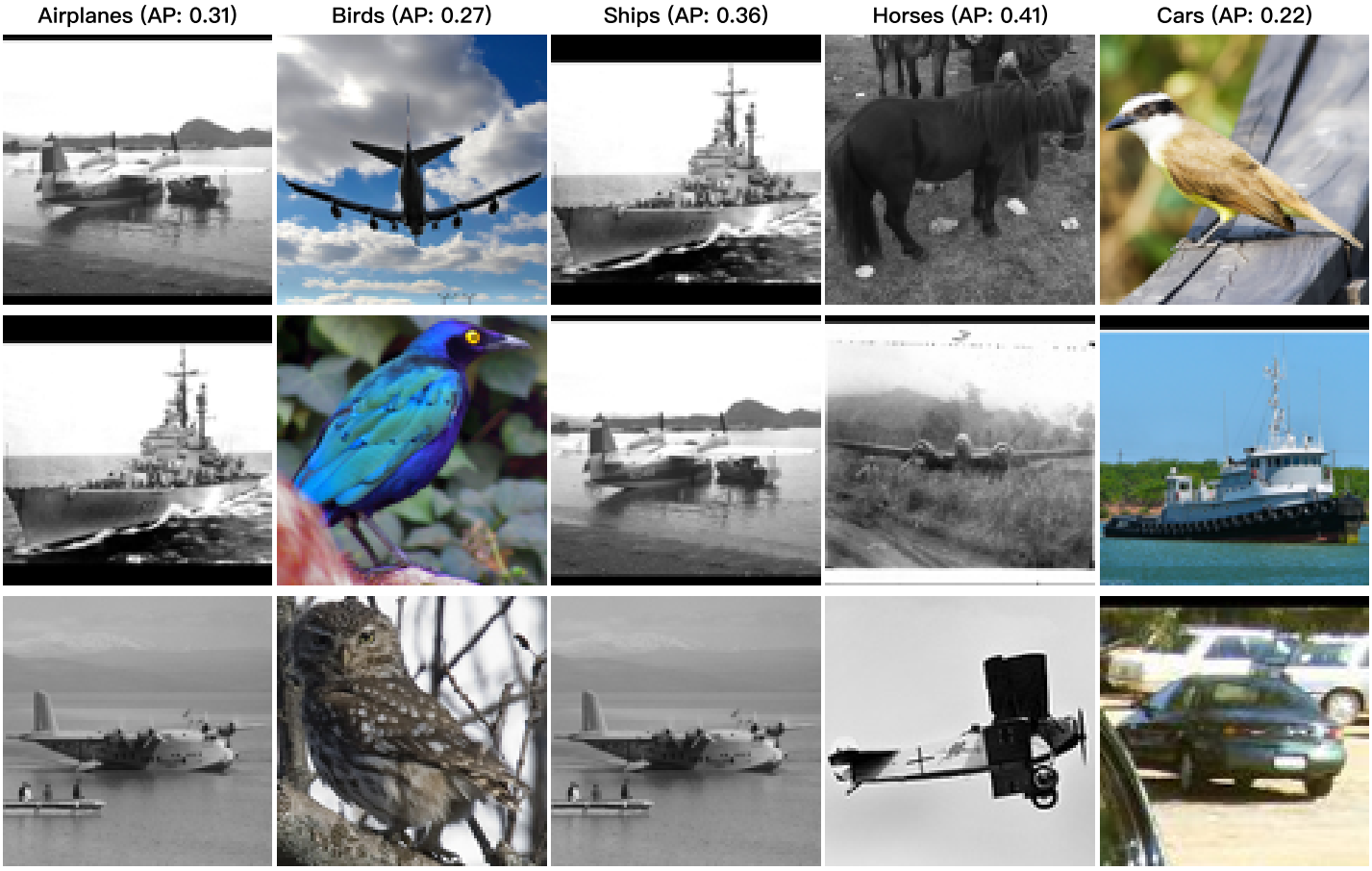


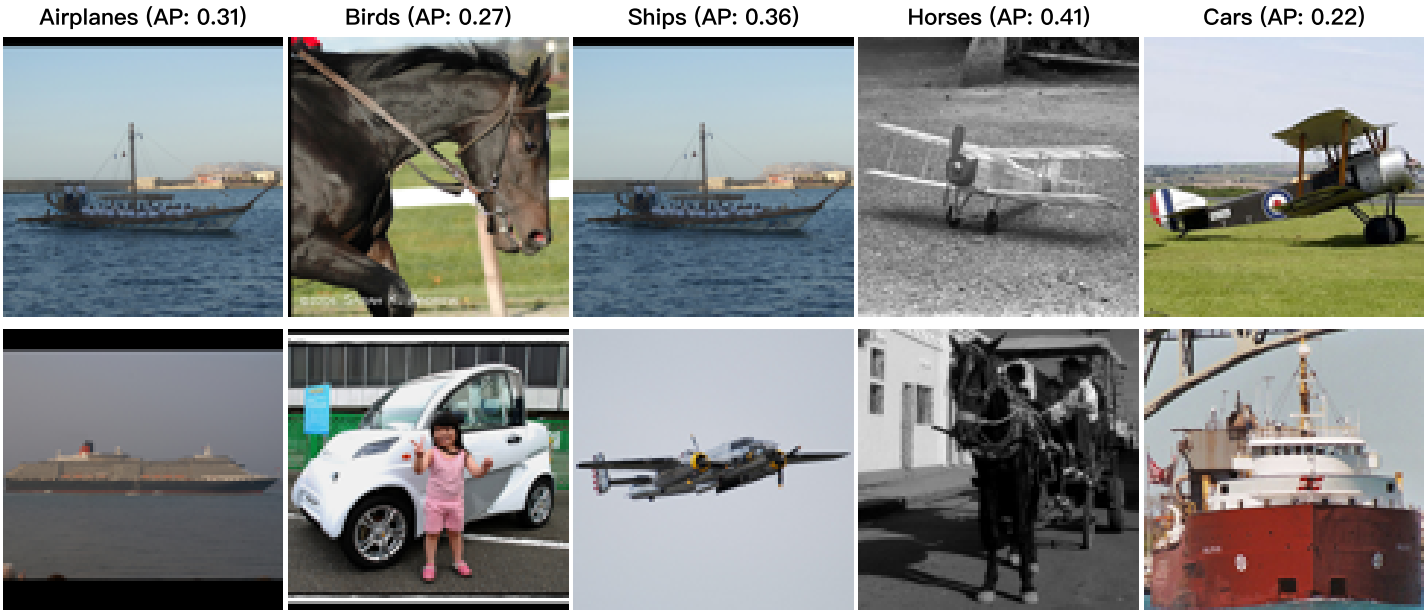
Settings

SIFT step size	5 px
SIFT sample method	key
Vocabulary size	1000 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

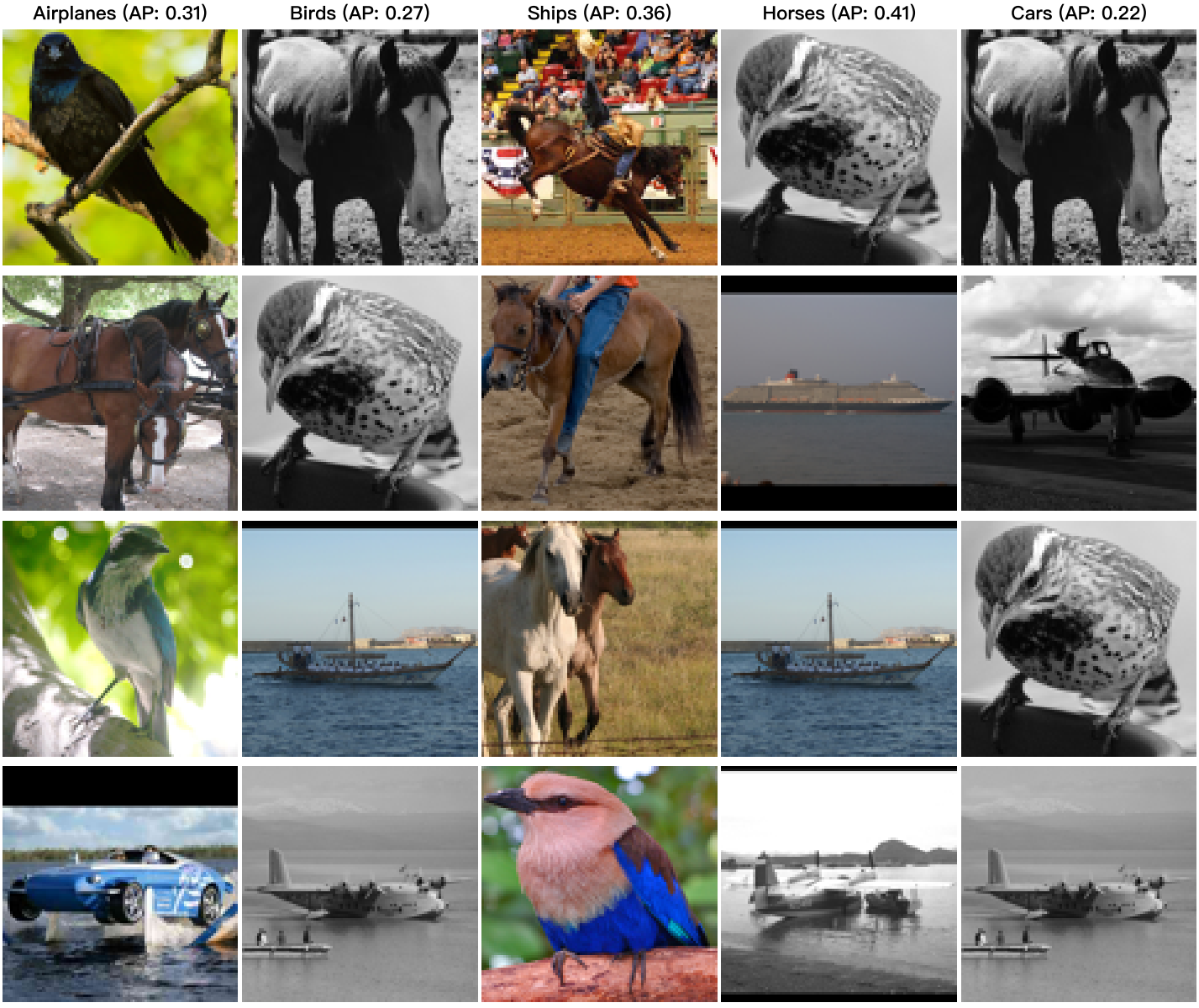
Prediction lists (MAP: 0.32)

Positive Order

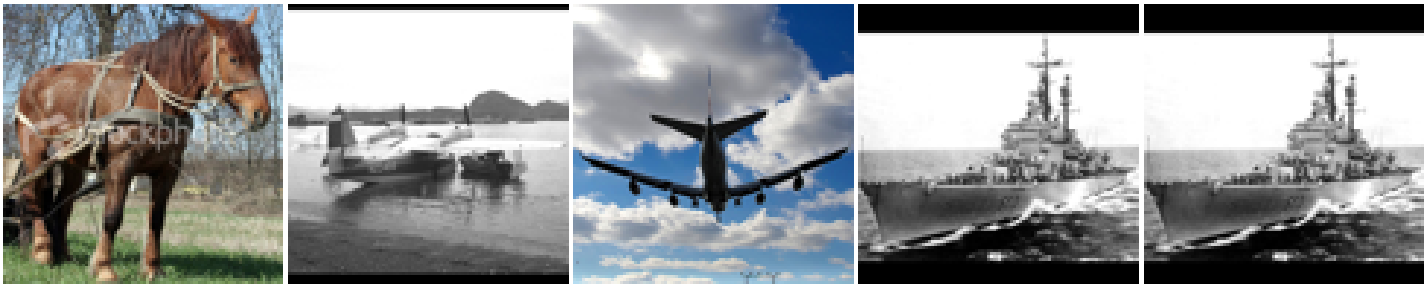




Reverse order



Airplanes (AP: 0.31) | Birds (AP: 0.27) | Ships (AP: 0.36) | Horses (AP: 0.41) | Cars (AP: 0.22)



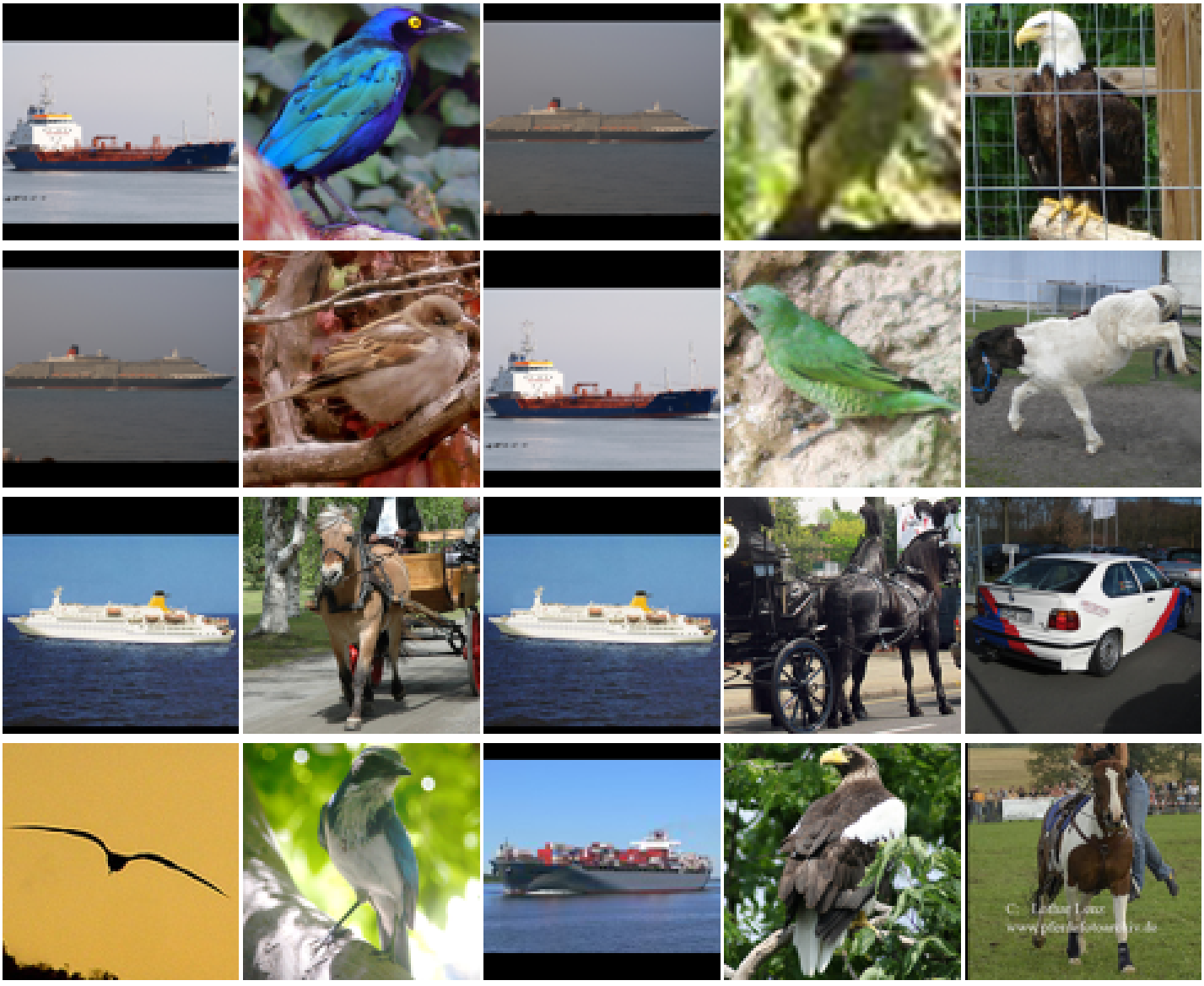
Settings

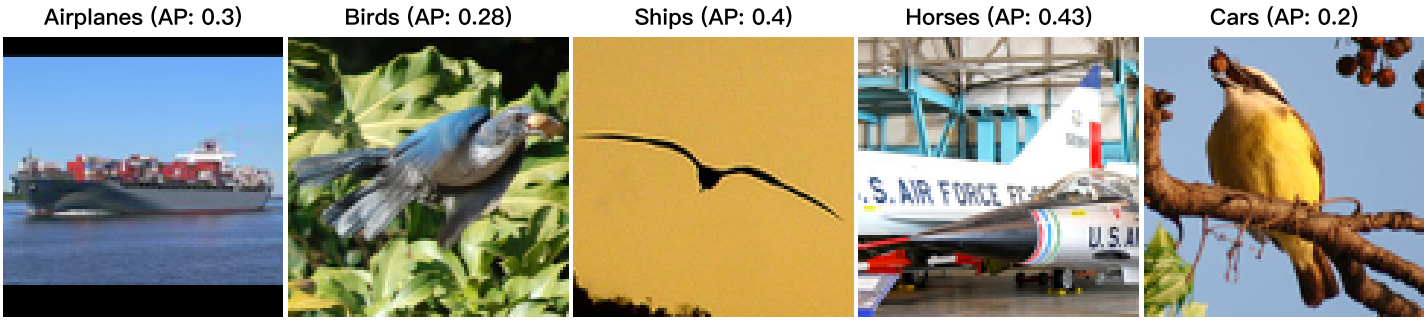
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	4000 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.32)

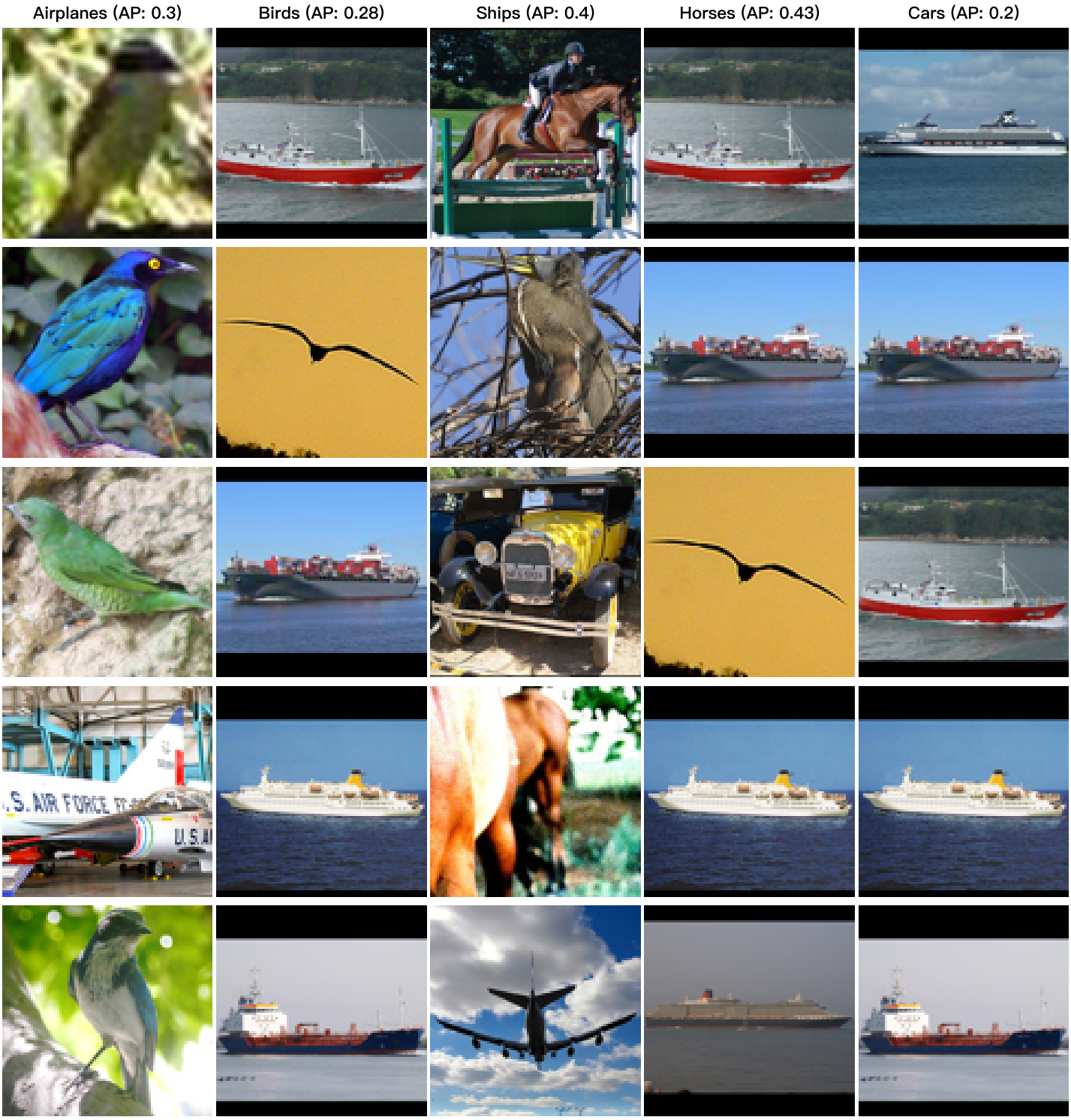
Positive Order

Airplanes (AP: 0.3) | Birds (AP: 0.28) | Ships (AP: 0.4) | Horses (AP: 0.43) | Cars (AP: 0.2)





Reverse order



Settings

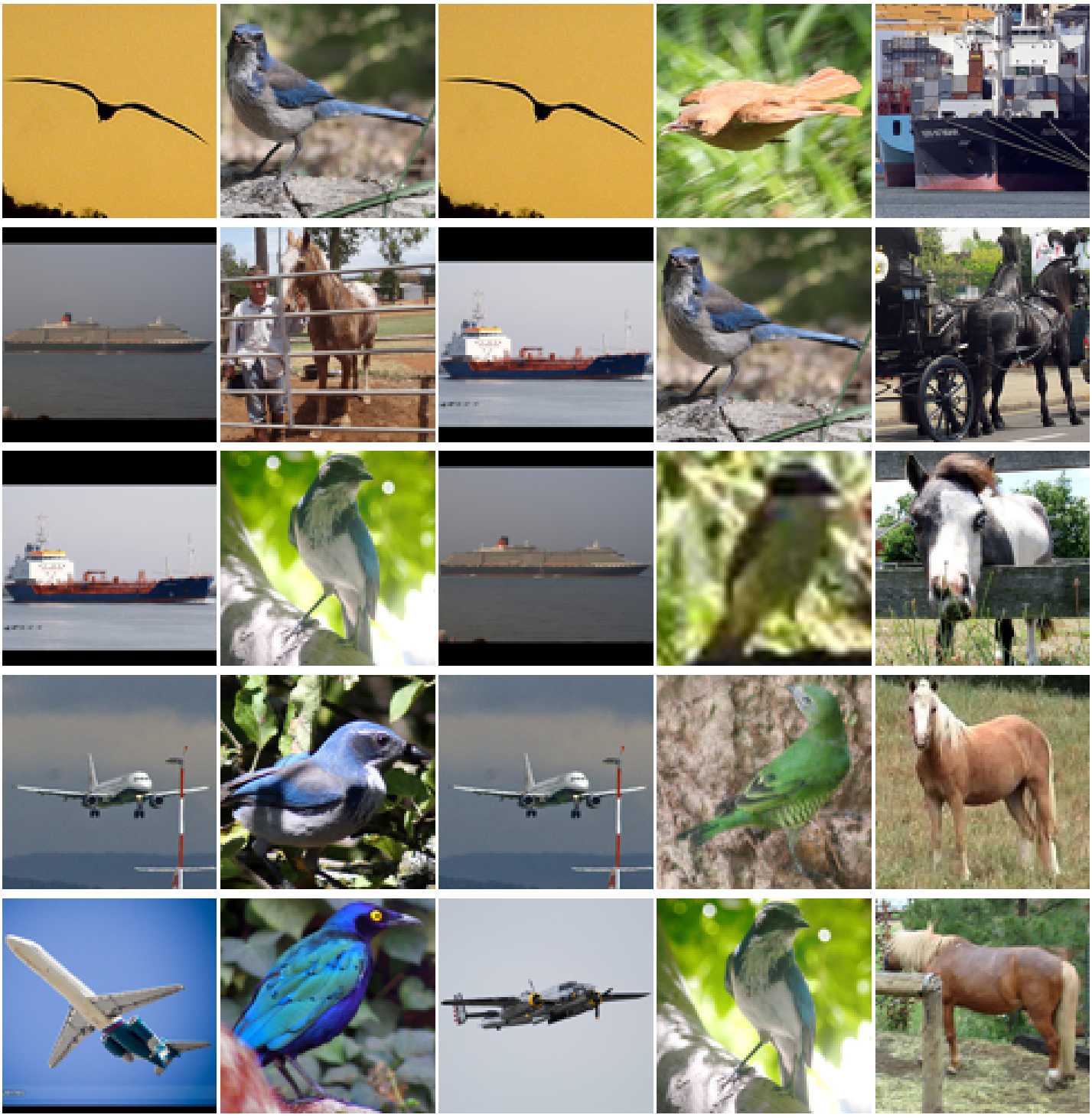
SIFT step size	5 px
SIFT sample method	key
Vocabulary size	4000 words

color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.32)

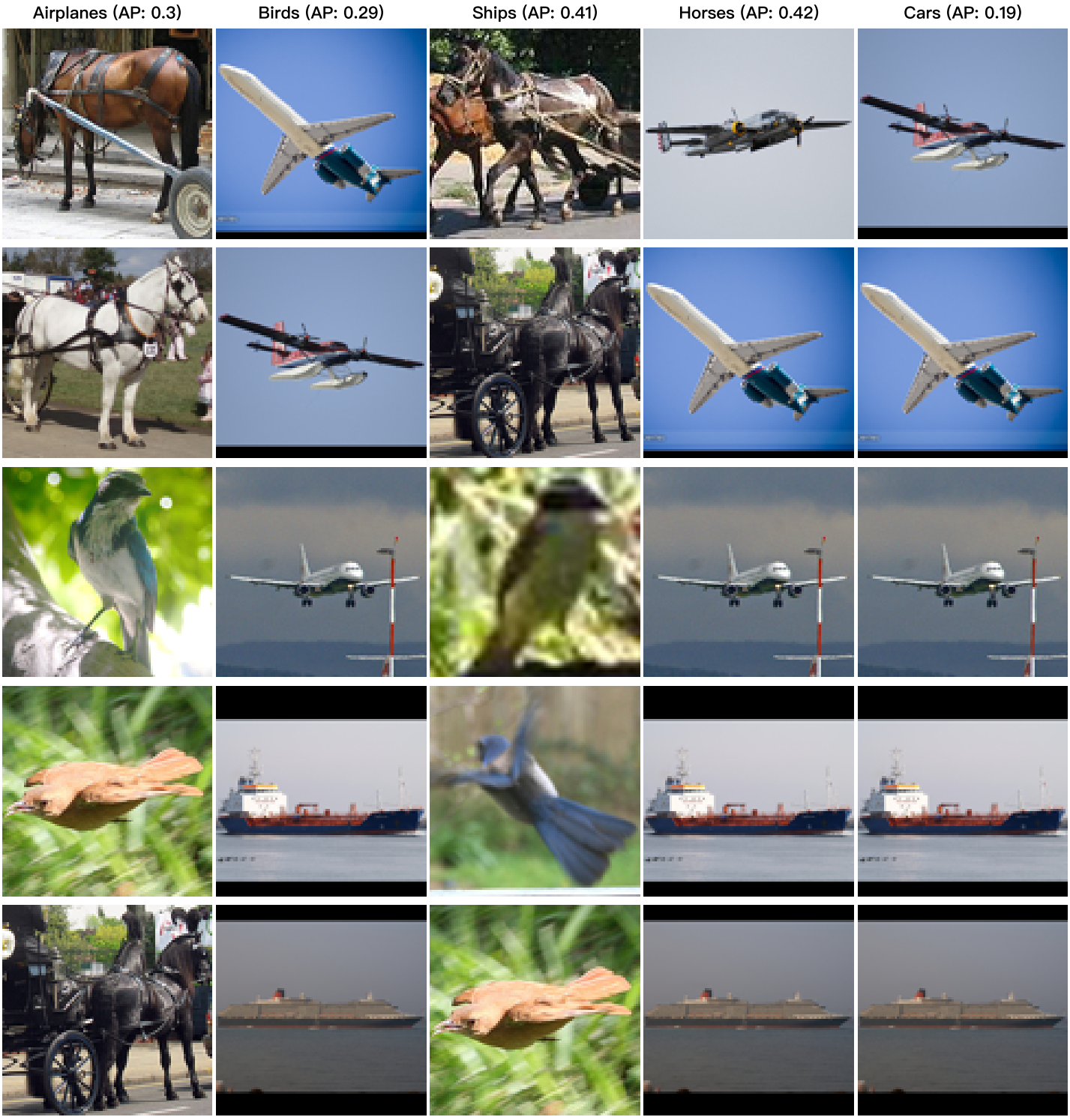
Positive Order

Airplanes (AP: 0.3)	Birds (AP: 0.29)	Ships (AP: 0.41)	Horses (AP: 0.42)	Cars (AP: 0.19)



Reverse order

Airplanes (AP: 0.3)	Birds (AP: 0.29)	Ships (AP: 0.41)	Horses (AP: 0.42)	Cars (AP: 0.19)

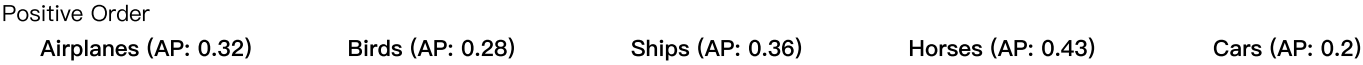


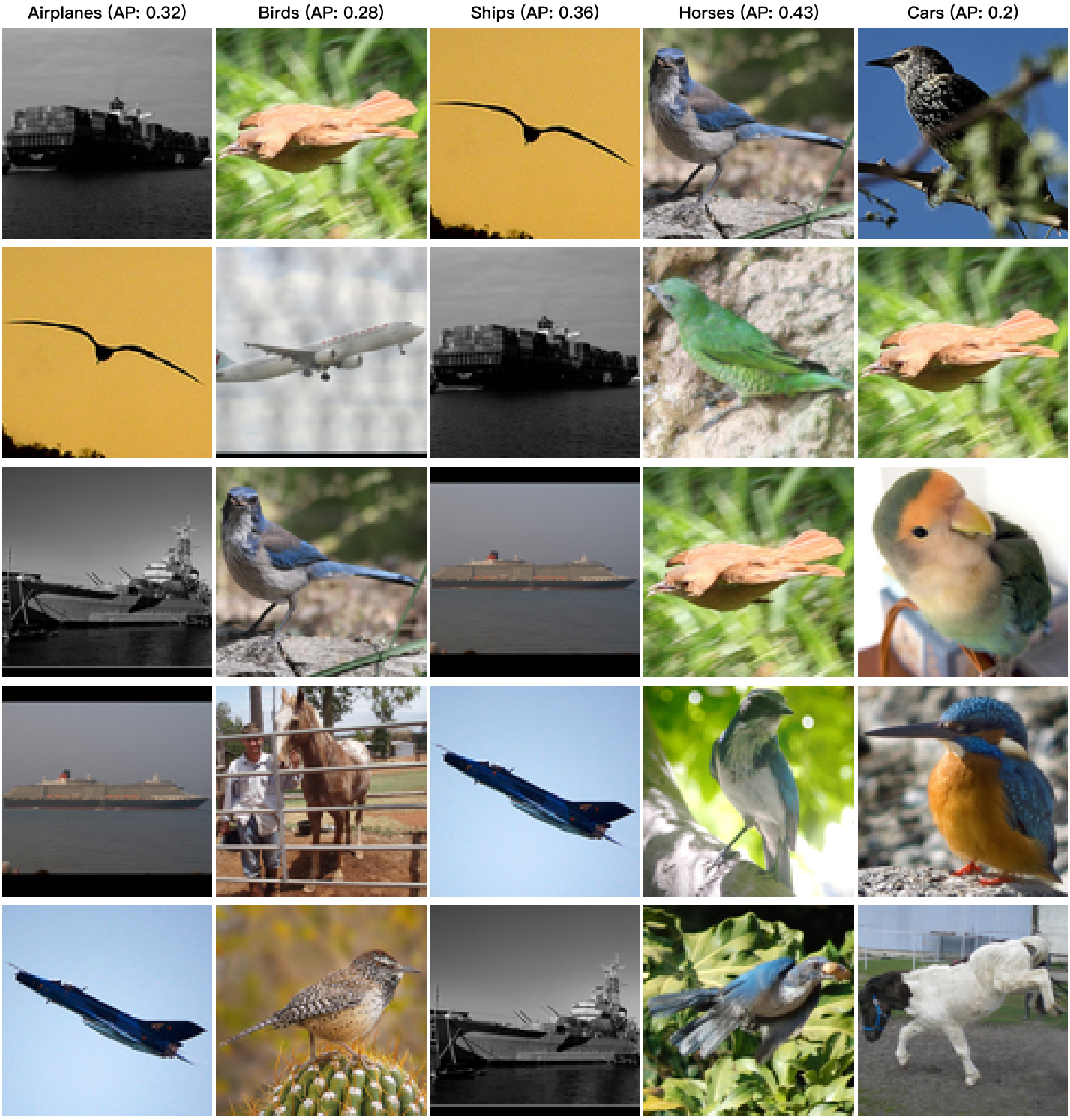
Settings

SIFT step size	5 px
SIFT sample method	key
Vocabulary size	4000 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

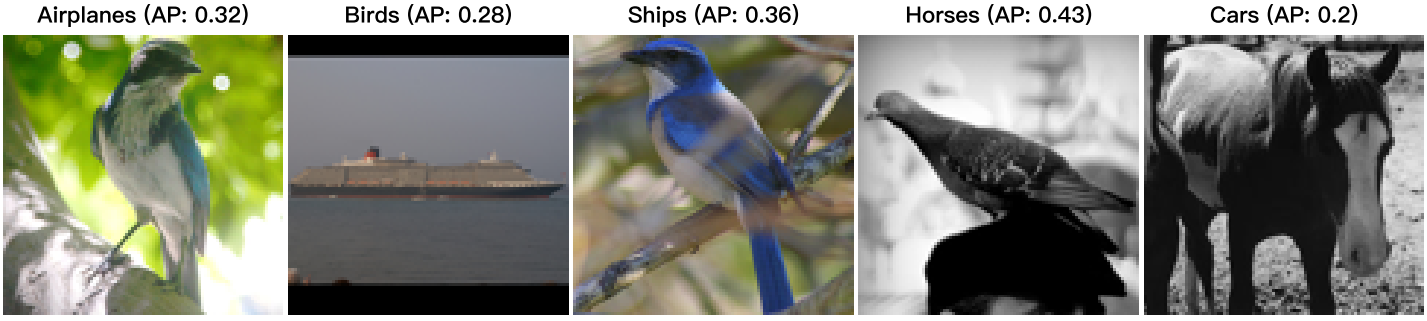
Prediction lists (MAP: 0.32)

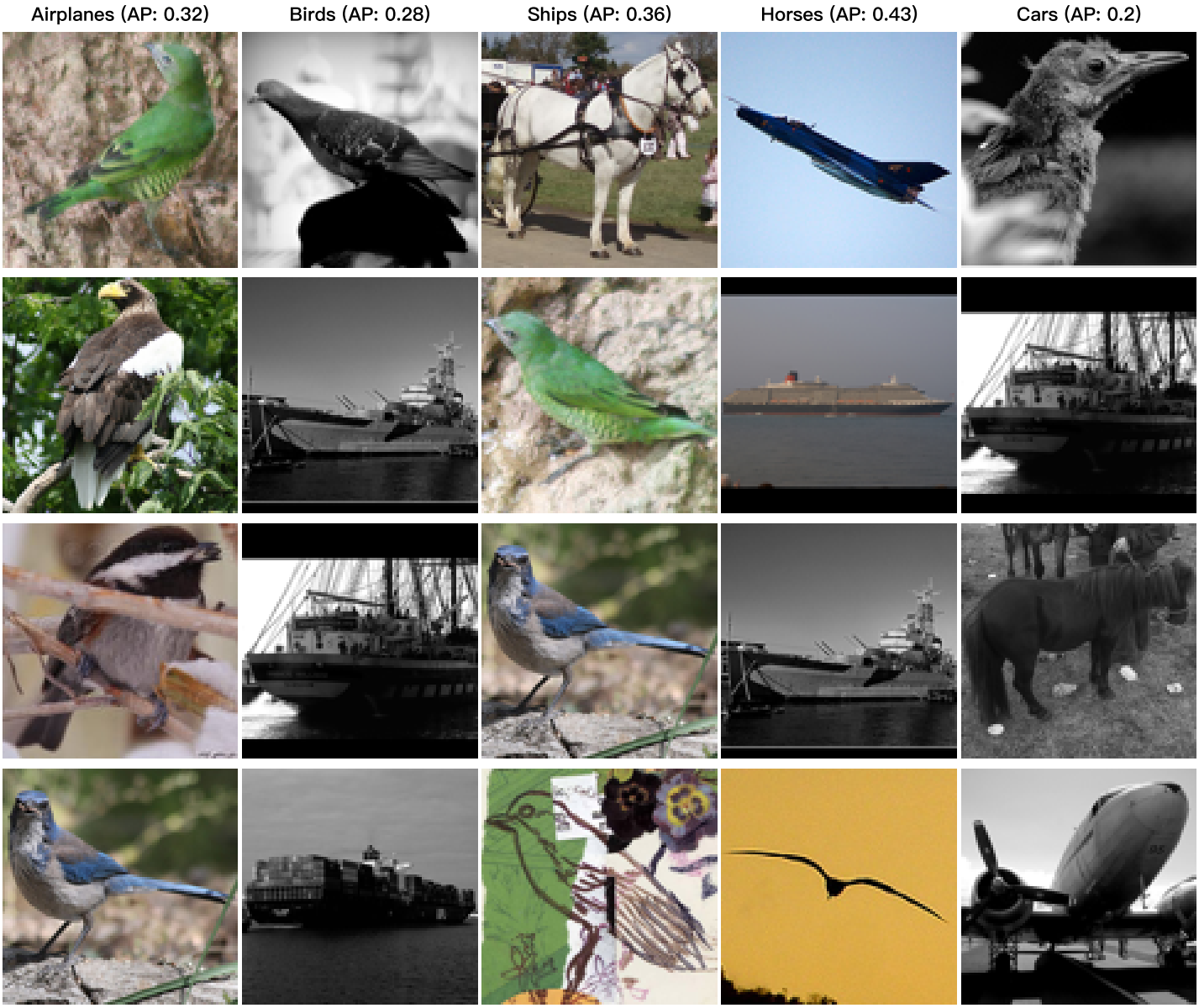
Positive Order





Reverse order



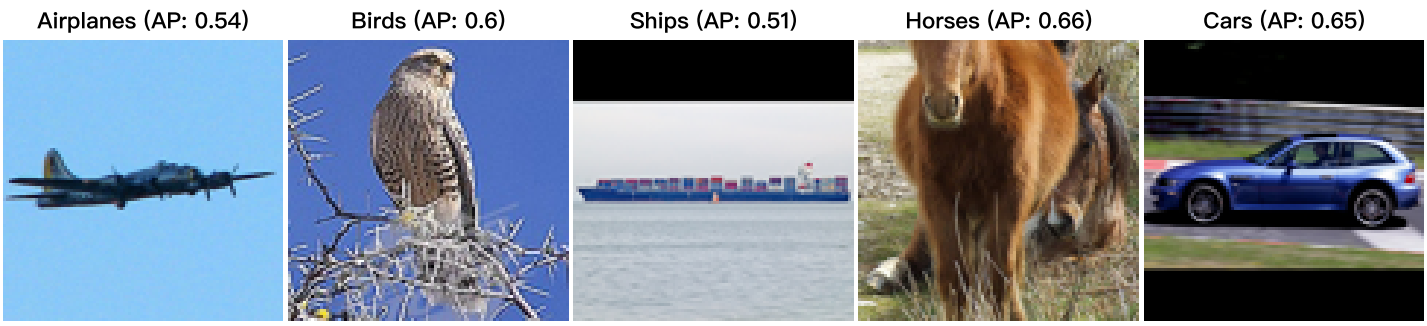


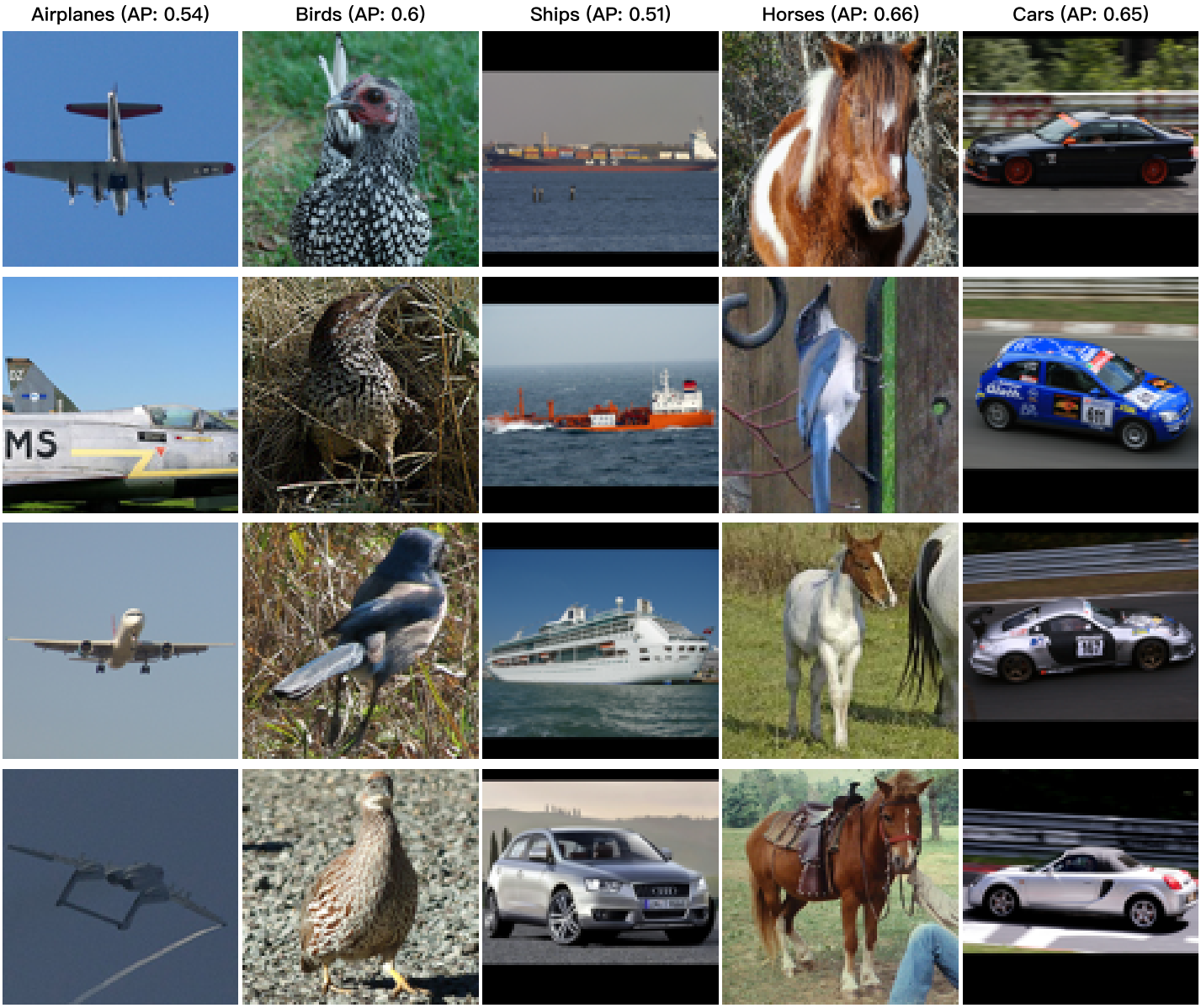
Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	400 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

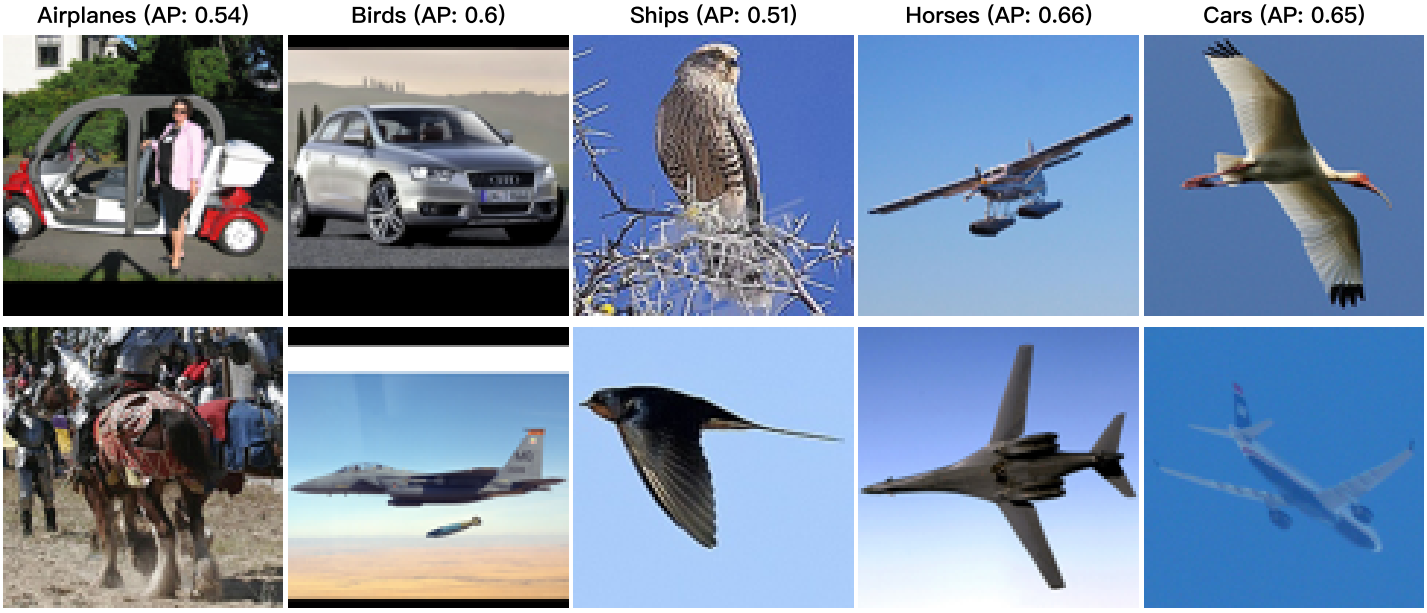
Prediction lists (MAP: 0.59)

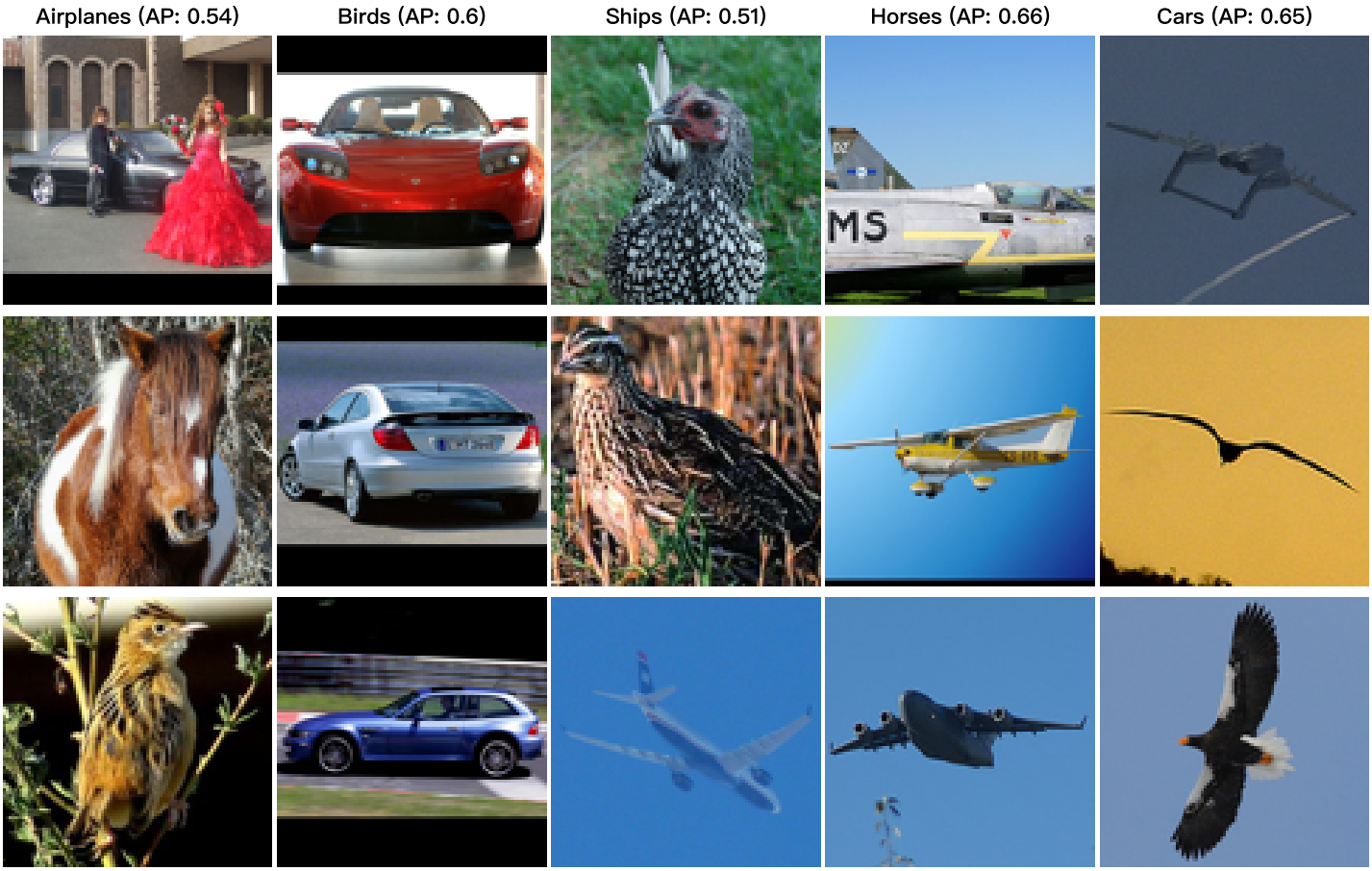
Positive Order





Reverse order



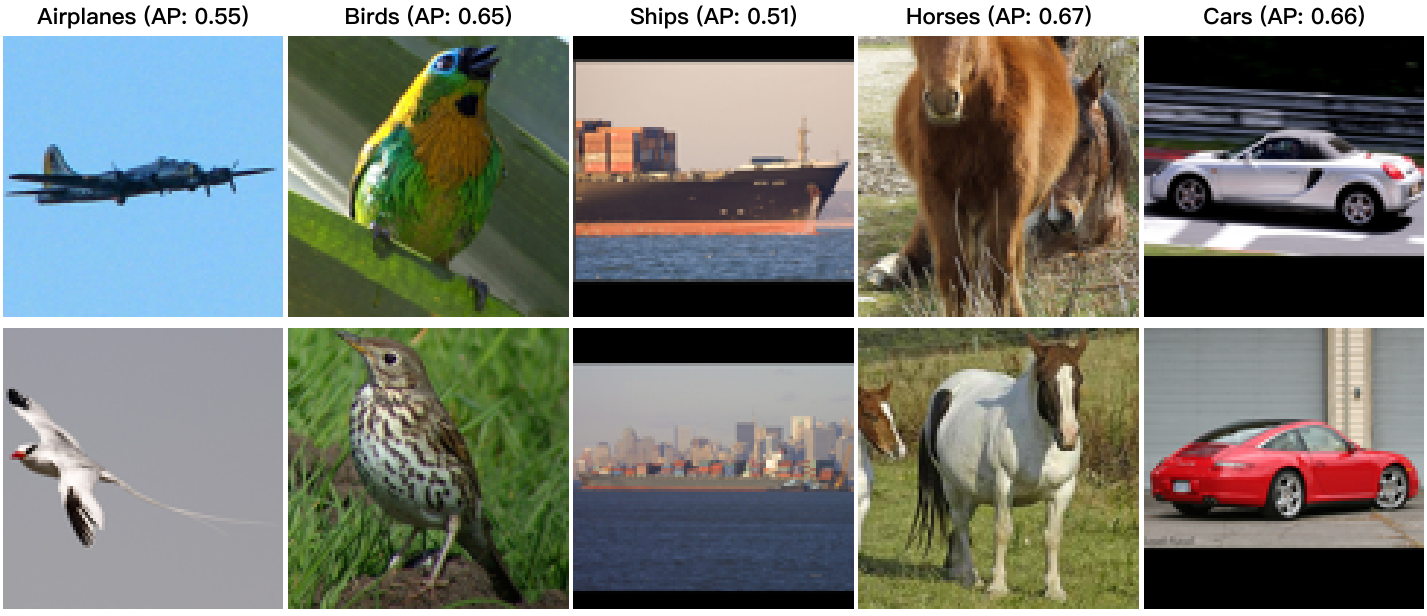


Settings

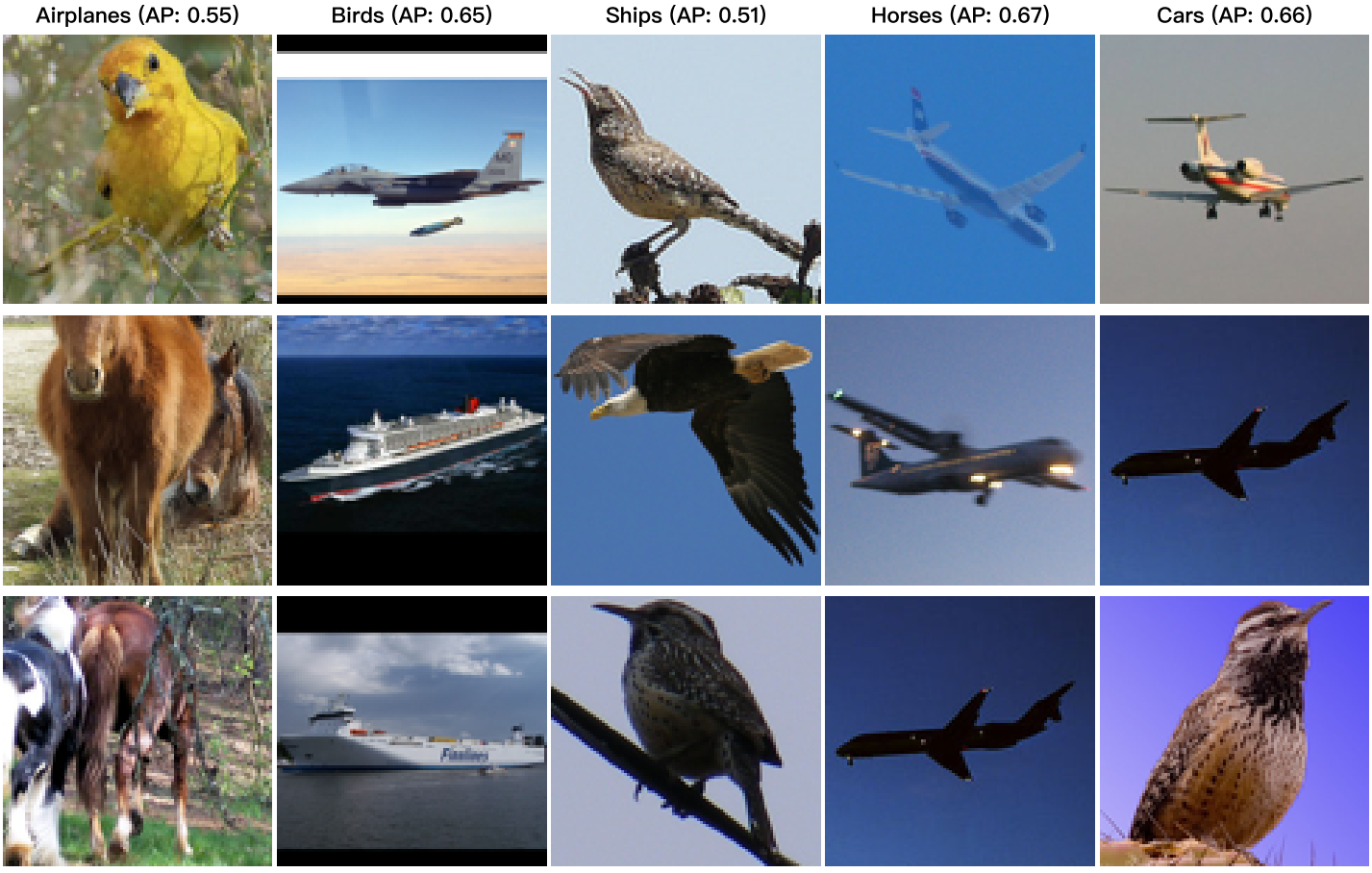
SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	400 words
color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

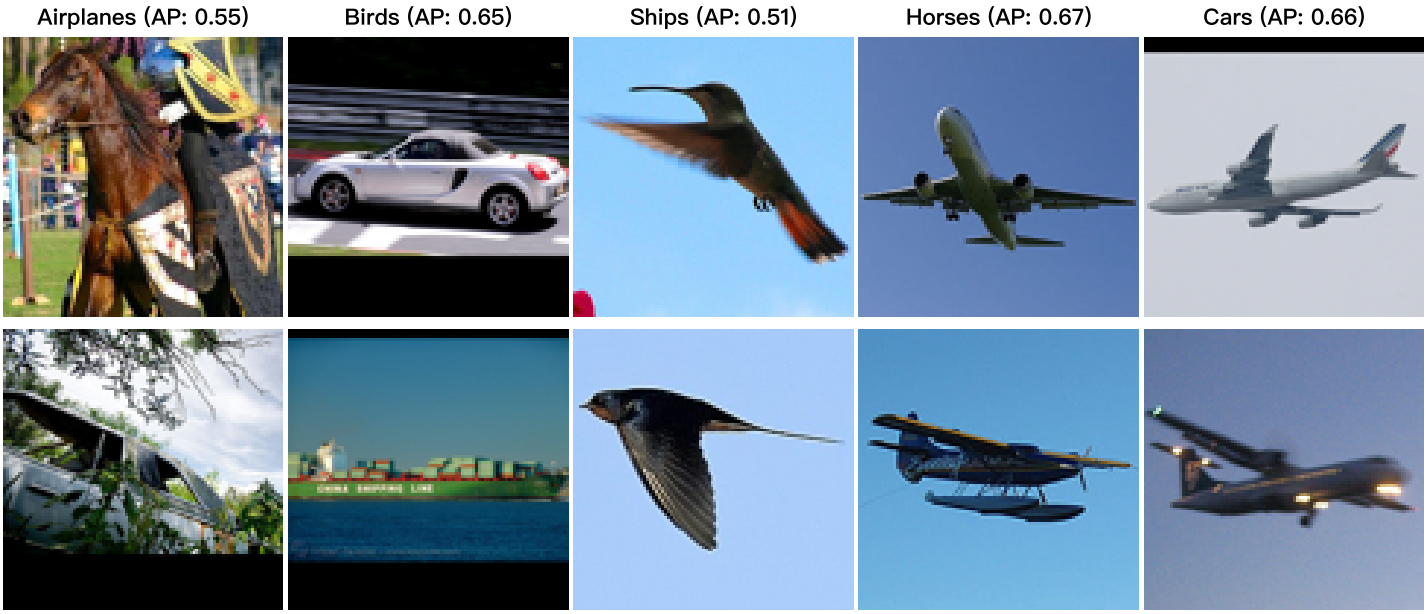
Prediction lists (MAP: 0.6)

Positive Order



Reverse order



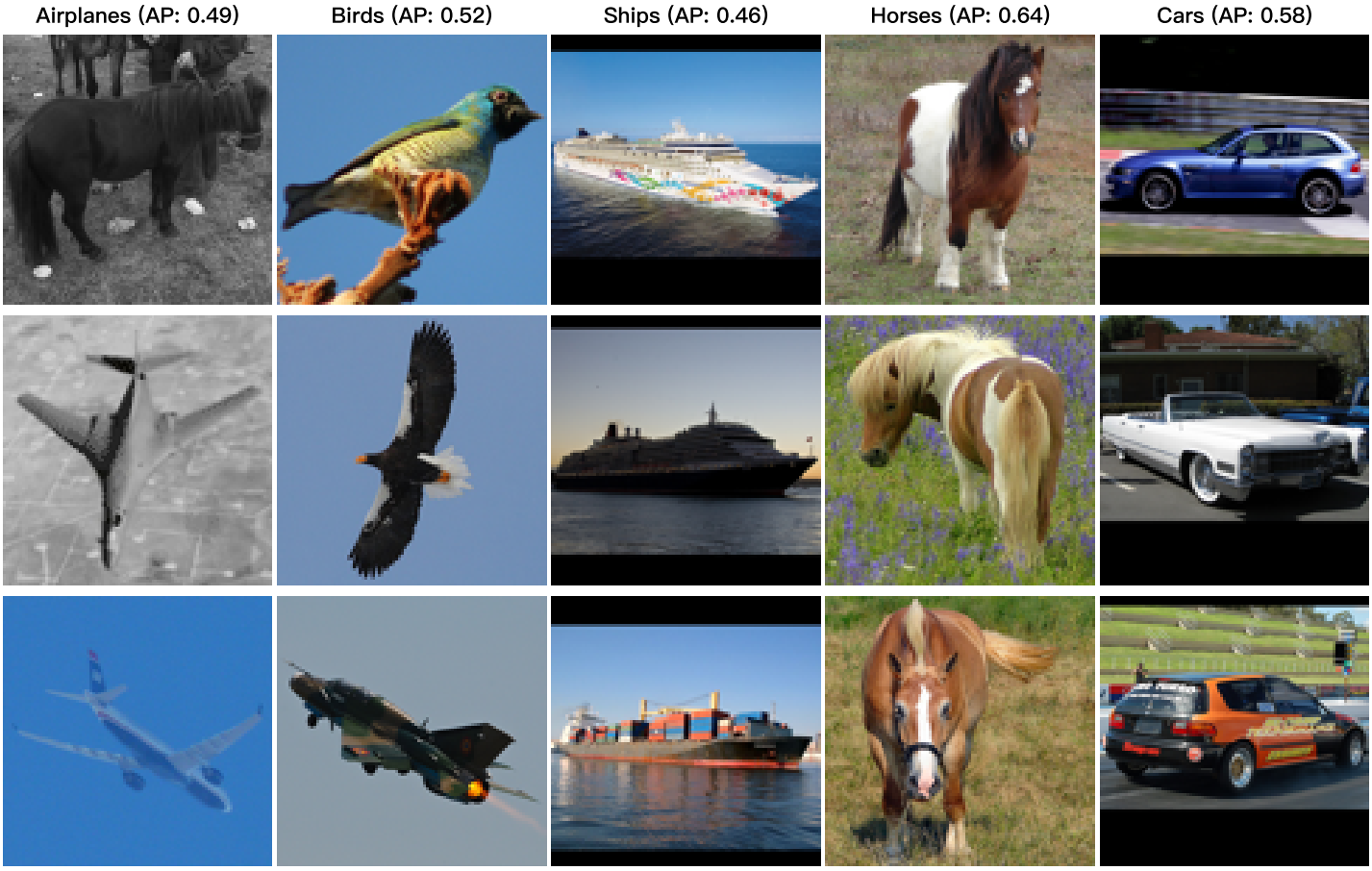


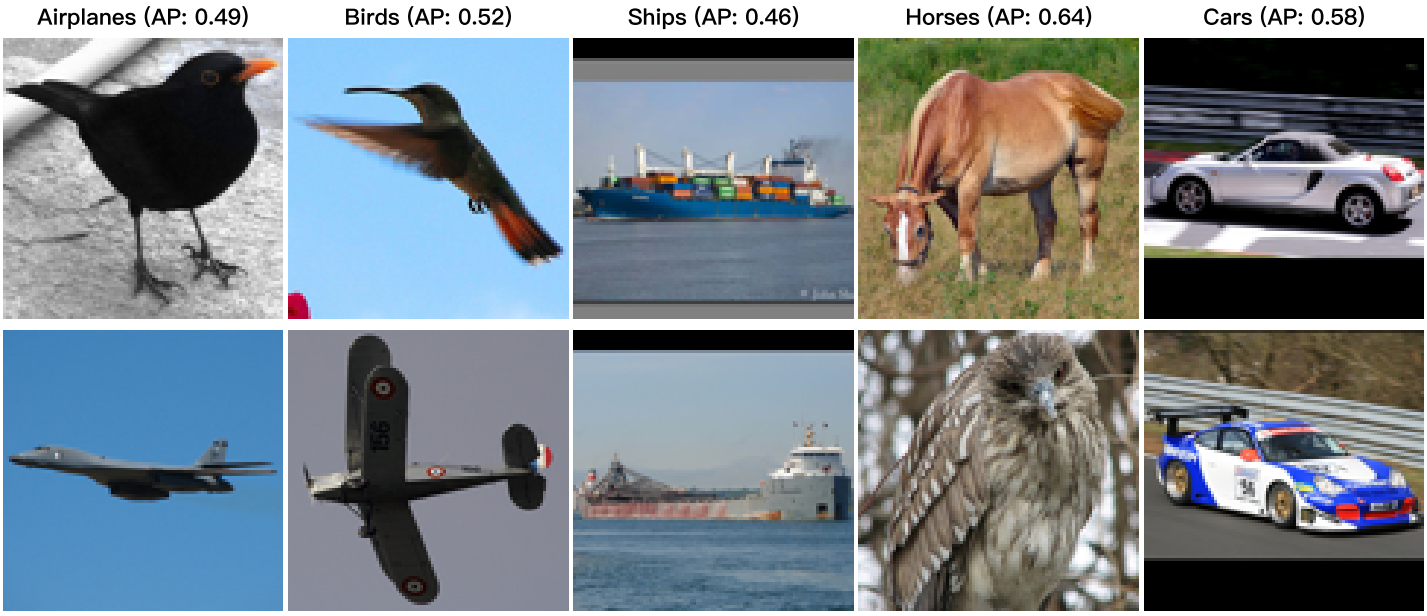
Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	400 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

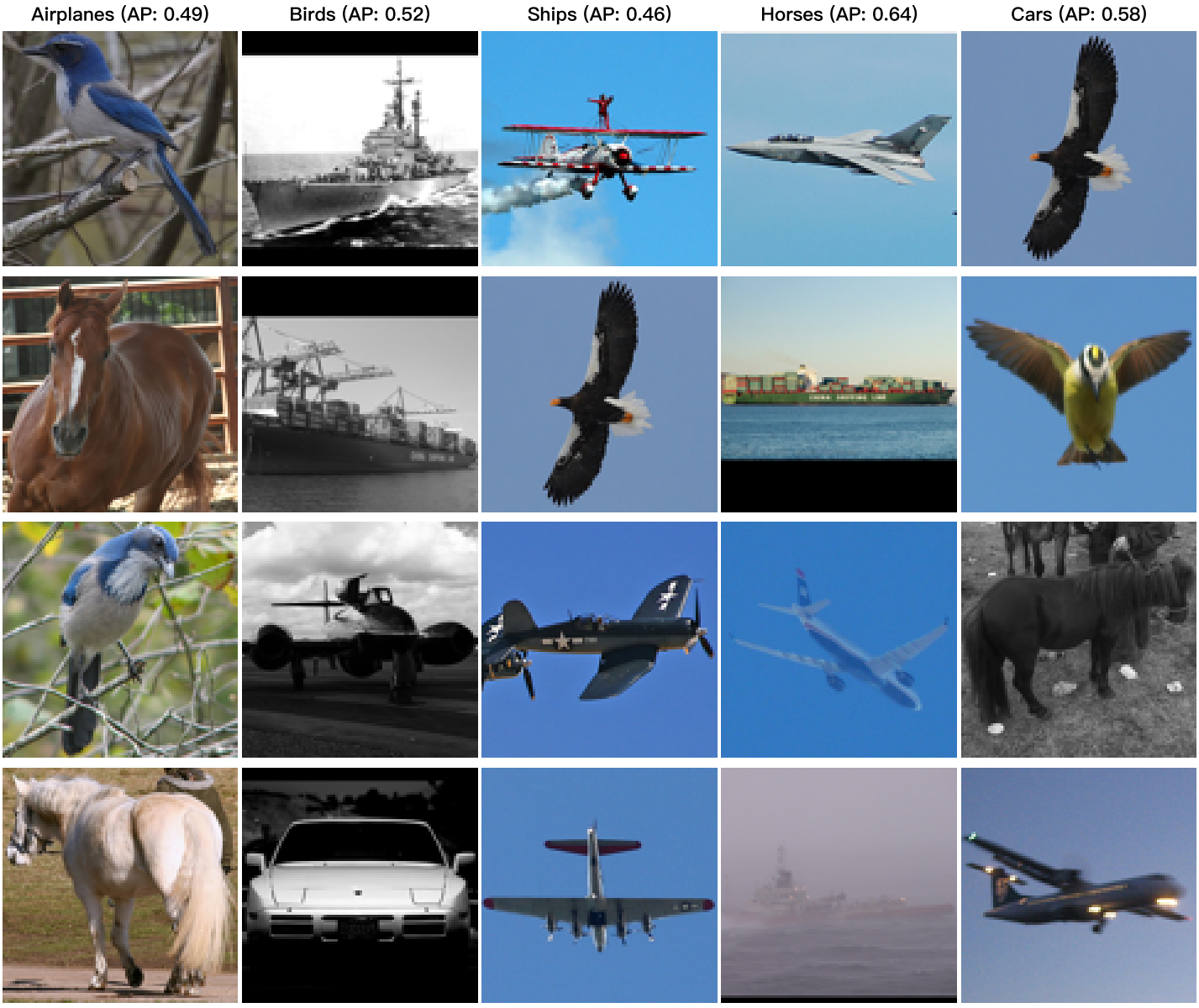
Prediction lists (MAP: 0.54)

Positive Order





Reverse order



Airplanes (AP: 0.49) | Birds (AP: 0.52) | Ships (AP: 0.46) | Horses (AP: 0.64) | Cars (AP: 0.58)

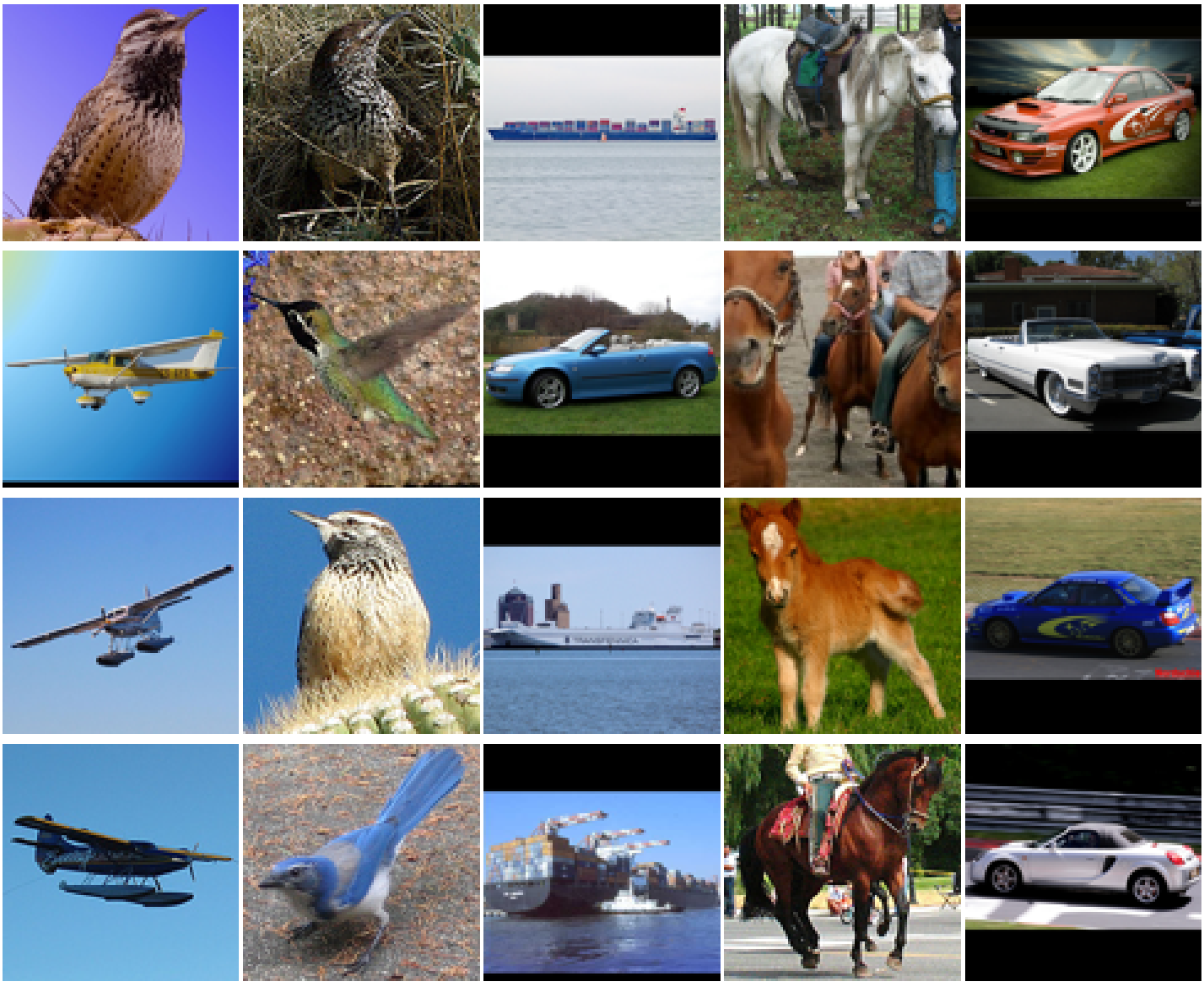
Settings

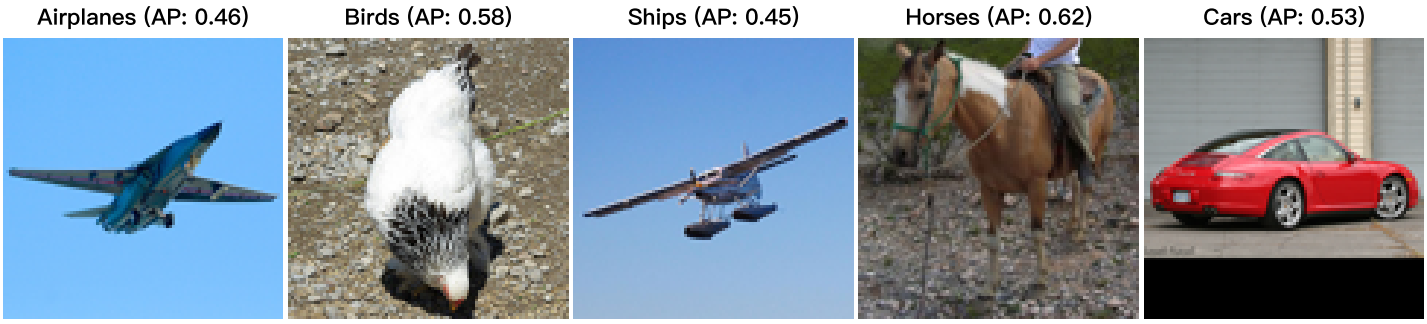
SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	1000 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.53)

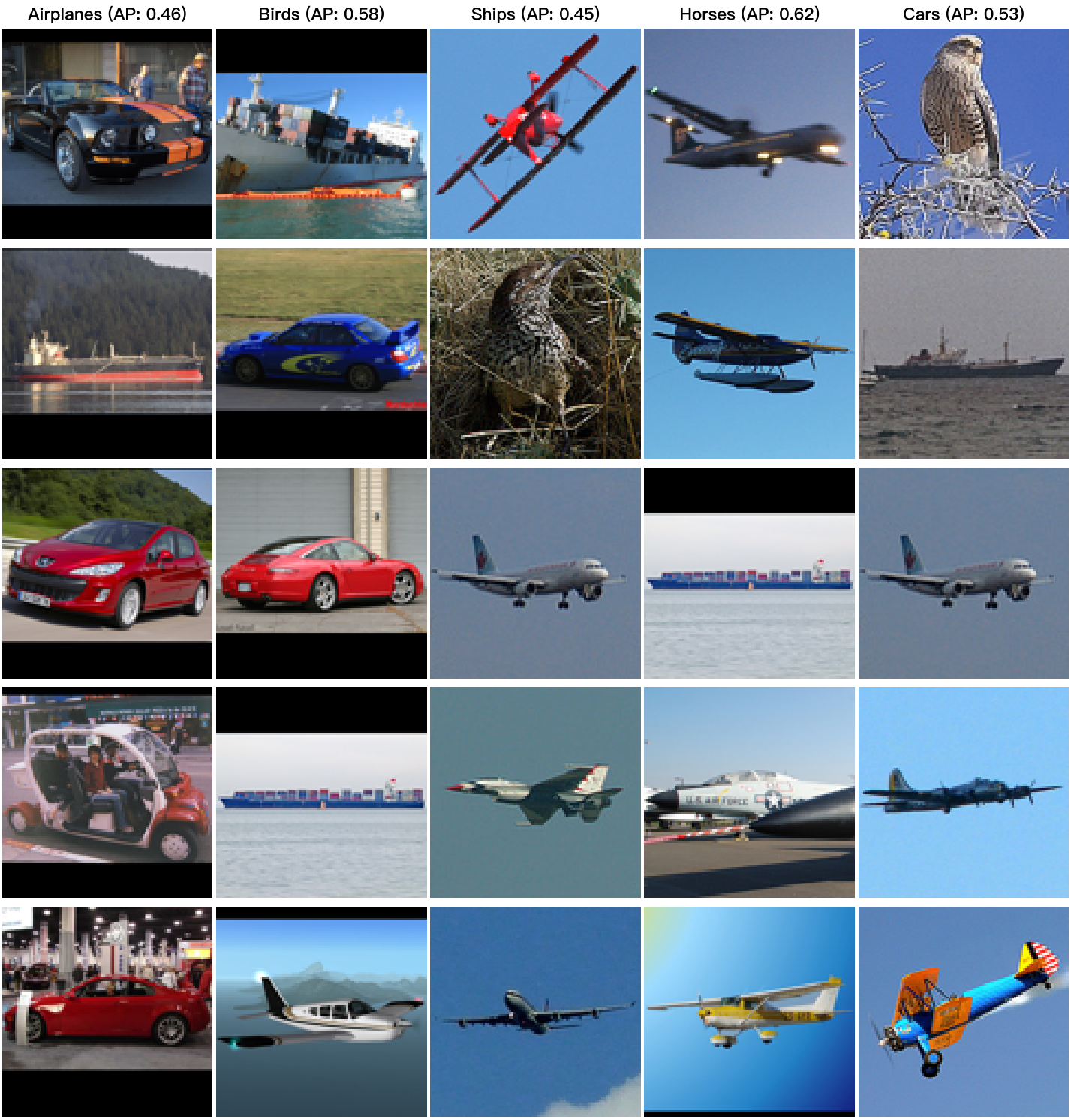
Positive Order

Airplanes (AP: 0.46) | Birds (AP: 0.58) | Ships (AP: 0.45) | Horses (AP: 0.62) | Cars (AP: 0.53)





Reverse order



Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	1000 words

color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

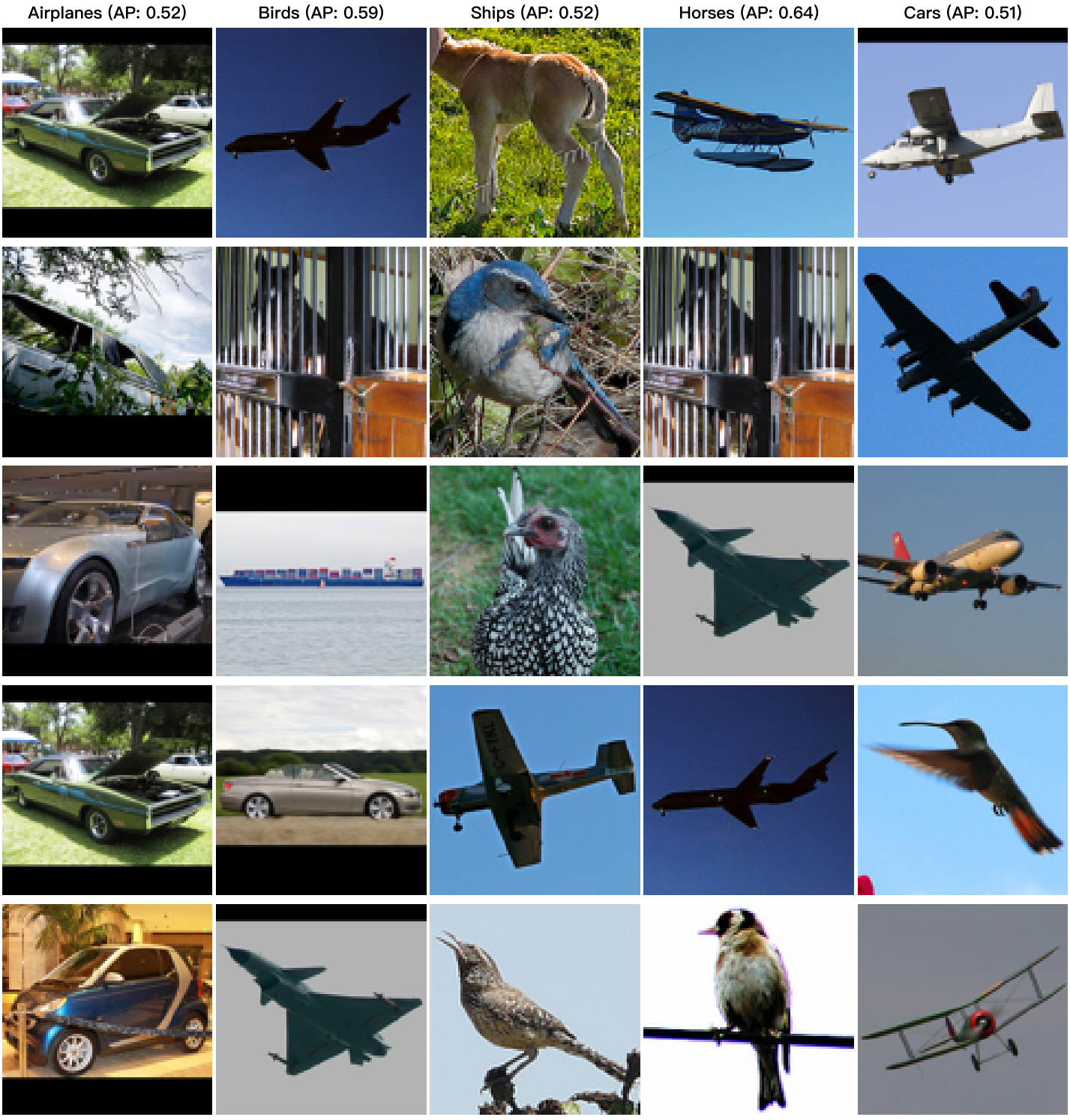
Prediction lists (MAP: 0.55)

Positive Order

Airplanes (AP: 0.52)	Birds (AP: 0.59)	Ships (AP: 0.52)	Horses (AP: 0.64)	Cars (AP: 0.51)

Reverse order

Airplanes (AP: 0.52)	Birds (AP: 0.59)	Ships (AP: 0.52)	Horses (AP: 0.64)	Cars (AP: 0.51)

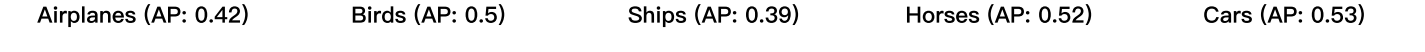


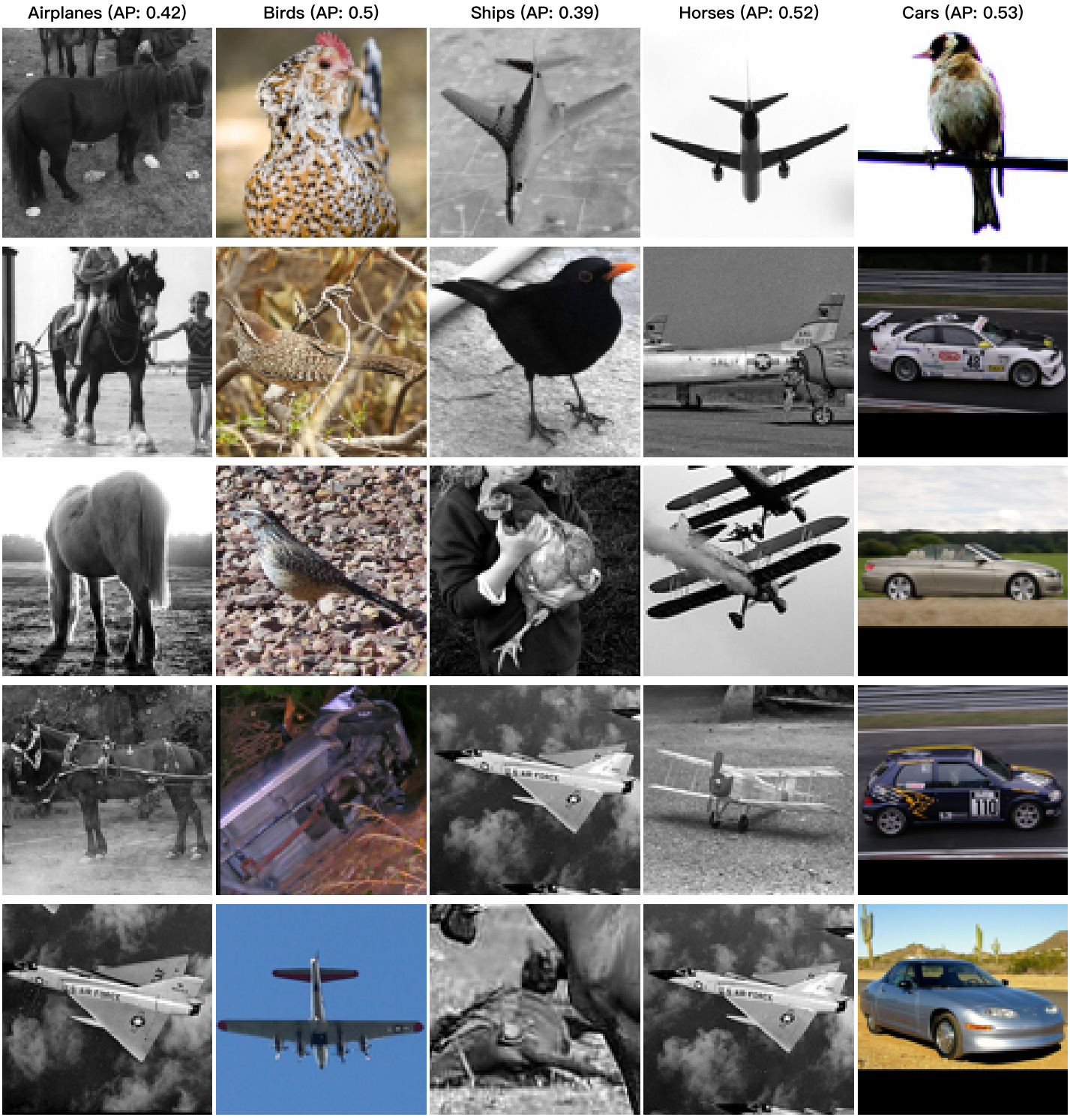
Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	1000 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

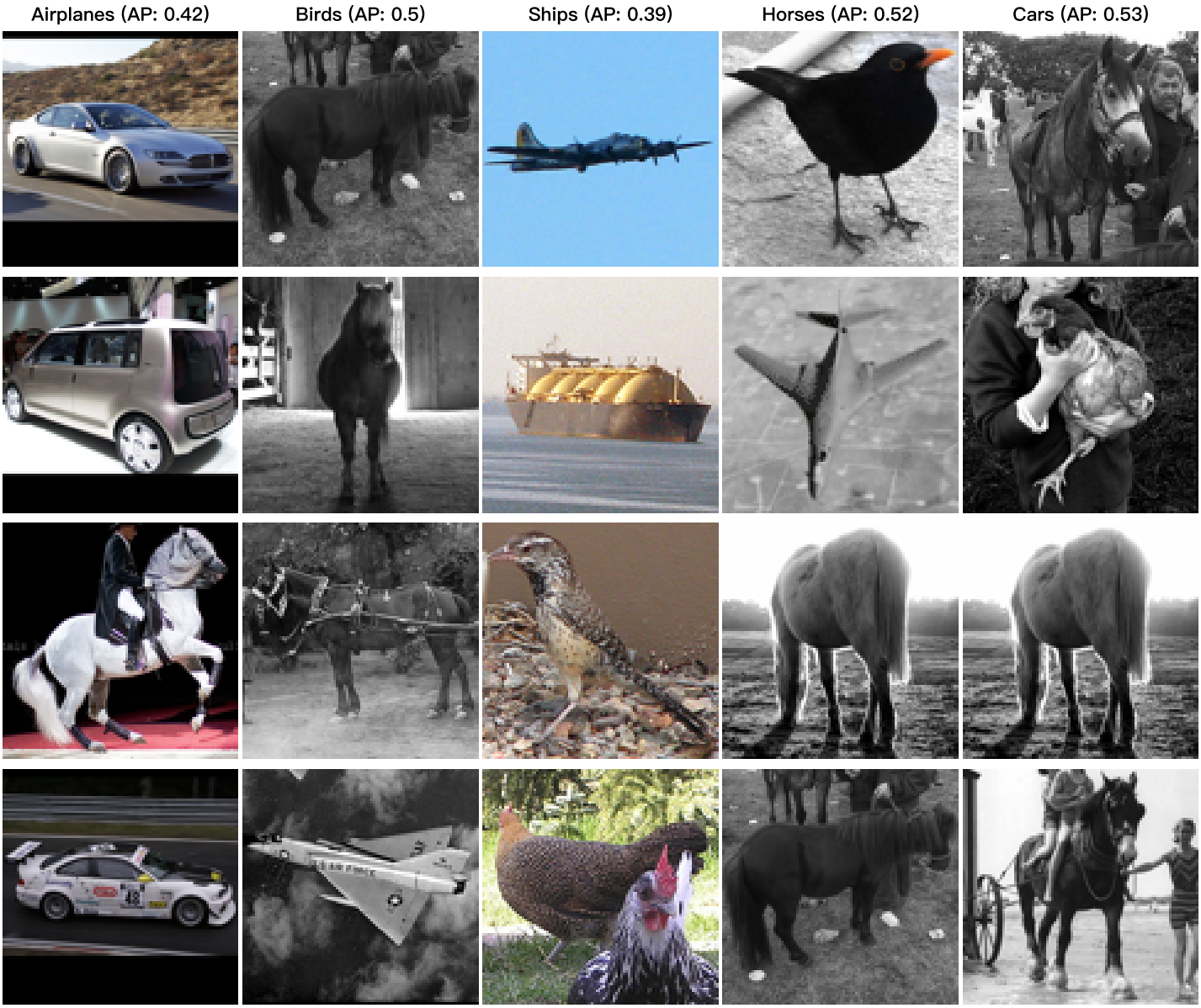
Prediction lists (MAP: 0.47)

Positive Order





Reverse order

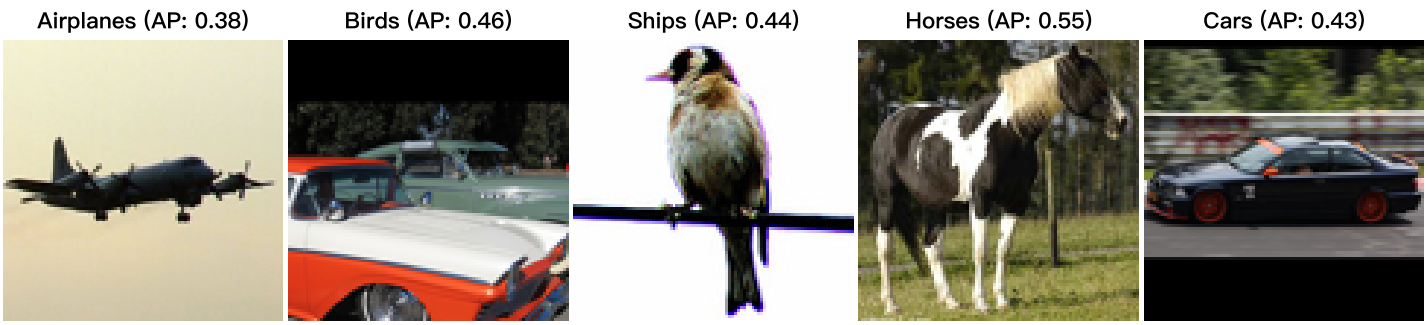


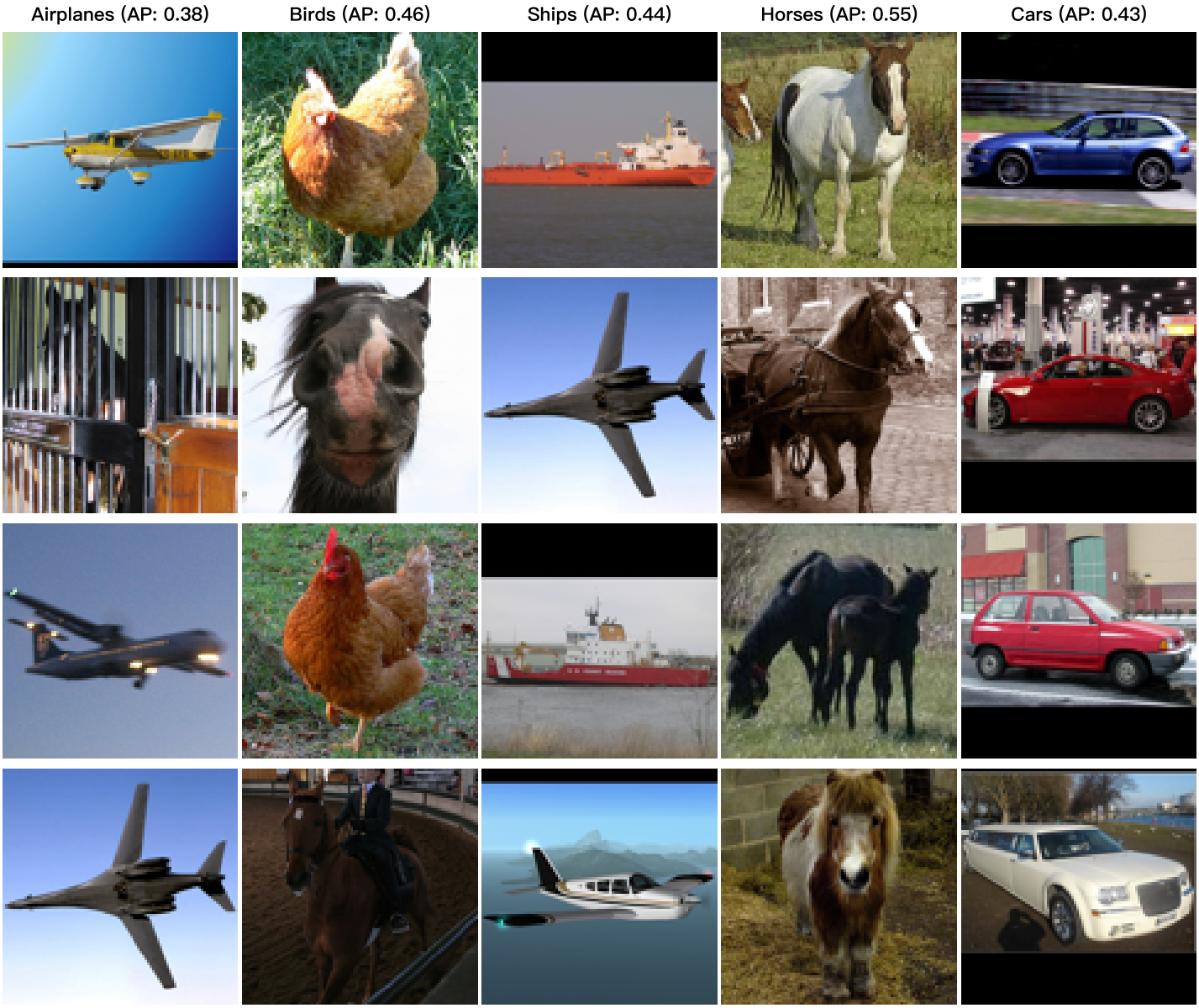
Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	4000 words
color space	grey
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

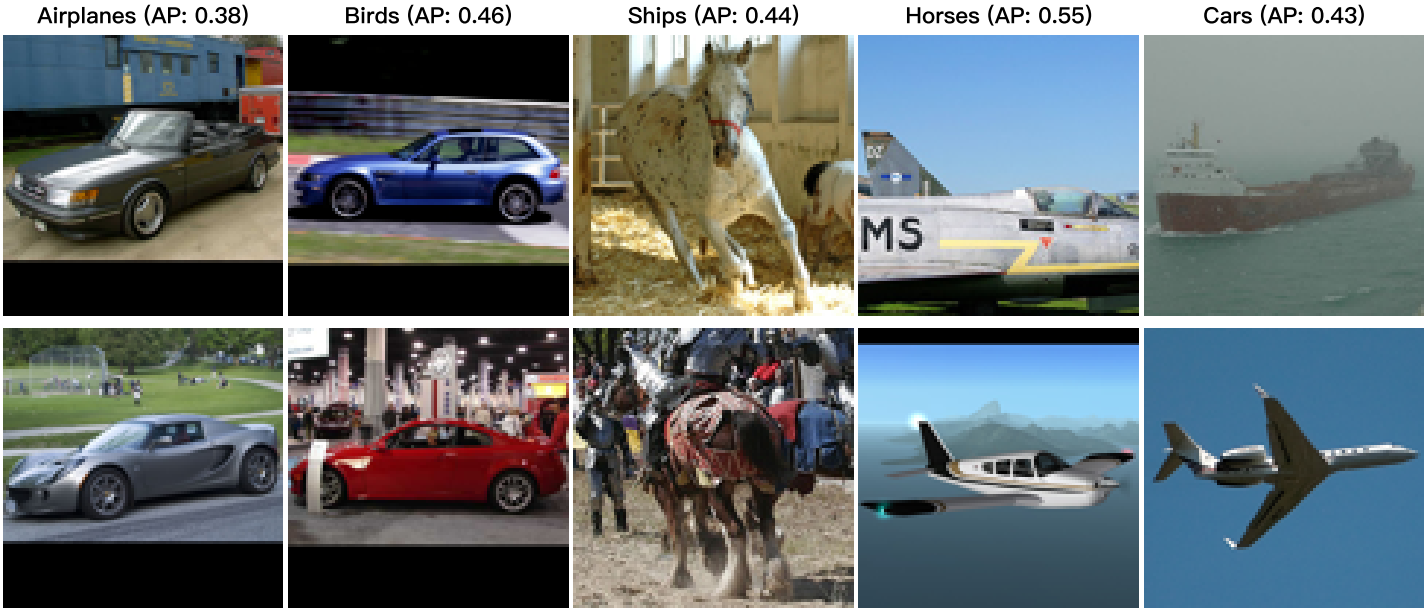
Prediction lists (MAP: 0.45)

Positive Order





Reverse order

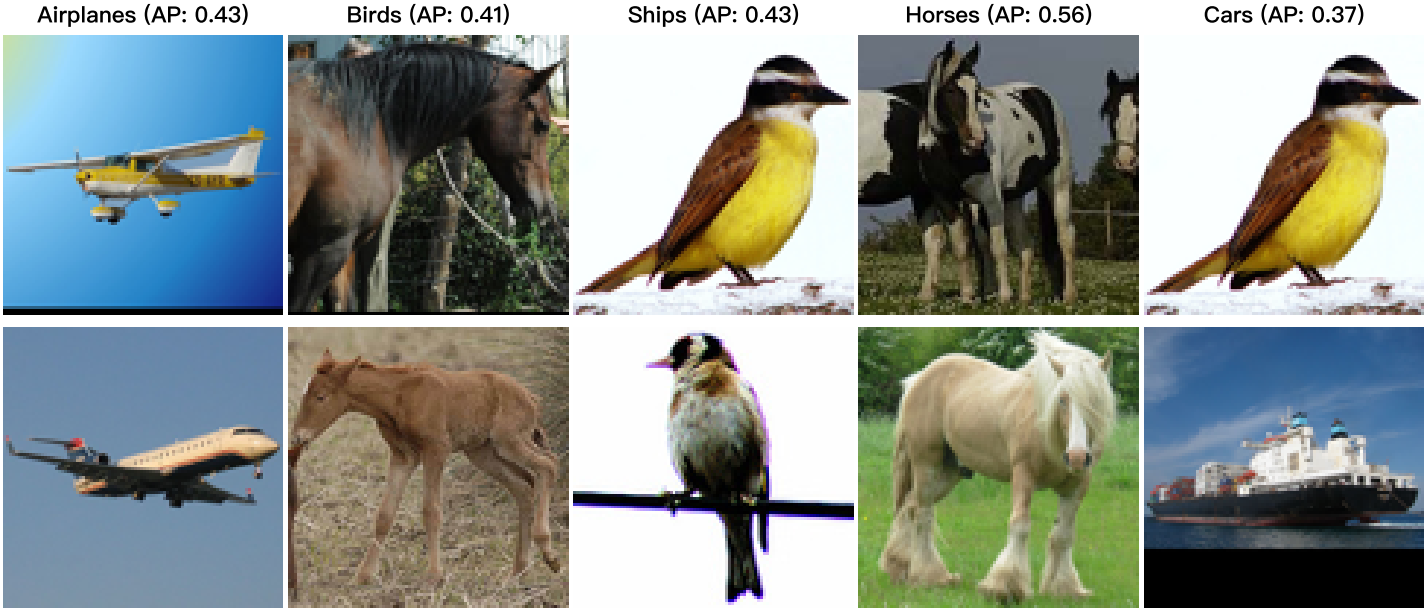


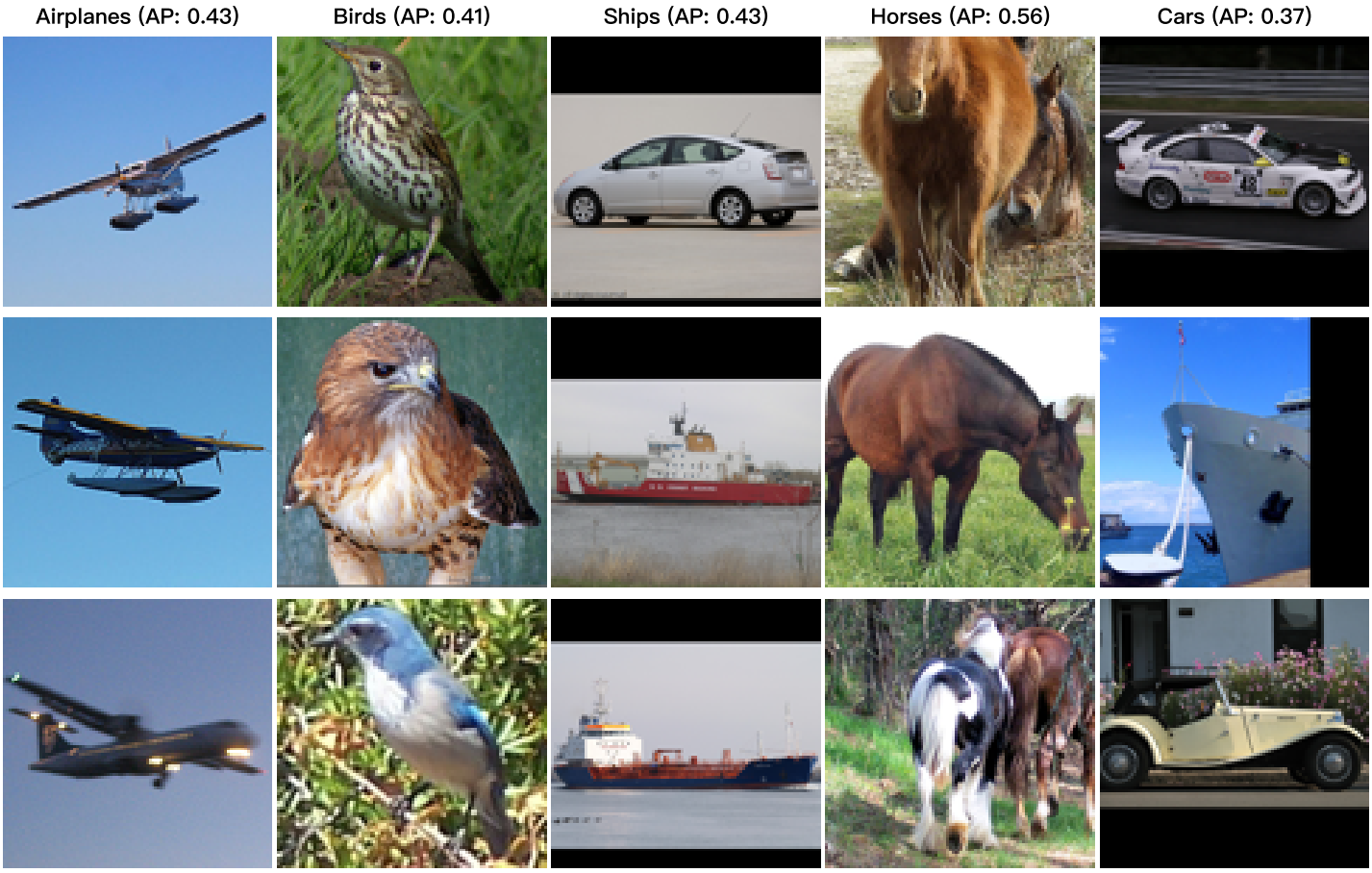
Settings

SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	4000 words
color space	rgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

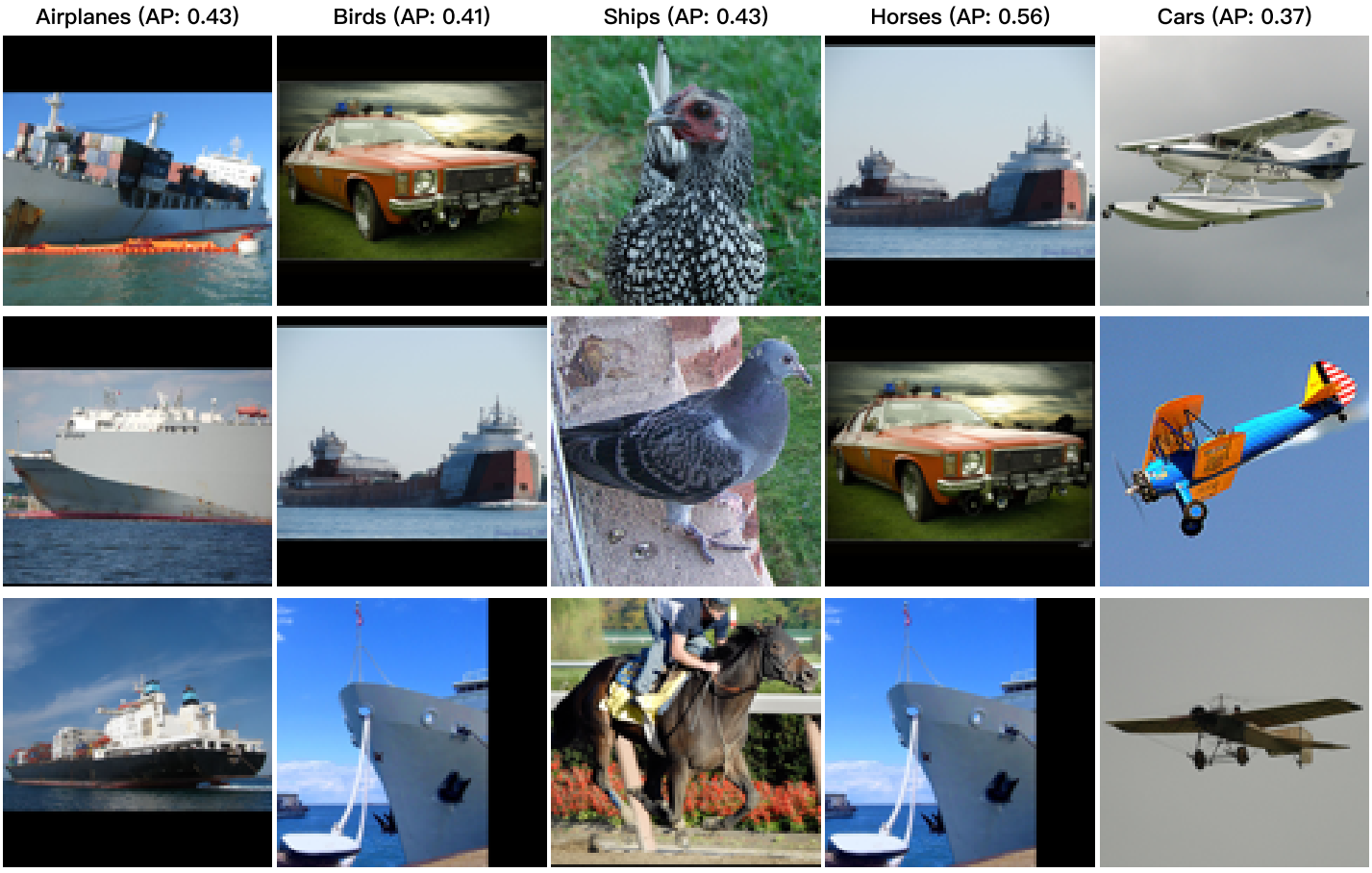
Prediction lists (MAP: 0.44)

Positive Order





Reverse order



Airplanes (AP: 0.43) | Birds (AP: 0.41) | Ships (AP: 0.43) | Horses (AP: 0.56) | Cars (AP: 0.37)

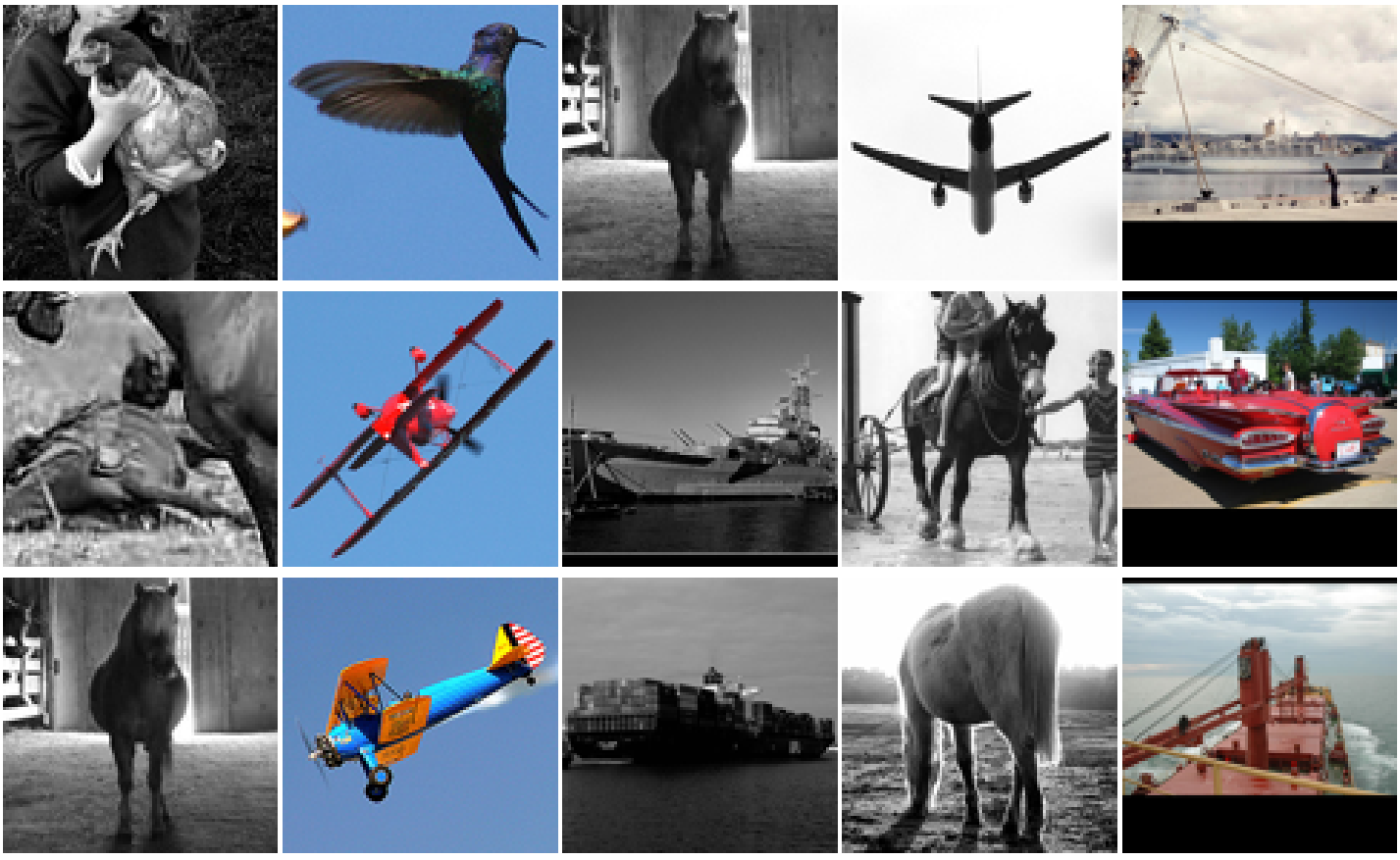
Settings

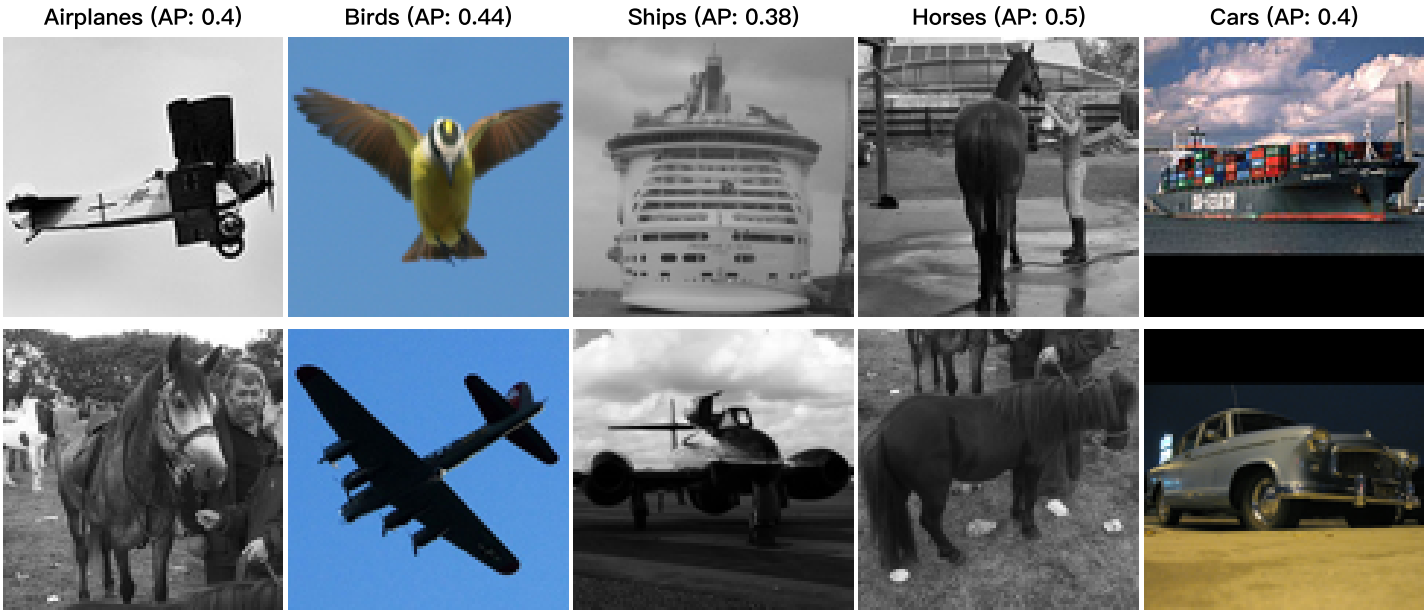
SIFT step size	5 px
SIFT sample method	dense
Vocabulary size	4000 words
color space	orgb
Vocabulary fraction	0.2
SVM training data	400 positive, 1600 negative per class
SVM kernel type	RBF

Prediction lists (MAP: 0.43)

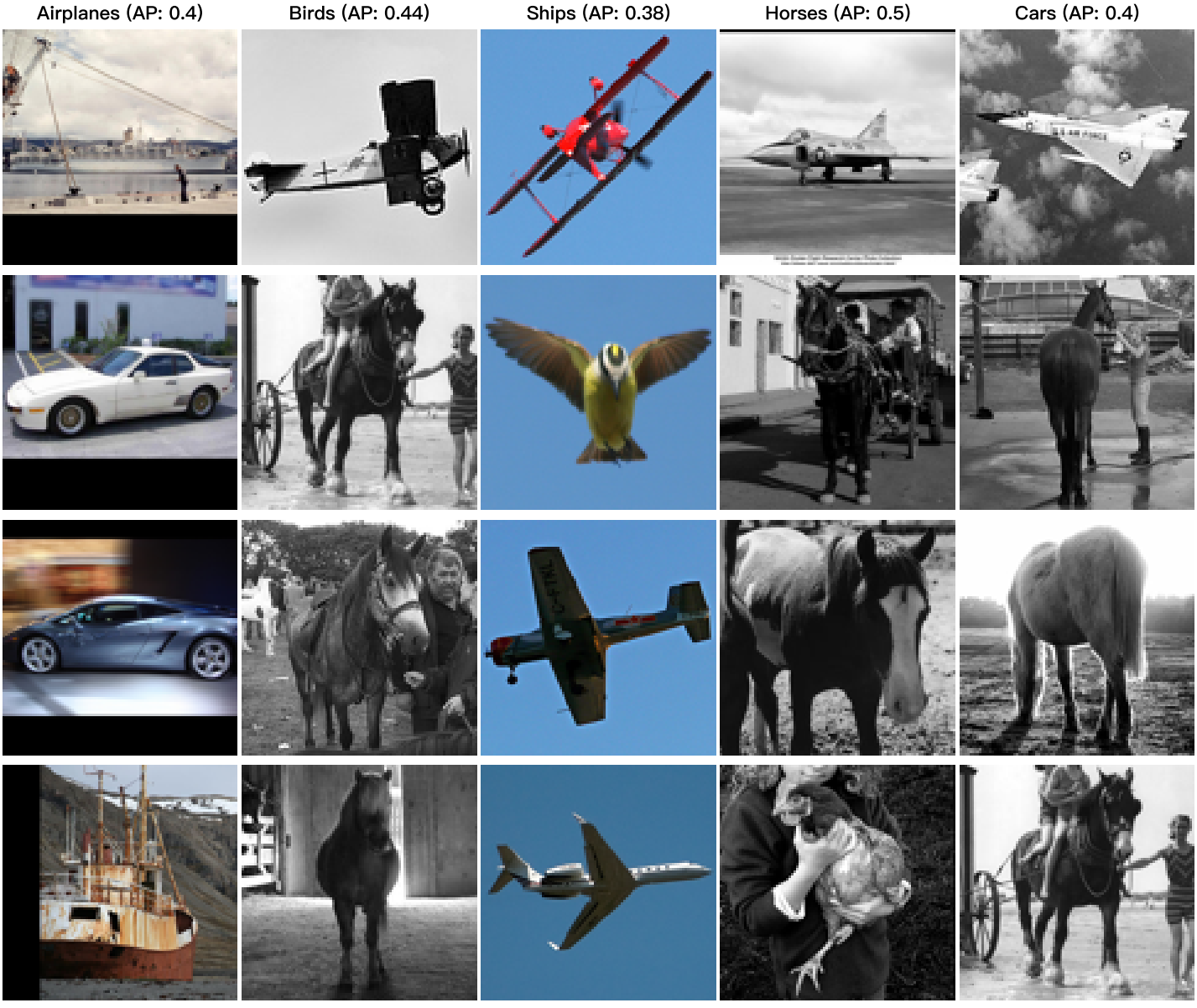
Positive Order

Airplanes (AP: 0.4) | Birds (AP: 0.44) | Ships (AP: 0.38) | Horses (AP: 0.5) | Cars (AP: 0.4)

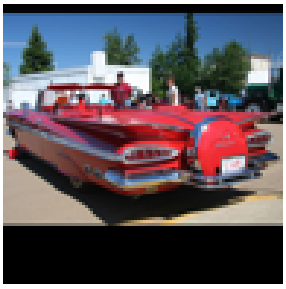




Reverse order



Airplanes (AP: 0.4)



Birds (AP: 0.44)

Ships (AP: 0.38)

Horses (AP: 0.5)

Cars (AP: 0.4)