

Абдрахманов Р.Е.^{1,2}, Алимова А.Н.^{1*}, Соколов С.В.¹,
Чалова Н.Р.¹, Нурсеитов Д.Б.¹

¹Институт информационных и вычислительных технологий, г. Алматы, Казахстан

²Казахский национальный исследовательский технический университет имени

К.И. Сатпаева, г. Алматы, Казахстан

*e-mail: anic2002@gmail.com

НАБОР ДАННЫХ И БЕНЧМАРК ДЛЯ ТОНКОЗЕРНИСТОЙ КЛАССИФИКАЦИИ ОХОТНИЧЬИХ ВОДОПЛАВАЮЩИХ ПТИЦ

Аннотация

Идентификация птиц по виду и полу имеет важное значение для учета добычи, мониторинга состояния охотничьих ресурсов и соблюдения установленных правил охоты. Существующие наборы данных и цифровые приложения преимущественно определяют вид, не учитывая половую принадлежность особей, что ограничивает их применение в охотничьем хозяйстве. В работе представлен специализированный набор данных GameWaterbird-41 для совместной классификации вида и пола охотничьих водоплавающих птиц и куликов, являющихся объектами охоты. Набор данных сформирован на основе материалов iNaturalist и включает 18 851 изображение 38 видов, распределенных по 41 классу. Проведен бенчмарк сверточных нейронных сетей и визуальных трансформеров в трех режимах: без дополнительного обучения, с обучением линейного классификатора и с полной тонкой настройкой. Рассмотрены модели общего назначения и модели, предварительно обученные на биологических изображениях. Наиболее высокое качество классификации получено при полной тонкой настройке на целевых данных. Набор данных и результаты бенчмарка могут быть использованы для совершенствования учета добычи и информационной поддержки управления охотничьими ресурсами.

Ключевые слова: тонкозернистое визуальное распознавание, классификация видов птиц, половой диморфизм, водоплавающие птицы, трансферное обучение, температурное масштабирование, аугментация при тестировании.

Абдрахманов Р.Е.^{1,2}, Алимова А.Н.^{1*}, Соколов С.В.¹, Чалова Н.Р.¹, Нурсеитов Д.Б.¹

¹Ақпараттық және есептеу технологиялары институты, Алматы қ., Қазақстан

²Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы қ., Қазақстан

АҢШЫЛЫҚ СУ ҚҰСТАРЫНЫҢ ТҮРЛЕРІН ҰСАҚ ТҮЙІРШІКТІ ЖІКТЕУГЕ АРНАЛҒАН ДЕРЕКТЕР ЖИЫНТЫҒЫ ЖӘНЕ БЕНЧМАРК

Аңдатпа

Құстарды түрі мен жынысына қарай сәйкестендіру ауланған құстардың есебін жүргізу, аңшылық ресурстардың жағдайын мониторингтеу және белгіленген аңшылық ережелерінің сақталуын қамтамасыз ету үшін маңызды. Қолданыстағы деректер жиынтықтары мен цифрлық қосымшалар негізінен құстың түрін ғана анықтауға бағытталған, ал жеке құстардың жыныстық ерекшелігін ескермейді. Бұл олардың аңшылық шаруашылығында қолданылу мүмкіндігін шектейді. Бұл жұмыста аңшылық нысандары болып табылатын су құстары мен шалшықшылардың түрі мен жынысын бірге жіктеуге арналған GameWaterbird-41 мамандандырылған деректер жиынтығы ұсынылады. Деректер жиынтығы iNaturalist материалдары негізінде құрастырылған және 41 класқа бөлінген 38 түрге жататын 18 851 кескінді қамтиды. Конволюциялық нейрондық желілер мен визуалды трансформерлерге үш режимде: қосымша оқытусыз, сызықтық классификаторды оқыту арқылы және модельді толықтай дәлдеп оқыту арқылы бенчмарк-бағалау жүргізілді. Жалпы мақсаттағы модельдер және биологиялық кескіндер негізінде алдын ала оқытылған модельдер қарастырылды. Жіктеудің ең жоғары сапа көрсеткіштеріне мақсатты деректер жиынтығында модельді толықтай дәлдеп оқыту барысында қол жеткізілді. Деректер жиынтығы мен бенчмарк нәтижелері ауланған жануарлардың

есебін жүргізу жүйесін жетілдіру және аншылық ресурстарды басқаруды ақпараттық қолдау мақсатында пайдаланылуы мүмкін.

Түйін сөздер: егжей-тегжейлі визуалды тану, құс түрлерін жіктеу, жыныстық диморфизм, су құстары, трансферлік оқыту, температуралық масштабтау, тестілеу кезіндегі аугментация.

R. Abdrakhmanov^{1,2}, A. Alimova^{1*}, S. Sokolov¹, N. Chalova¹, D. Nurseitov¹

¹The Institute of Information and Computational Technologies, Almaty, Kazakhstan

²K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan

DATASET AND BENCHMARK FOR FINE-GRAINED CLASSIFICATION OF HUNTING WATERFOWL

Abstract

Identification of birds by species and sex is important for harvest accounting, monitoring of game resource state, and compliance with established hunting regulations. Existing datasets and digital applications predominantly identify species without accounting for the sex of individuals, which limits their applicability in game management. This paper presents a specialized dataset, GameWaterbird-41, for combined classification of species and sex of hunting waterfowl and waders that are game species. The dataset was composed based on materials from iNaturalist and includes 18,851 images of 38 species distributed across 41 classes. A benchmark of convolutional neural networks and vision transformers was conducted in such classification modes as zero-shot classification mode, classification mode with linear classifier training, and classification mode with full fine-tuning. General-purpose models and models pre-trained on biological images were considered. The highest classification quality was obtained with full fine-tuning on the target data. The dataset and benchmark results can be used to improve game harvest accounting and information support for game resource management.

Keywords: fine-grained visual recognition, bird species classification, sexual dimorphism, waterfowl, transfer learning, temperature scaling, test-time augmentation.

Введение

Идентификация вида и пола птицы по фотографии является классической задачей тонкозернистого визуального распознавания [1, 2]. В контексте охотничьего хозяйства данная задача имеет практическое значение, поскольку нормативные документы, регулирующие охоту, определяют не только перечень видов, разрешенных к добыче, но и половую принадлежность особей. Для многих видов уток характерен половой диморфизм. Например, самцы и самки кардинально различаются по окраске, что позволяет использовать их визуальные признаки для идентификации. У других видов, в частности у гусей, оба пола внешне неразличимы, однако сам вид может быть разрешен или запрещен к добыче. Надежный автоматический классификатор способен помочь охотникам правильно определять птицу в полевых условиях, а субъектам охотничьего хозяйства масштабировать мониторинг популяций на основе научных данных.

В настоящее время учет добываемых водоплавающих птиц в охотничьем хозяйстве преимущественно ведется по укрупненным таксономическим группам (утки, кулики) [3]. При этом нормативные документы регламентируют добычу конкретных видов, а в ряде случаев учитывают и половую принадлежность особей [4, 5]. В связи с этим разработка автоматизированных методов идентификации вида и пола птиц по фотографиям представляет практический интерес для повышения качества учета добычи, мониторинга охотничьих ресурсов и информационной поддержки принятия управленческих решений в охотничьем хозяйстве.

Существующие крупномасштабные наборы данных для распознавания птиц, такие как CUB-200-2011 [1], NABirds [2], Birdsnap [6] и наборы данных iNaturalist [7], ориентированы преимущественно на идентификацию вида среди сотен и тысяч классов. При этом CUB-200-2011, NABirds и Birdsnap сформированы главным образом на основе орнитофауны Северной Америки. Существующие цифровые приложения для идентификации птиц, включая Merlin Bird ID и iNaturalist, имеют более широкий географический охват, однако не ориентированы

специально на задачи учета охотничьих видов Казахстана и, как правило, не предусматривают определение пола особи. В отличие от них, разработанный в настоящем исследовании набор данных ориентирован на виды водоплавающих птиц и куликов, являющиеся объектами охоты и обитающие на территории Республики Казахстан или мигрирующих через нее, а для видов с выраженным половым диморфизмом предусматривает классификацию с учетом пола особи.

Предварительно обученные (фундаментальные) модели, такие как CLIP [8] и его биологическая модификация BioCLIP [9], показывают высокую точность при распознавании вида без дополнительного обучения (zero-shot). Такие модели предварительно обучаются на больших наборах изображений общего назначения и затем могут быть адаптированы для решения новых задач. Однако вопрос о том, способны ли такие модели решать более сложную задачу совместной классификации вида и пола без доменно-специфической тонкой настройки, остается открытым.

Несмотря на существование крупных наборов данных и современных моделей компьютерного зрения, специализированные наборы данных, ориентированные на совместную классификацию вида и пола охотничьих водоплавающих птиц, в настоящее время отсутствуют. Кроме того, недостаточно изучены возможности современных фундаментальных моделей при решении данной задачи. Целью настоящего исследования является разработка специализированного набора данных GameWaterbird-41 и проведение сравнительного анализа современных моделей компьютерного зрения для совместной классификации вида и пола охотничьих водоплавающих птиц как основы для создания цифрового определителя птиц.

Разработанный набор данных и результаты сравнительного анализа моделей могут быть использованы при создании цифровых систем поддержки охотничьего хозяйства, обеспечивающих автоматическое определение вида и пола птиц по фотографиям, полученным в полевых условиях. Использование подобных систем позволит повысить качество учета добычи, обеспечить получение объективной информации о видовом составе охотничьих ресурсов и расширить возможности их мониторинга.

Методология исследования

Тонкозернистое распознавание птиц. Набор данных CUB-200-2011 [1] является одним из используемых бенчмарков для тонкозернистого распознавания птиц по морфологическим признакам на уровне вида. Он включает 11 788 изображений 200 видов птиц Северной Америки. Набор данных NABirds [2] расширяет данный подход до 555 визуальных категорий, включая различия по полу и возрасту и содержит 48 000 изображений, собранных пользователями. Birdsnap [6] охватывает 500 видов птиц Северной Америки. Платформа iNaturalist [7] расширила область применения методов тонкозернистого распознавания до тысяч таксономических классов различных групп живых организмов. В отличие от перечисленных наборов данных, разработанный в настоящем исследовании набор данных включает значительно меньшее число видов и изображений, но ориентирован на решение специализированной регуляторно-экологической задачи классификации охотничьих водоплавающих птиц и куликов, являющихся объектами охоты, и дополнительно содержит информацию о половой принадлежности особей.

Фундаментальные модели для распознавания объектов дикой природы. Под моделью в работе понимается обученная на данных программа, которая по изображению оценивает вероятность принадлежности объекта к каждому из классов. Модель CLIP [8] показала, что визуально-языковое предобучение на 400 млн пар «изображение–текст» обеспечивает высокую эффективность классификации без дополнительного обучения на целевом наборе данных с использованием текстовых описаний классов. BioCLIP [9] представляет собой специализированную модификацию CLIP, адаптированную для биологических задач. Модель предварительно обучена на наборе данных TreeOfLife-10M, охватывающем 450 000 таксонов, и достигла высокой точности при распознавании биологических объектов. Модель

DINOv2 [10] основана на самонадзорном обучении и использует набор данных из 142 млн изображений без применения текстовых аннотаций. Получаемые признаки дают высокую эффективность по сравнению с другими подходами при решении задач тонкозернистой классификации как при использовании линейного классификатора, так и после тонкой настройки модели.

Визуальные трансформеры. Архитектура Vision Transformer (ViT) [11] показала, что трансформерные модели при масштабном предварительном обучении могут превосходить сверточные нейронные сети при анализе изображений. В работе [12] исследовано влияние аугментации данных и методов регуляризации на качество обучения моделей семейства ViT (AugReg). Аугментация данных представляет собой искусственное расширение набора данных за счет преобразований изображений, таких как сдвиг, отражение, поворот и изменение яркости. Полученные авторами контрольные точки ViT-B/16 показали, что модель хорошо обобщается при тонкой настройке.

Калибровка модели и ТТА. Модель называют калиброванной, если заявленный ею уровень уверенности соответствует фактической доле правильных предсказаний. Например, среди предсказаний с вероятностью 90% правильными должны быть примерно девять из десяти. В исследовании [13] показано, что один обучаемый параметр температуры, применяемый к логитам, то есть выходным числовым значениям модели до их преобразования в вероятности, эффективно калибрует апостериорные вероятности без потери точности классификации. Такой подход называется температурным масштабированием и заключается в корректировке вероятностей нейронной сети с помощью параметра температуры, подбираемого на валидационных данных без переобучения самой модели. Направление ошибки калибровки зависит от используемой функции потерь. Модели, обученные со сглаживанием меток и функцией потерь Focal Loss [14], как правило, характеризуются заниженной уверенностью, при которой вероятности распределяются по классам более равномерно, чем это соответствует фактической достоверности предсказаний, и требуют применения температуры $T < 1$ для корректировки распределения вероятностей. Именно это наблюдается у всех дообученных моделей в настоящем исследовании, оптимальные значения температуры находятся в диапазоне от 0,49 до 0,69. Аугментация при тестировании (ТТА) представляет собой применение преобразований изображения не на этапе обучения, а на этапе распознавания. При этом для одного исходного изображения формируется несколько аугментированных вариантов, а полученные для них предсказания усредняются. Такой подход снижает дисперсию предсказаний и повышает их устойчивость к изменениям изображения.

Классификация пола и вида птиц. Автоматическое определение пола птиц привлекло значительно меньше внимания, чем идентификация вида. Большинство исследователей либо рассматривают его как дополнительную задачу предсказания атрибута [2], либо ограничиваются мономорфными видами, исключая необходимость определения пола. Разработанный в настоящем исследовании набор данных включает пол в качестве основной цели классификации для диморфных видов, что делает его одним из немногих бенчмарков, ориентированных на совместную классификацию вида и пола птиц.

Набор данных GameWaterbird-41. Каталог видов сформирован на основе официального перечня животных, являющихся объектами регулируемой охоты в Республике Казахстан [4]. В него включены все водоплавающие и болотно-луговые птицы, являющиеся объектами охоты: речные и нырковые утки (семейство Anatidae), гуси (Anser spp.), крохали, пеганки и кулики (кроншнепы, веретенники, вальдшнеп и родственные виды). Всего в собранном наборе данных включено 38 видов. С точки зрения задачи классификации эти виды делятся на две группы:

– Полово-диморфные виды (25 видов). Самцы и самки имеют выраженные различия во внешнем облике и имеют разное оперение. В эту группу входят большинство речных уток (кряква, шилохвость, чирки, свиязь, широконоска и др.) и нырковых уток (чернети, нырки,

гоголь, морянка и др.). Для этих видов определение пола по фотографии возможно и имеет регуляторное значение.

– Мономорфные виды (13 видов). Самцы и самки практически не отличаются по внешнему виду и имеют одинаковое оперение. В эту группу входят все гуси рода *Anser*, а также вальдшнеп, кроншнепы и веретенники. Для этих видов классификация выполняется только на уровне вида.

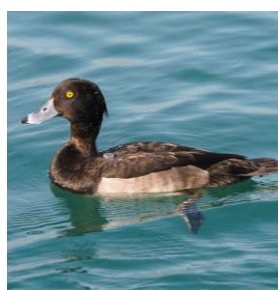
Изображения получены из базы данных iNaturalist [15] с ограничением по уровню качества Research Grade (верифицированные наблюдения) и открытых лицензий (CC0, CC-BY, CC-BY-SA, CC-BY-NC, CC-BY-NC-SA). На рисунке 1 первые два изображения демонстрируют выраженный половой диморфизм у кряквы. Самец кряквы имеет яркое переливчатое оперение, тогда как для самки характерна бурая покровительственная окраска. Изображение белолобого гуся демонстрирует типичную проблему межвидового сходства среди мономорфных видов, затрудняющую их автоматическую классификацию. Загруженные изображения были обработаны детектором объектов YOLOv11x [16], предварительно обученным на наборе данных COCO.



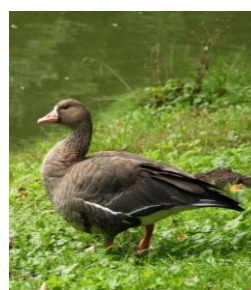
Кряква ♂
Anas platyrhynchos



Кряква ♀
Anas platyrhynchos



Хохл. чернеть ♂
Aythya fuligula



Белолобый гусь
Anser albifrons



Большой кроншнеп
Numenius arquata

Рисунок 1. Примеры изображений из GameWaterbird-41, полученных из iNaturalist

В итоговый набор данных включались только изображения, которые удовлетворяли следующим критериям:

- на изображении присутствует только одна птица;
- ограничивающий прямоугольник птицы занимает не менее 20% площади изображения.

Применение этих критериев позволило исключить изображения с несколькими особями, удаленными объектами или частично перекрытыми птицами. Координаты ограничивающего прямоугольника сохраняются для каждого изображения и в дальнейшем используются на этапе инференса для автоматической обрезки области птицы.

Платформа iNaturalist позволяет пользователям указывать половую принадлежность особи («самец», «самка» или «не определен»). Для 25 диморфных видов в набор данных включаются только изображения с подтвержденной меткой пола «самец» или «самка». Изображения с меткой «не определен» или без указания пола исключались с целью снижения уровня шума в аннотациях. Для 13 мономорфных видов в набор данных включались все изображения, удовлетворяющие критериям качества, независимо от наличия информации о поле.

После сбора и фильтрации виды или половые группы, содержащие менее 50 изображений, объединялись в единый класс «прочие». В результате была сформирована следующая структура классов:

- 14 из 25 диморфных видов содержат не менее 50 изображений как самцов, так и самок. Для каждого такого вида сформировано два отдельных класса, что в сумме составляет 28 классов (14×2);

– 12 из 13 мономорфных видов содержат не менее 50 изображений и представлены 12 отдельными классами;

– оставшиеся 12 видов (11 диморфных и 1 мономорфный), не удовлетворяющие минимальному порогу по числу изображений, объединены в один класс «прочие».

Итоговая таксономия набора данных содержит 41 класс (28 + 12 + 1), а общий объем набора данных составляет 18 851 изображение. Размер отдельных классов варьируется от 50 до 1 010 изображений, что отражает неравномерную представленность различных видов в базе данных iNaturalist. Набор данных был разделен на обучающую (75%), валидационную (15%) и тестовую (10%) выборки методом стратифицированной выборки по классам. В результате обучающая выборка содержит 14 182 изображения, валидационная 2 804 изображения, а тестовая 1 865 изображений. Сводные характеристики набора данных представлены в таблице 1, а распределение изображений по классам набора данных GameWaterbird-41 представлено на рисунке 2.

Таблица 1. Сводные характеристики набора данных GameWaterbird-41

Характеристика	Значение
Всего изображений	18 851
Всего видов	38
Полово-диморфные	25
Мономорфные	13
Всего классов	41
Половые классы (14 × 2)	28
Классы мономорфных видов	12
Прочие	1
Изображений в классе (мин / макс)	50 / 1 010
Обучение / Валидация / Тест	14 182 / 2 804 / 1 865
Источник изображений	iNaturalist (глобально)
Детектор	YOLOv11x, bbox ≥ 20% кадра

GameWaterbird-41 — распределение по классам (41 класс)

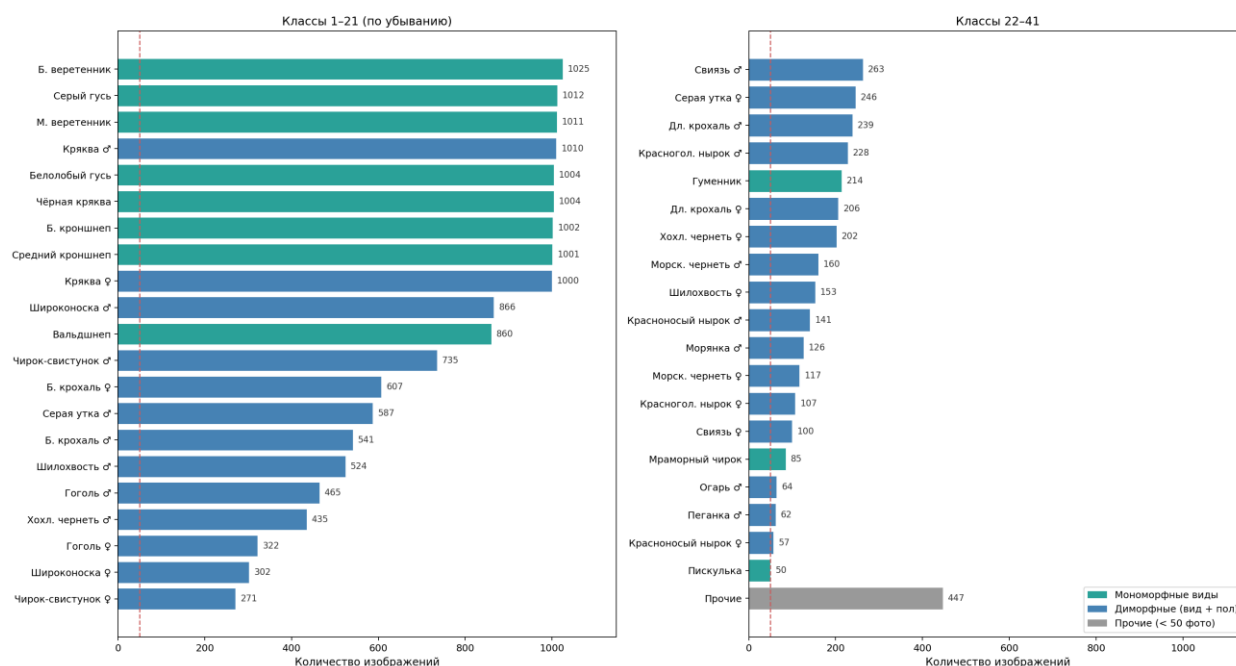


Рисунок 2. Распределение изображений по классам GameWaterbird-41

Для повышения читаемости диаграмма разделена на две панели, а классы отсортированы по убыванию количества изображений. Красная пунктирная линия означает порог в 50 изображений, используемый при формировании классов: виды и пол, содержащие меньшее число изображений, объединяются в класс «прочие». Зеленым цветом обозначены мономорфные виды, синим полово-диморфные классы («вид + пол»).

Постановка экспериментов

В данном исследовании оценивались шесть семейств моделей, охватывающих два типа архитектур (сверточные нейронные сети и визуальные трансформеры) и три подхода к предварительному обучению. Характеристики используемых моделей описаны в таблице 2.

Таблица 2. Модели, включенные в бенчмарк

<i>Модель</i>	<i>Архитектура</i>	<i>Количество параметров</i>	<i>Модель предобучена на датасете</i>	<i>Входное разрешение в пикселях</i>
<i>ResNet-50 [17]</i>	<i>CNN</i>	<i>25M</i>	<i>ImageNet-1K (надзор.)</i>	<i>224</i>
<i>EfficientNet-B3 [18]</i>	<i>CNN</i>	<i>12M</i>	<i>ImageNet-1K (надзор.)</i>	<i>224</i>
<i>CLIP ViT-B/16 [8]</i>	<i>ViT-B/16</i>	<i>150M</i>	<i>Изобр.–текст (400M)</i>	<i>224</i>
<i>BioCLIP ViT-B/16 [9]</i>	<i>ViT-B/16</i>	<i>150M</i>	<i>Биологич. фото (10M)</i>	<i>224</i>
<i>ViT-B/16 AugReg [12]</i>	<i>ViT-B/16</i>	<i>86M</i>	<i>ImageNet-21K (надзор.)</i>	<i>224</i>
<i>DINOv2 ViT-B/14 [10]</i>	<i>ViT-B/14</i>	<i>87M</i>	<i>Самонадзор. (142M)</i>	<i>518</i>

Для оценки возможностей различных подходов рассматривались три режима обучения. Модели CLIP и BioCLIP оценивались без доменно-специфического обучения. В этом режиме сравниваются эмбединги (векторы признаков) изображения и текстовых описаний классов. Эмбединг представляет собой числовое представление изображения, формируемое моделью, при котором визуально похожие изображения имеют близкие векторы признаков. Были протестированы различные варианты текстовых подсказок (научное название, русское название и их комбинации), после чего для каждой модели фиксировался наилучший результат. Поскольку данные модели не имеют механизма предсказания пола, оценка проводилась только на уровне вида. Предсказание считалось верным, если вид был определен правильно, вне зависимости от пола особи.

Для моделей CLIP и BioCLIP также исследовался режим обучения линейного классификатора. При этом визуальный кодировщик использовался с замороженными весами, то есть его параметры не изменялись, а обучался только добавленный поверх него линейный классификатор. Линейный классификатор обучался для распознавания всех 41 классов, при этом веса базовой сети оставались неизменными. Такой режим позволяет оценить, насколько хорошо замороженное признаковое пространство разделяет классы «вид + пол» без полной адаптации модели.

На заключительном этапе шесть моделей дообучались в режиме полной тонкой настройки для решения задачи классификации 41 класса. Поверх каждой базовой модели добавлялась классификационная головка, состоящая из слоя LayerNorm и полносвязного слоя Linear(feat_dim, 41). Параметры базовой сети и классификационной головы обновлялись совместно с дифференциальными скоростями обучения. Параметры обучения, приведенные ниже, необходимы для воспроизведения экспериментов, однако для интерпретации полученных результатов их детальное понимание не требуется.

Все модели обучались с использованием единой конфигурации. В качестве оптимизатора применялся AdamW с коэффициентом затухания весов 0,05. Скорость обучения для классификационной головы составляла 10^{-3} , а для базовой сети 10^{-5} . Обучение проводилось в течение 25 эпох с использованием косинусного отжига и 2-эпохным линейным прогревом. Для компенсации дисбаланса классов (от 50 до 1 010 изображений на класс) применялись WeightedRandomSampler и функция потерь [14] с параметром $\gamma = 2,0$. Весовые коэффициенты классов задавались обратно пропорционально их частоте, а для дополнительного повышения устойчивости обучения использовались сглаживание меток со значением 0,1.

Аугментация обучающих данных включала случайное масштабированное кадрирование (масштаб 0,7–1,0), случайное горизонтальное отражение, поворот изображения в диапазоне $\pm 15^\circ$ и случайное изменение яркости, контрастности и насыщенности в диапазоне $\pm 0,2$. Все преобразования применялись только к области изображения, ограниченной прямоугольником, содержащим птицу. На этапе тестирования использовалась единственная центральная обрезка изображения.

Все модели, кроме DINOv2 ViT-B/14, использовали входное разрешение 224×224 пикселя. Для модели DINOv2 ViT-B/14 использовалось разрешение 518×518 пикселей, соответствующее ее исходной архитектуре ViT-B/14 с размером патча 14×14 и решеткой признаков 37×37 . Обучение всех моделей проводилось на одной видеокарте NVIDIA RTX 4080 SUPER с объемом 16 ГБ.

После полной тонкой настройки моделей выполнялось температурное масштабирование. Для этого все логиты модели делились на скалярный параметр температуры, после чего к ним применяется нормирующее преобразование softmax:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}.$$

Значение температуры подбиралось на валидационной выборке путем минимизации отрицательного логарифмического правдоподобия. При этом веса модели не изменялись, поскольку процедура выполнялась после завершения обучения. Температурное масштабирование не изменяет класс с максимальной вероятностью, поэтому точность Top-1 остается неизменной, но можно скорректировать уровень уверенности модели в предсказаниях. Во всех проведенных экспериментах оптимальное значение температуры оказалось меньше единицы. Это указывает на заниженную уверенность моделей, и характерной особенностью обучения с использованием сглаживания меток и функции потерь. В отличие от типичной для нейросетевых классификаторов избыточной уверенности, которая требует температуры больше единицы для сглаживания распределения вероятностей, здесь наблюдается обратный эффект: температура меньше единицы делает распределение вероятностей более выраженным.

На этапе тестирования дополнительно исследовалось влияние аугментации. Для каждого тестового изображения формировалось восемь детерминированных вариантов, включающих центральные обрезки трех масштабов, их горизонтальные отражения, вариант с изменением цветовых характеристик и вариант с поворотом изображения. Итоговые предсказания получали путем усреднения логитов всех вариантов изображения. При совместном использовании методов усреднение логитов выполнялось перед применением температурного масштабирования.

Качество моделей оценивалось по показателям точности Top-1 и Top-5 на тестовой выборке, содержащей 1 865 изображений. Для оценки качества калибровки вероятностных предсказаний использовалась ожидаемая ошибка калибровки (ECE), вычисляемая по 15 равным интервалам [13]. По смыслу ECE представляет собой средневзвешенное расхождение между заявленной уверенностью модели и фактической долей правильных ответов. Для её

вычисления предсказания разбиваются на интервалы по уровню уверенности, после чего в каждом интервале заявленная уверенность сравнивается с фактической точностью. Нулевое значение ECE соответствует идеальной калибровке. Для режима без дополнительного обучения и режима обучения только линейного классификатора значение ECE не рассчитывалось, поскольку показатели калибровки несопоставимы между режимами.

Метрики качества. Основным показателем качества классификации является точность Top-1, определяемая как доля изображений, для которых верный класс получил наибольшую оценку среди всех 41 класса. Дополнительно рассчитывается точность Top-5 – доля изображений, для которых верный класс вошел в число пяти наиболее вероятных. Второй показатель показывает, насколько близко модель была к правильному ответу в тех случаях, когда все же ошиблась.

$$\text{Top} - 1 = \frac{M_1}{N}, \quad \text{Top} - 5 = \frac{M_5}{N},$$

где N – размер тестовой выборки (1 865 изображений), M_1 – число изображений, для которых верный класс получил наибольшую оценку, M_5 – число изображений, для которых верный класс попал в первую пятерку.

Для анализа качества распознавания отдельных классов используются две метрики: полнота (Recall) и точность (Precision). При расчете этих метрик каждый класс k рассматривается независимо от остальных, и все тестовые изображения распределяются по одной из четырех категорий:

- TP (истинно-положительные) изображения класса k , правильно отнесенные моделью к классу k ;
- FN (ложно-отрицательные) изображения класса k , ошибочно отнесенные моделью к другому классу;
- FP (ложно-положительные) изображения других классов, ошибочно отнесенные моделью к классу k ;
- TN (истинно-отрицательные) изображения других классов, которые модель к классу k и не относил.

На основе этих категорий рассчитываются Recall и Precision:

$$\text{Recall} = \frac{TP}{(TP + FN)}, \quad \text{Precision} = \frac{TP}{(TP + FP)}.$$

Recall отвечает на вопрос: «Какую долю изображений птиц данного класса модель распознала правильно?», тогда как Precision отвечает на вопрос: «Какая доля изображений птиц, отнесенных моделью к данному классу, действительно принадлежит этому классу?». Обе метрики определяются непосредственно по матрице ошибок, где Recall вычисляется по строке соответствующего класса, а Precision по его столбцу.

Например, для самки серой утки (*Mareca strepera*) из 24 тестовых изображений модель верно распознала 19 (TP), пять отнесла к другим классам (FN), а еще девять изображений других видов ошибочно отнесла к этому классу (FP). В результате Recall составила $19/24=79\%$, а Precision $19/28=68\%$. Такое расхождение свидетельствует о том, что к классу самки серой утки ошибочно относятся самки других видов, прежде всего кряквы, что согласуется с наблюдением о криптической окраске самок.

Термин «точность» используется в статье в двух значениях: точность Top-1 – доля правильных классификаций на всей тестовой выборке, а Precision – доля правильно классифицированных изображений среди всех изображений, отнесенных моделью к данному классу. Во избежание разночтений далее для второй метрики сохраняется английский термин Precision, а для доли правильно распознанных изображений данного класса используется термин «Полнота» (recall).

Результаты экспериментов

В таблице 3 представлены результаты сравнительного анализа (бенчмарка) всех моделей на тестовой выборке, содержащей 1 865 изображений, относящихся к 41 классу. Для моделей, прошедших полную тонкую настройку, приведены значения точности Top-1, Top-5 и ожидаемой ошибки калибровки (ECE), полученные при совместном использовании температурного масштабирования и восьмикратной аугментации на этапе тестирования (TTA×8), обеспечивших наилучшее качество классификации. В режиме без дополнительного обучения (zero-shot) оценка проводилась только на уровне вида: модель предсказывала вид по текстовым подсказкам без учета пола особи, а предсказание считалось верным при корректном определении вида вне зависимости от пола. Это делает результаты несопоставимыми напрямую с дообученными моделями, которые выполняют классификацию по полному набору классов «вид+пол». Значение ожидаемой ошибки калибровки (ECE) указывается только для дообученных моделей после полной тонкой настройки.

Строки таблицы сгруппированы по режимам обучения. Колонки Top-1 и Top-5 отражают долю правильно классифицированных изображений: чем выше значение, тем лучше качество модели. Колонка ECE характеризует расхождение между уровнем уверенности модели и фактической точностью её предсказаний: чем меньше значение, тем лучше калибровка модели.

Таблица 3. Результаты бенчмарка моделей на тестовой выборке GameWaterbird-41

Модель	Режим	Top-1 (%)	Top-5 (%)	ECE
CLIP ViT-B/16	Zero-shot	47,51	87,02	–
BioCLIP ViT-B/16	Zero-shot	67,99	90,83	–
CLIP ViT-B/16	Линейный классификатор	31,80	69,71	–
BioCLIP ViT-B/16	Линейный классификатор	54,10	82,52	–
ResNet-50	Тонкая настройка (T+TTA)	74,64	93,46	0,051
EfficientNet-B3	Тонкая настройка (T+TTA)	81,34	95,55	0,040
BioCLIP ViT-B/16	Тонкая настройка (T+TTA)	90,99	97,16	0,011
CLIP ViT-B/16	Тонкая настройка (T+TTA)	91,90	97,43	0,023
ViT-B/16 AugReg	Тонкая настройка (T+TTA)	93,78	98,39	0,013
DINOv2 ViT-B/14	Тонкая настройка (T+TTA)	95,34	98,55	0,010

Полученные результаты показывают, что тонкая настройка (дообучение), то есть продолжение обучения предварительно обученной модели на собственных данных, является необходимым этапом для достижения высокой точности классификации. Лучший результат в режиме без дополнительного обучения демонстрирует модель BioCLIP (67,99%), однако ее точность уступает на 27 процентных пунктов по сравнению с лучшей дообученной моделью DINOv2 (95,34%). Результаты свидетельствуют о том, что предобучение на общих биологических наборах данных улучшает качество извлекаемых признаков, но не заменяет доменную адаптацию при совместной классификации вида и пола.

Для модели CLIP обучение только линейного классификатора (31,80%) привело к снижению точности по сравнению с режимом zero-shot (47,51%). Данный результат обусловлен различием постановки задачи, в режиме zero-shot классификация выполнялась только на уровне вида (без учета пола), тогда как линейный классификатор обучался различать 41 класс, включая половое разделение. При использовании замороженного пространства признаков CLIP линейный классификатор не разделяет полные классы «вид+пол» без сквозной адаптации параметров модели. Для BioCLIP снижение точности менее выраженное (с 67,99% до 54,10%), что указывает на более информативное признаковое пространство, сформированное в процессе предобучения на биологических изображениях.

Наилучшие результаты среди всех дообученных моделей продемонстрировала модель DINOv2 ViT-B/14, достигшая точности Top-1, равное 95,34% и Top-5, равной 98,55%. За ней

следуют ViT-B/16 AugReg (93,78%) и CLIP ViT-B/16 (91,90%). Модель DINOv2 использует исходное разрешение изображения 518×518 пикселей вместо 224×224 и самонадзорное предобучение, основанное на извлечении плотных признаков, что обеспечивает преимущество при тонкозернистой классификации.

Сверточные нейронные сети (CNN) показывают более низкую точность по сравнению с визуальными трансформерами. Модели ResNet-50 и EfficientNet-B3, несмотря на высокие результаты на наборе данных ImageNet, достигли точности лишь 74,64% и 81,34% соответственно. Это может быть связано с тем, что механизм глобального внимания визуальных трансформеров лучше учитывает тонкие различия в окрасе и структуре оперения, характерные для задач тонкозернистой классификации.

Оценка калибровки модели. Калибровка модели является необходимым условием для практического применения модели в полевых условиях. Охотник, консультирующийся с моделью, ожидает, что заявленный уровень уверенности соответствует фактической точности. Без калибровки исходные оценки уверенности DINOv2 систематически занижены. На рисунке 3 приведен пример классификации самца чирка-свистунка (*Anas crecca*), где без калибровки ($T=1,0$) правильному классу присваивается лишь 51,8% вероятности, хотя он лидирует среди всех альтернатив. После использования температурного масштабирования ($T=0,494$) вероятность правильного класса возрастает до 95,9%, при этом ранжирование классов не меняется, а корректируются только абсолютные значения оценок уверенности. В таблице 4 представлены результаты оценки влияния температурного масштабирования и TTA на показатели качества модели DINOv2 ViT-B/14. Значение параметра температуры $T = 0,494$ было подобрано на валидационной выборке.



Рисунок 3. Влияние температурного масштабирования на уверенность модели для самца чирка-свистунка (*Anas crecca*)

Таблица 4. Влияние температурного масштабирования и TTA на DINOv2 ViT-B/14

Конфигурация	Top-1	Top-5	ECE
Базовая ($T = 1,0$, без TTA)	95,17	98,61	0,308
+ Темп. ($T = 0,494$)	95,17	98,61	0,013
+ TTA×8 ($T = 1,0$)	95,34	98,55	0,310
+ Темп. + TTA×8	95,34	98,55	0,010

Температурное масштабирование снижает значение ECE с 0,308 до 0,013 без изменения точности Top-1, что подтверждает корректность ранжирования классов моделью при заниженном уровне уверенности исходных предсказаний. Иными словами, до калибровки заявленная моделью уверенность в среднем расходилась с фактической точностью примерно

на 31 процентный пункт, тогда как после калибровки это расхождение сократилось до 1,3 пункта. Применение ТТА повышает точность Top-1 на 0,17 процентного пункта по сравнению с использованием одного варианта изображения. Совместное применение обеих техник дает лучшее значение ECE, равное 0,010, при сохранении прироста точности, полученного за счет ТТА.

Анализ по группам видов. В таблице 5 представлены значения точности Top-1 модели DINOv2 ViT-B/14 с использованием температурного масштабирования и аугментации на этапе тестирования (Т+ТТА) для различных групп видов.

Таблица 5. Точность Top-1 модели DINOv2 ViT-B/14 (Т+ТТА) по группам видов

Группа видов	Top-1, %
Кулики (<i>Numenius</i> , <i>Limosa</i> , <i>Scolopax</i>)	95–100
Речные утки, самцы (6 классов)	93–100
Речные утки, самки (6 классов)	79–100
Речные утки, мономорфные (<i>Anas zonorhyncha</i>)	95
Нырковые утки, самцы (6 классов)	93–100
Нырковые утки, самки (5 классов)	80–100
Нырковые утки, мономорфные (<i>Marmaronetta angustirostris</i>)	88
Крохали (<i>Mergus spp.</i> , 4 класса)	91–97
Пеганки (<i>Tadorna spp.</i> , 2 класса)	100
Гуси (<i>Anser spp.</i> , 4 вида)	93–100

Как видно из таблицы 5, наиболее высокие значения точности достигнуты для куликов. Для вальдшнепа (*Scolopax rusticola*) и большого веретенника (*Limosa limosa*) точность достигает 100%, ни для одного вида куликов точность не снижается ниже 95%. Это обусловлено характерными морфологическими признаками этих видов, такими как особенности облика и размеров, а также отсутствием близких визуально сходных классов в наборе данных.

Для полово-диморфных видов точность классификации самцов выше, чем у самок. Самцы уток, как правило, имеют яркое контрастное брачное оперение, тогда как самки большинства видов окрашены в однотонные бурые тона. Это приводит к более низкой и вариабельной точности по классификации самок (79–100%) относительно самцов (91–100%). Наиболее сложными для распознавания оказались самка серой утки (*Mareca strepera*, 79%) и самка красноголового нырка (*Aythya ferina*, 80%), что обусловлено их криптической окраской и высоким сходством с другими видами в пределах своих таксономических групп.

Для гусей основная часть ошибок связана с взаимной путаницей между близкими видами. Точность Top-1 варьируется от 93% для белолобого гуся (*Anser albifrons*) до 96% для серого гуся (*Anser anser*). Чаще ошибки возникают между *A. albifrons* и пискулькой (*A. erythropus*). 5 из 100 тестовых снимков *A. albifrons* были классифицированы как *A. erythropus*. Это связано с тем, что диагностические признаки, такие как размер белого пятна на лбу или наличие желтого глазного кольца, часто неразличимы на исходных фотографиях, что ограничивает точность классификации по визуальным данным. При интерпретации результатов для отдельных классов следует учитывать объем тестовой выборки. Для ряда редких видов тестовая выборка содержит не более пяти изображений, например, для *Sibirionetta formosa*, *Oxyura leucocephala*, самка *Mergellus albellus*. В таких случаях даже одна ошибка классификации может изменить значение точности на 20–100 процентных пунктов. Поэтому диапазоны, приведенные в таблице 5, отражают общую тенденцию, тогда как оценки для редких классов ограничены малым объемом тестовой выборки. Полная нормализованная по строкам матрица ошибок модели DINOv2 ViT-B/14 (ТТА×8) на тестовой выборке GameWaterbird-41, содержащей 1 865 изображений, представлена на рисунках 4–6. Поскольку при 41 классе единая матрица становится трудночитаемой в пределах ширины страницы, она разделена на три блока в соответствии с таксономической принадлежностью классов: речные

утки (Рисунок 4), нырковые утки (Рисунок 5), а также крохали, пеганки, гуси, кулики и класс «прочие» (Рисунок 6).

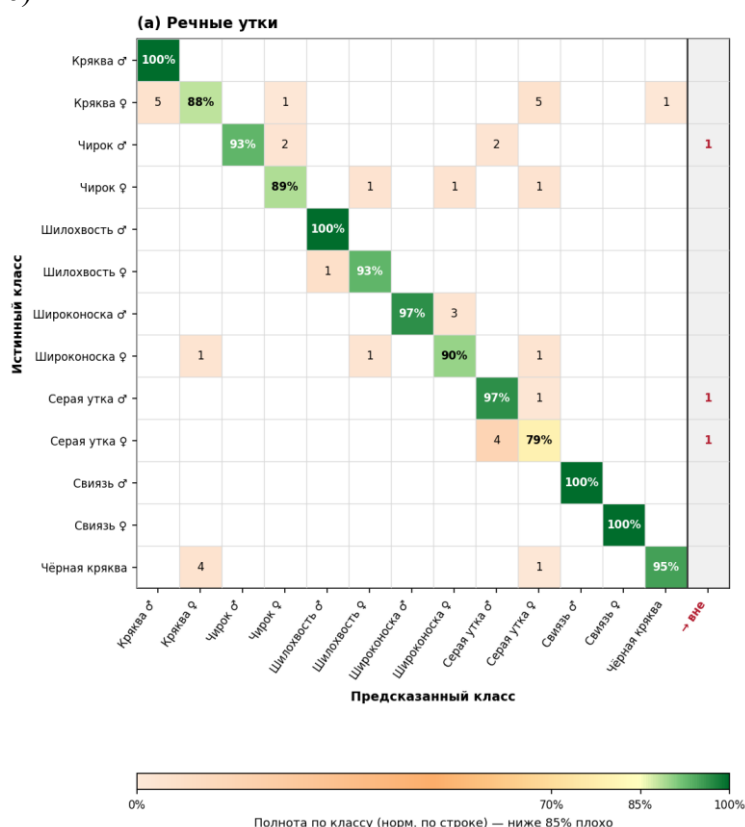


Рисунок 4. Матрица ошибок (блок 1 из 3), речные утки

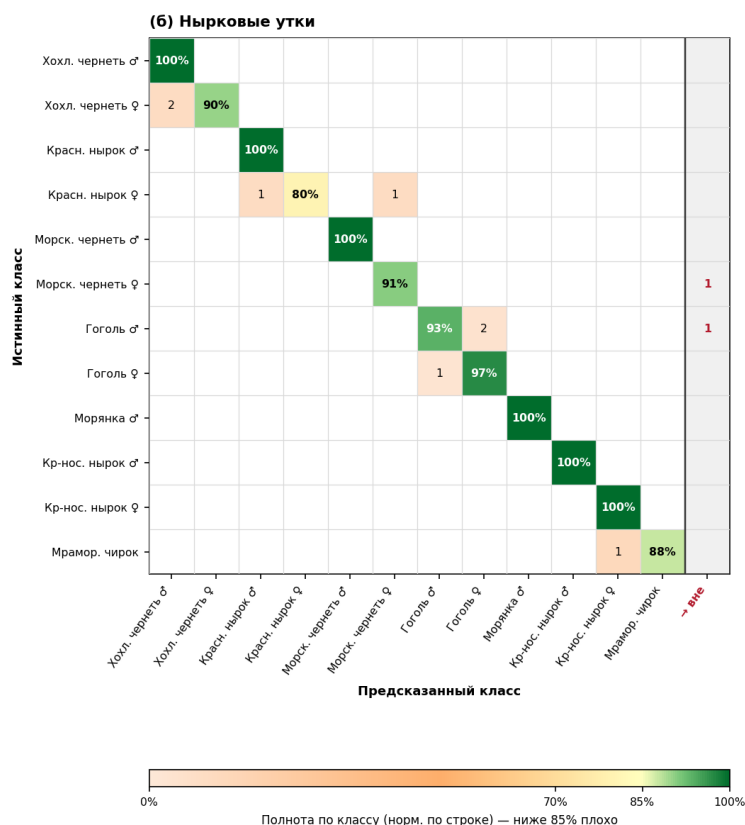


Рисунок 5. Матрица ошибок (блок 2 из 3), нырковые утки

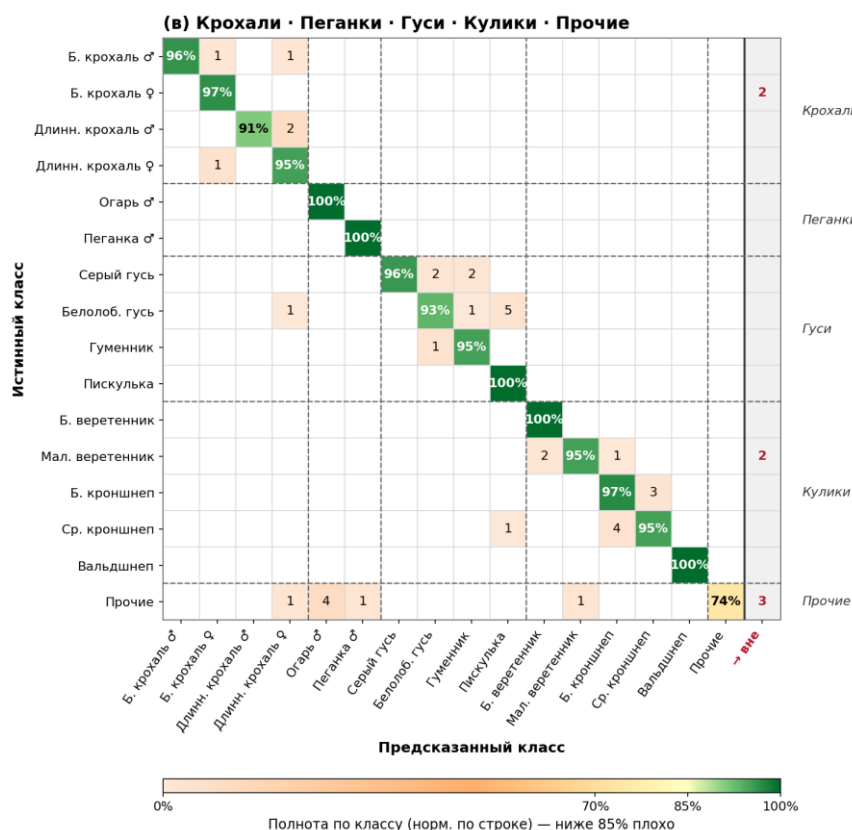


Рисунок 6. Матрица ошибок (блок 3 из 3), крохали, пеганки, гуси, кулики, прочие

На диагонали представлены значения полноты (recall) для каждого класса, вне диагонали количество ошибок классификации (≥ 1). Белые ячейки соответствуют отсутствию ошибок классификации, а пунктирные линии разделяют таксономические группы внутри рисунка 6. Строки матрицы нормализованы по всем 41 классу, а не внутри блока. Ошибки, выходящие за пределы блока, собраны в отдельном столбце «→ вне», поэтому сумма значений в каждой строке по-прежнему составляет 100%. Доля таких ошибок составляет 12 из 91 (13%), то есть основная часть ошибок приходится на классы внутри одного таксономического блока. Путаница наблюдается внутри группы гусей (*Anser albifrons* ↔ *A. erythropus*), а также между самками уток со сходной покровной окраской.

Дискуссия

Полученные результаты показывают, что при решении задачи совместной классификации вида и пола птиц определяющим фактором является не специализация предобучения, а дообучение на целевых данных. Разрыв между лучшим результатом в режиме без дополнительного обучения (BioCLIP, 67,99%) и лучшей дообученной моделью (DINOv2 ViT-B/14, 95,34%) превышает 27 процентных пунктов. При этом BioCLIP, предобученная на биологических изображениях, существенно превосходит CLIP общего назначения в режиме без дообучения (67,99% против 47,51%), тогда как после полной тонкой настройки разрыв между ними сокращается примерно до одного процентного пункта. Наилучший результат показывает модель DINOv2, обученная без разметки и без какой-либо биологической специализации. Это указывает на то, что признаки, необходимые для различения пола и близкородственных видов, извлекаются преимущественно из целевой выборки, а не наследуются из доменного предобучения.

Второй результат касается природы оставшихся ошибок, которые в большей степени связаны с определением пола, чем вида. Полнота для самцов лежит в диапазоне 91–100%,

тогда как для самок 79–100%. Анализ отдельных классов показывает, что самки не только хуже распознаются, но чаще становятся источником ложноположительных классификаций. Так, для самки серой утки полнота составляет 79%, а precision всего лишь 68%, поскольку к данному классу модель ошибочно относит самок кряквы и других речных уток. Это согласуется с биологическими особенностями видов: самки большинства уток обладают покровительственной окраской, тогда как брачное оперение самцов отличается контрастностью и выраженной видоспецифичностью. Для практических задач это означает, что наибольшую сложность представляет определение пола особей птиц с неяркой окраской.

Третий результат позволяет определить пределы применимости предложенного подхода. Для гусей рода *Anser* ошибки носят межвидовой характер и сосредоточены в паре *Anser albifrons* и *A. erythropus*, где 5 из 100 тестовых изображений белолобого гуся ошибочно классифицированы как пискулька. Диагностические признаки этих видов, такие как размер белого пятна на лбу и наличие желтого окологлазного кольца, на любительских фотографиях часто неразличимы. Это указывает на то, что ограничение обусловлено не столько особенностями модели, сколько недостаточной информативностью исходных изображений. Необходимые диагностические признаки отсутствуют или выражены недостаточно, поэтому увеличение объема выборки или усложнение модели не гарантирует устранения подобных ошибок. Схожая закономерность отмечается и в других прикладных исследованиях по распознаванию биологических объектов, где точность, достигнутая на курируемых наборах данных, снижается при переходе к изображениям, полученным в естественных условиях.

Отдельного внимания заслуживает калибровка модели. Сырые выходы дообученных моделей систематически занижают уверенность. Ожидаемая ошибка калибровки составляет 0,308, а температурное масштабирование снижает её до 0,013, без изменения предсказаний. Для рассматриваемой задачи это имеет практическое значение, поскольку решение о добыче принимается по пороговому значению уверенности. Некалиброванная модель не позволяет корректно интерпретировать этот показатель, так как заявленные 50% могут соответствовать и 30%, так и 95% вероятности правильной классификации. Практическая ценность работы состоит не только в достигнутой точности, но и в том, что показатель уверенности становится интерпретируемым и пригодным для построения правил принятия решения.

Ограничения работы определяют направления дальнейших исследований. Изображения получены с платформы iNaturalist и сняты преимущественно в благоприятных условиях, тогда как в полевой обстановке птицы нередко наблюдаются на большом расстоянии, против света или в полете. Для ряда редких классов тестовая выборка содержит менее пяти изображений, что не позволяет делать надежных выводов об их распознавании. Набор данных не различает переходное (эклипсное) оперение, при котором самцы уток внешне сближаются с самками, что характерно для части охотничьего сезона. Наконец, для мономорфных видов определение пола по фотографии невозможно в принципе, поэтому данный случай сознательно выведен за рамки постановки задачи. Дальнейшая работа предполагает расширение набора данных изображениями, полученными в полевых условиях, введение классов эклипсного оперения, а также проверку устойчивости моделей на данных, независимых от платформы iNaturalist.

Заключение

В настоящей работе представлен набор данных GameWaterbird-41, содержащий 18 851 изображение 38 охотничьих видов водоплавающих птиц и куликов, являющихся объектами охоты, объединенных в 41 класс с учетом вида и пола. Также проведен сравнительный анализ шести семейств моделей компьютерного зрения в режимах без дополнительного обучения (zero-shot), линейного классификатора и полной тонкой настройки.

Полученные результаты демонстрируют, что дообученные визуальные трансформеры существенно превосходят CNN-модели и фундаментальные модели, используемые без дополнительного обучения. Наилучший результат показывает модель DINOv2 ViT-B/14, достигнув точности 95,34% по показателю Top-1 и 98,55% по показателю Top-5 при

совместном применении температурного масштабирования и аугментации на этапе тестирования (ТТА). Обучение с использованием сглаживания меток и функции потерь Focal Loss приводит к заниженной оценке уверенности модели, что эффективно компенсируется температурным масштабированием.

Сложной группой для классификации остаются гуси рода *Anser*, для которых высокое межвидовое сходство ограничивает точность распознавания по фотографическим данным. Дальнейшие исследования будут направлены на расширение набора данных за счет малочисленных классов, а также на систематическую ручную проверку разметки каждого класса, прежде всего для визуально схожих пар видов, аннотации которых наиболее подвержены ошибкам идентификации.

Благодарность

Исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан в рамках проекта «Разработка цифровой системы управления охотничьими хозяйствами Республики Казахстан с применением методов машинного обучения» (ИРН AP26104615).

Список использованных источников

- [1] Wah C., Branson S., Welinder P., Perona P., Belongie S. The Caltech-UCSD Birds-200-2011 Dataset. – 2011. – Technical Report № CNS-TR-2011-001. – 8 p. – URL: <https://authors.library.caltech.edu/records/cvm3y-5hh21>.
- [2] Van Horn G., Branson S., Farrell R., Haber S., Barry J., Ipeirotis P., Perona P., Belongie S. Building a Bird Recognition App and Large Scale Dataset with Citizen Scientists: The Fine Print in Fine-Grained Dataset Collection // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2015. – P. 595–604. – DOI: [10.1109/CVPR.2015.7298658](https://doi.org/10.1109/CVPR.2015.7298658).
- [3] Республики Казахстан. Законы. Об охране, воспроизводстве и использовании животного мира: нормативные правовые акты: принят Парламентом Республики Казахстан 9 июля 2004 года № 593. – URL: <https://adilet.zan.kz/rus/docs/Z040000593> (дата обращения: 13.07.2026).
- [4] Правила охоты: утв. приказом и.о. Министра сельского хозяйства Республики Казахстан от 27 февраля 2015 года № 18-03/157: зарегистрирован в Министерстве юстиции Республики Казахстан 20 мая 2015 года № 11091. – URL: <https://adilet.zan.kz/rus/docs/V1500011091> (дата обращения: 13.07.2026).
- [5] Налоговый кодекс Республики Казахстан: Кодекс Республики Казахстан от 18 июля 2025 года № 214-VIII ЗПК: по состоянию на 01.07.2026. – URL: <https://adilet.zan.kz/rus/docs/K2500000214> (дата обращения: 13.07.2026).
- [6] Berg T., Liu J., Lee S.W., Alexander M.L., Jacobs D.W., Belhumeur P.N. Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2014. – P. 2019–2026. DOI: [10.1109/CVPR.2014.259](https://doi.org/10.1109/CVPR.2014.259).
- [7] Van Horn G., Cole E., Beery S., Wilber K., Belongie S., Mac Aodha O. Benchmarking Representation Learning for Natural World Image Collections // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2021. – P. 12884–12893. DOI: [10.1109/CVPR46437.2021.01269](https://doi.org/10.1109/CVPR46437.2021.01269).
- [8] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. Learning Transferable Visual Models from Natural Language Supervision // *Proceedings of the 38th International Conference on Machine Learning*. – 2021. – Vol. 139. – P. 8748–8763. – URL: <https://proceedings.mlr.press/v139/radford21a.html> (дата обращения: 13.07.2026). DOI: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020).
- [9] Stevens S., Wu J., Thompson M.J., Campolongo E.G., Song C.H., Carlyn D.E., Dong L., Dahdul W.M., Stewart C., Berger-Wolf T., Chao W.-L., Su Y. BioCLIP: A Vision Foundation Model for the Tree of Life // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2024. – P. 19412–19424. – DOI: [10.1109/CVPR52733.2024.01836](https://doi.org/10.1109/CVPR52733.2024.01836).
- [10] Oquab M., Darcet T., Moutakanni T., Vo H. V., Szafraniec M., Khalidov V., Fernandez P., Haziza D., Massa F., El-Nouby A., Assran M., Ballas N., Galuba W., Howes R., Huang P.-Y., Li S.- W., Misra I., Rabbat M., Sharma V., Synnaeve G., Xu H., Jégou H., Mairal J., Labatut P., Joulin A., Bojanowski P. DINOv2:

Learning Robust Visual Features without Supervision // *Transactions on Machine Learning Research*. – 2024. – 32 p. – URL: <https://openreview.net/forum?id=a68SUt6zFt> (дата обращения: 13.07.2026). DOI: [10.48550/arxiv.2304.07193](https://doi.org/10.48550/arxiv.2304.07193).

[11] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. *An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale* // *Proceedings of the International Conference on Learning Representations (ICLR)*. – 2021. – 21 p. – URL: <https://openreview.net/pdf?id=YicbFdNTTy> (дата обращения: 13.07.2026). DOI: [10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929).

[12] Steiner A., Kolesnikov A., Zhai X., Wightman R., Uszkoreit J., Beyer L. *How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers* // *Published in Transactions on Machine Learning Research*. – 2022. – 16 p. – URL: <https://openreview.net/pdf?id=4nPswr1KcP> (дата обращения: 13.07.2026). DOI: [10.48550/arXiv.2106.10270](https://doi.org/10.48550/arXiv.2106.10270).

[13] Guo C., Pleiss G., Sun Y., Weinberger K. Q. *On Calibration of Modern Neural Networks* // *Proceedings of the 34th International Conference on Machine Learning*. – 2017. – Vol. 70. – P. 1321–1330. – URL: <https://proceedings.mlr.press/v70/guo17a.html> (дата обращения: 13.07.2026).

[14] Lin T.-Y., Goyal P., Girshick R., He K., Dollar P. *Focal Loss for Dense Object Detection* // *Proceedings of the IEEE International Conference on Computer Vision*. – 2017. – P. 2999–3007. DOI: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).

[15] iNaturalist. – URL: <https://www.inaturalist.org> (дата обращения: 13.07.2026).

[16] Jocher G. *Ultralytics YOLO*. – URL: <https://github.com/ultralytics/ultralytics> (дата обращения: 13.07.2026).

[17] He K., Zhang X., Ren S., Sun J. *Deep Residual Learning for Image Recognition* // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. – 2016. – P. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).

[18] Tan M., Le Q.V. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* // *Proceedings of the 36th International Conference on Machine Learning*. – 2019. – Vol. 97. – P. 6105–6114. URL: <https://proceedings.mlr.press/v97/tan19a.html> (дата обращения: 17.08.2026).

References

[1] Wah C., Branson S., Welinder P., Perona P., Belongie S. *The Caltech-UCSD Birds-200-2011 Dataset*. – 2011. – Technical Report № CNS-TR-2011-001. – 8 p. – URL: <https://authors.library.caltech.edu/records/cvm3y-5hh21>.

[2] Van Horn G., Branson S., Farrell R., Haber S., Barry J., Ipeirotis P., Perona P., Belongie S. *Building a Bird Recognition App and Large Scale Dataset with Citizen Scientists: The Fine Print in Fine-Grained Dataset Collection* // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2015. – P. 595–604. – DOI: [10.1109/CVPR.2015.7298658](https://doi.org/10.1109/CVPR.2015.7298658).

[3] Respublika Kazakhstan. *Ob okhrane, vosпроизводстве i ispol'zovanii zhivotnogo mira* [On the Protection, Reproduction and Use of Wildlife]: Law of the Republic of Kazakhstan No. 593 dated July 9, 2004. Available at: <https://adilet.zan.kz/rus/docs/Z040000593> (accessed July 13, 2026). (In Russian).

[4] *Pravila okhoty* [Hunting Rules]: approved by Order of the Acting Minister of Agriculture of the Republic of Kazakhstan No. 18-03/157 dated February 27, 2015; registered with the Ministry of Justice of the Republic of Kazakhstan on May 20, 2015, No. 11091. Available at: <https://adilet.zan.kz/rus/docs/V1500011091> (accessed July 13, 2026). (In Russian).

[5] *Nalogovyi kodeks Respubliki Kazakhstan* [Tax Code of the Republic of Kazakhstan]: Code of the Republic of Kazakhstan No. 214-VIII ZRK dated July 18, 2025, as amended on July 1, 2026. Available at: <https://adilet.zan.kz/rus/docs/K2500000214> (accessed July 13, 2026). (In Russian).

[6] Berg T., Liu J., Lee S.W., Alexander M.L., Jacobs D.W., Belhumeur P.N. *Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds* // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2014. – P. 2019–2026. DOI: [10.1109/CVPR.2014.259](https://doi.org/10.1109/CVPR.2014.259).

[7] Van Horn G., Cole E., Beery S., Wilber K., Belongie S., Mac Aodha O. *Benchmarking Representation Learning for Natural World Image Collections* // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2021. – P. 12884–12893. DOI: [10.1109/CVPR46437.2021.01269](https://doi.org/10.1109/CVPR46437.2021.01269).

[8] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. *Learning Transferable Visual Models from Natural Language Supervision*

// *Proceedings of the 38th International Conference on Machine Learning*. – 2021. – Vol. 139. – P. 8748–8763. – URL: <https://proceedings.mlr.press/v139/radford21a.html> (accessed July 13, 2026). DOI: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020).

[9] Stevens S., Wu J., Thompson M.J., Campolongo E.G., Song C.H., Carlyn D.E., Dong L., Dahdul W.M., Stewart C., Berger-Wolf T., Chao W.-L., Su Y. BioCLIP: A Vision Foundation Model for the Tree of Life // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2024. – P. 19412–19424. – DOI: [10.1109/CVPR52733.2024.01836](https://doi.org/10.1109/CVPR52733.2024.01836).

[10] Oquab M., Darcet T., Moutakanni T., Vo H. V., Szafraniec M., Khalidov V., Fernandez P., Haziza D., Massa F., El-Nouby A., Assran M., Ballas N., Galuba W., Howes R., Huang P.-Y., Li S.- W., Misra I., Rabbat M., Sharma V., Synnaeve G., Xu H., Jégou H., Mairal J., Labatut P., Joulin A., Bojanowski P. DINOv2: Learning Robust Visual Features without Supervision // *Transactions on Machine Learning Research*. – 2024. – 32 p. – URL: <https://openreview.net/forum?id=a68SUt6zFt> (accessed July 13, 2026). DOI: [10.48550/arxiv.2304.07193](https://doi.org/10.48550/arxiv.2304.07193).

[11] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale // *Proceedings of the International Conference on Learning Representations (ICLR)*. – 2021. – 21 p. – URL: <https://openreview.net/pdf?id=YicbFdNTTy> (accessed July 13, 2026). DOI: [10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929).

[12] Steiner A., Kolesnikov A., Zhai X., Wightman R., Uszkoreit J., Beyer L. How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers // *Published in Transactions on Machine Learning Research*. – 2022. – 16 p. – URL: <https://openreview.net/pdf?id=4nPswr1KcP> (accessed July 13, 2026). DOI: [10.48550/arXiv.2106.10270](https://doi.org/10.48550/arXiv.2106.10270).

[13] Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks // *Proceedings of the 34th International Conference on Machine Learning*. – 2017. – Vol. 70. – P. 1321–1330. – URL: <https://proceedings.mlr.press/v70/guo17a.html> (accessed July 13, 2026).

[14] Lin T.-Y., Goyal P., Girshick R., He K., Dollar P. Focal Loss for Dense Object Detection // *Proceedings of the IEEE International Conference on Computer Vision*. – 2017. – P. 2999–3007. DOI: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).

[15] iNaturalist. – URL: <https://www.inaturalist.org> (accessed July 13, 2026).

[16] Jocher G. Ultralytics YOLO. – URL: <https://github.com/ultralytics/ultralytics> (accessed July 13, 2026).

[17] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. – 2016. – P. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).

[18] Tan M., Le Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // *Proceedings of the 36th International Conference on Machine Learning*. – 2019. – Vol. 97. – P. 6105–6114. URL: <https://proceedings.mlr.press/v97/tan19a.html> (accessed August 17, 2026).