Анализ семантического разрыва при классификации изображений по функциональному назначению

А.Н. Жданова [1,], А.В. Куприянов [1,2]*
[1] Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе, д. 34;
[2] Институт систем обработки изображений, НИЦ «Курчатовский институт», 443001, Россия, г. Самара, ул. Молодогвардейская, д. 151;
** автор для связи*

Аннотация

Рассматривается задача классификации изображений по функциональному назначению, отличающаяся от традиционной объектной классификации тем, что категория определяется функциональной ролью изображения и зависит от состава объектов сцены, их взаимного расположения и контекста. Формализована проблема семантического разрыва между объектным признаковым пространством, формируемым свёрточными нейронными сетями, предобученными на объектных датасетах, и пространством функциональных категорий. Предложена таксономия функциональных категорий изображений с формальными критериями разграничения, основанными на относительной значимости объектов и их взаимном расположении.

Экспериментальные исследования проведены на выборке из 1722 реальных изображений профилей пользователей социальных сетей с экспертной разметкой (коэффициент согласованности Коэна $\kappa = 0{,}89$). При использовании архитектуры ResNet50 с двухэтапным переносом обучения достигнуты значения Accuracy $= 0{,}85$ и $F1 = 0{,}83$. Выполнен анализ структуры ошибок классификации, показавший, что 67 % ошибочных решений связаны с пограничными случаями, в которых объектный состав изображения не определяет функциональную категорию однозначно.

Проведён количественный анализ признакового пространства, выявивший существенное снижение разделимости категорий, различающихся отношениями между объектами, по сравнению с категориями, различающимися объектным составом. Полученные результаты подтверждают наличие семантического разрыва между объектным и функциональным представлениями изображений и указывают на необходимость явного учёта отношений между объектами при решении задач функциональной классификации.

Ключевые слова: классификация изображений, функциональные категории, семантический разрыв; перенос обучения, свёрточные нейронные сети, признаковое пространство.

Цитирование: **Жданова, А.Н.** Анализ семантического разрыва при классификации изображений по функциональному назначению / А.Н. Жданова, А.В. Куприянов // Компьютерная оптика. – 2026. – Т. 50, № 5. – 2079. – doi: 10.18287/COJ2079.

Citation: Zhdanova AN, Kupriyanov AV. Semantic gap analysis in image classification by functional purpose. Computer Optics 2026; 50(5): 2079. doi: 10.18287/COJ2079.

Введение

В задачах анализа пользовательского контента, управления мультимедийными коллекциями, интеллектуального поиска и рекомендательных систем возникает необходимость классификации изображений по их функциональному назначению. Анализ содержимого изображений является важным компонентом систем поиска по изображениям, организации мультимедийных коллекций и автоматической обработки пользовательского контента [1 – 3]. В отличие от объектной классификации, в таких задачах категория определяется ролью изображения и зависит от состава сцены, взаимного расположения объектов и контекста.

Изображения с одинаковым набором объектов могут относиться к различным категориям. Например, фотография человека и животного может использоваться как персональный портрет, изображение владельца с питомцем или фотография животного. При этом объектный состав изображения остаётся неизменным, тогда как функциональная категория определяется композицией сцены и относительной значимостью объектов.

Стандартные бенчмарки компьютерного зрения, такие как ImageNet [4], CIFAR [5] и COCO [6], ориентированы преимущественно на объектную классификацию, при которой категория определяется присутствием объектов на изображении. На основе таких наборов данных разработаны современные архитектуры глубокого обучения, демонстрирующие высокое качество распознавания изображений [7].

Вместе с тем в ряде прикладных задач объектный состав изображения оказывается недостаточным для определения его категории. В качестве примера можно рассмотреть изображения профилей пользователей

социальных сетей, используемые в настоящей работе. Фотография человека с собакой может быть отнесена к категории «портрет владельца», «портрет с питомцем» или «изображение животного» в зависимости от композиции сцены и соотношения объектов.

Подобные ситуации отражают проблему семантического разрыва между объектными признаками, формируемыми нейросетями, предобученными на датасетах типа ImageNet, и признаками, необходимыми для функциональной классификации изображений. Например, категории «портрет» и «портрет с питомцем» могут содержать одинаковый набор объектов, но различаться их ролью в композиции изображения.

Термин «семантический разрыв» (semantic gap) был введён в области поиска изображений по содержанию: в работе [1] он определён как несоответствие между информацией, которую можно извлечь из визуальных данных, и той интерпретацией, которую эти же данные получают у пользователя в конкретной ситуации. В настоящей работе это понятие конкретизируется применительно к задаче функциональной классификации: под семантическим разрывом понимается расхождение между признаковым описанием изображения, формируемым свёрточной сетью, обученной на объектной задаче, и признаками, определяющими функциональную категорию изображения, – относительной значимостью объектов, их взаимным расположением и композицией сцены.

Большинство существующих исследований посвящено задачам объектной классификации и детекции объектов. В работе [8] рассмотрена эволюция методов детекции от статистических критериев к современным нейросетевым архитектурам. Вместе с тем при переходе к прикладным задачам анализа изображений существенное значение приобретают особенности предметной области и структура данных. В работе [9] показано, что качество распознавания снижается при отличии тестовых данных от обучающей выборки. Данное обстоятельство представляет интерес и для задач функциональной классификации, в которых признаки, сформированные на объектных датасетах, используются для решения задач иного типа.

Ряд работ посвящён интерпретации содержимого изображений, выходящей за пределы объектной разметки. Так, в работе [10] решается задача распознавания действий животных по изображению: категория определяется не самим присутствием объекта, а его состоянием, что сближает данную постановку с функциональной классификацией. Вопросы переносимости признаков, сформированных при обучении на объектных наборах данных, систематически исследованы в работе [11], где показано, что универсальность свёрточных признаков убывает с ростом глубины слоя, а качество переноса снижается по мере увеличения различий между исходной и целевой задачами. В работе [12] на широком наборе целевых задач установлено, что выигрыш от переноса уменьшается для задач, семантика которых отличается от объектной классификации. Эти результаты мотивируют применяемую в настоящей работе двухэтапную стратегию дообучения и постановку вопроса о границах применимости объектных признаков.

Необходимость учёта отношений между объектами при интерпретации сцены отражена в направлении, связанном с детекцией визуальных отношений и графами сцены. В работе [13] графы сцены использованы для поиска изображений, в работе [14] предложен метод детекции визуальных отношений вида «субъект – предикат – объект», а набор данных Visual Genome [15] содержит плотную разметку объектов, их атрибутов и попарных отношений. Перечисленные работы, однако, направлены на явное восстановление структуры сцены, а не на классификацию изображения в целом по его функциональной роли.

Наиболее близкой к рассматриваемой является задача распознавания намерения (intent), стоящего за публикацией изображения. В работе [16] представлен набор данных Intentonomy и показано, что категории намерений слабо выводимы из объектной разметки, а качество их распознавания существенно уступает качеству объектной классификации. В работе [17] исследована связь выбора фотографии профиля с личностными характеристиками пользователей социальных сетей. Вместе с тем задача автоматической классификации изображений профилей по функциональному назначению: с формальными критериями категорий и количественным анализом ограничений объектных признаков – насколько известно авторам, ранее систематически не рассматривалась.

Целью настоящей работы является исследование особенностей классификации изображений по функциональному назначению и анализ семантического разрыва между объектным и функциональным признаковыми пространствами.

Формально рассматриваемая постановка остаётся задачей многоклассовой классификации изображений: меняется не тип искомого отображения, а структура разметки, поскольку функциональная категория определяется объектным составом сцены и отношениями между объектами и контекстом. Научная новизна работы состоит в следующем: предложена формализация задачи функциональной классификации, в которой явно выделены объектный состав, межобъектные отношения и контекст; разработана таксономия функциональных категорий с формальными критериями их разграничения; количественно охарактеризован семантический разрыв через оценку разделимости категорий, анализ структуры ошибок классификатора, перенесённого с объектной задачи, и сравнение с подходом, использующим только объектные признаки.

В рамках исследования решаются следующие задачи:

1. формализация отличий функциональной классификации от объектной с математической постановкой;
2. разработка таксономии функциональных категорий с формальными критериями разграничения;

3. систематический анализ структуры ошибок с целью выявления источников семантического разрыва;
4. количественная оценка ограничений переноса объектных признаков.

1. Формализация задачи функциональной классификации

Пусть $\mathcal{X}$ – пространство изображений, $\mathcal{O}$ – множество объектных категорий, $\mathcal{F}$ – множество функциональных категорий. Объектная классификация определяет отображение $f_o: \mathcal{X} \to \mathcal{O}$, при котором принадлежность изображения классу определяется наличием на нём соответствующего объекта.

Функциональная классификация определяет отображение $f_f: \mathcal{X} \to \mathcal{F}$, при котором категория изображения зависит от объектного состава сцены, взаимного расположения объектов, их относительной значимости и контекста. Формально:

$$f_f(x) = g(O(x), R(x), C(x)), \tag{1}$$

где $O(x)$ – множество объектов на изображении; $R(x)$ – отношения между объектами: размеры, позиции, перекрытия; $C(x)$ – контекстные признаки: фон, освещение, стиль.

Основное отличие заключается в необходимости учёта объектов сцены $O(x)$, отношений между ними $R(x)$ и контекстных признаков $C(x)$ при определении функциональной категории изображения.

Предобученные на ImageNet свёрточные сети формируют признаковое пространство $\mathcal{Z}_o$, оптимизированное для объектной классификации. При переносе на функциональную задачу возникает семантический разрыв: признаки $z \in \mathcal{Z}_o$ могут не содержать информации, необходимой для различения функциональных категорий.

Семантический разрыв, формализованный через введённые компоненты $O(x)$, $R(x)$ и $C(x)$, проявляется в том, что признаки, оптимизированные для объектного состава, не обязательно сохраняют информацию об отношениях между объектами и контексте. В результате функциональные категории, которые различаются преимущественно компонентой $R(x)$ при одинаковом объектном составе, оказываются плохо разделимыми в признаковом пространстве. Данная ситуация характерна для категорий, различающихся ролью объектов при одинаковом наборе (см. таксономию ниже). Количественная оценка этого эффекта приведена в параграфе 3, а сравнение с детекционным подходом иллюстрирует потерю качества при использовании только объектных признаков.

В рассматриваемой работе под функциональной категорией понимается класс изображения, определяемый его ролью в процессе визуальной коммуникации пользователя. Один и тот же набор объектов может относиться к разным категориям в зависимости от их взаимного расположения, относительной значимости и композиции.

Рассмотрим пример. Пусть x_1 – изображение человека крупным планом, x_2 – изображение человека и собаки, где человек доминирует, x_3 – изображение человека и собаки, где собака доминирует. В объектном пространстве:

$$O(x_1) = \{\text{человек}\}, O(x_2) = O(x_3) = \{\text{человек,собака}\}. \tag{2}$$

Поскольку изображения x_2 и x_3 содержат одинаковый набор объектов, признаки, формируемые сетью, предобученной на объектной задаче, могут оказаться близкими, несмотря на различие их функционального назначения.

На основе анализа 1722 реальных изображений профилей социальных сетей предложена таксономия из шести категорий (табл. 1), определяемых не объектным составом, а функциональной ролью. Порог 15 % был определён в ходе предварительного анализа выборки и согласован экспертами при разработке схемы разметки.

Табл. 1. Таксономия функциональных категорий изображений

Категория	Описание	Функция	Доля в выборке
I	Индивидуальный портрет – изображение одного человека, занимающего центральную часть кадра (>50 % площади)	прямая визуальная идентификация	34,2 %
II	Групповой портрет – изображение двух и более людей	демонстрация социальной принадлежности	28,7 %
III	Комбинированный портрет – человек и животное, оба визуально значимы (животное занимает >15 % площади)	демонстрация связи с питомцем	3,3 %
IV	Изображение животного – животное без присутствия людей или с минимальным присутствием (<15 % площади)	символическая репрезентация	5,8 %
V	Непортретное изображение – отсутствие людей и животных: пейзажи, объекты, абстракции	гетерогенная категория	18,9 %
VI	Пустое изображение – стандартный шаблон платформы	тривиальная категория с характерным визуальным паттерном	9,1 %

Категории III и IV могут содержать одинаковый набор объектов, однако различаются их ролью в композиции изображения. По этой причине признаки, эффективные для объектной классификации, оказываются менее информативными при разделении функциональных категорий.

2. Метод классификации

В качестве базовой архитектуры использована сеть ResNet50 [7], являющаяся одной из наиболее распространённых архитектур для решения задач классификации изображений. Архитектура содержит 50 свёрточных слоёв, объединённых остаточными соединениями, и около 25,6 млн обучаемых параметров.

Для адаптации объектных признаков к задаче функциональной классификации применена двухэтапная стратегия переноса обучения.

На первом этапе использовались веса сети, предварительно обученной на наборе данных ImageNet. Слои conv1-conv4, отвечающие за извлечение низкоуровневых признаков (границ, текстур и локальных структур), фиксировались. Исходный классификационный слой заменялся выходным слоем размерности 6, соответствующей числу функциональных категорий. Обучение выполнялось в течение 50 эпох со скоростью обучения 10^{-3}.

На втором этапе осуществлялась адаптация высокоуровневых признаков к особенностям целевой задачи. Для этого размораживались слои conv4_x и conv5_x, содержащие десять остаточных блоков. Дообучение модели проводилось в течение 30 эпох со скоростью обучения 10^{-4}.

Используемый подход основан на предположении о различной переносимости признаков различных уровней. Низкоуровневые признаки, описывающие локальные структуры изображения, обладают высокой универсальностью и могут использоваться в различных задачах компьютерного зрения. В свою очередь, высокоуровневые представления в большей степени зависят от специфики предметной области и требуют дополнительной адаптации при переходе от объектной классификации к функциональной.

На рис. 1 представлена архитектура классификатора на основе ResNet50 с двухэтапным переносом обучения. Слои conv1–conv3 сохраняют параметры, полученные при обучении на наборе ImageNet, тогда как слои conv4–conv5 адаптируются к решаемой задаче.



Рис. 1. Архитектура классификатора на основе ResNet50 с двухэтапным переносом обучения

Выборка характеризуется существенным дисбалансом классов: отношение численности наиболее представленного класса к наименее представленному составляет 10,5:1.

Для его компенсации использовалась взвешенная функция потерь на основе кросс-энтропии:

$$L = -\sum_{i=1}^{N} w_{y_i} \log p(y_i \mid x_i), w_c = \frac{N}{C \cdot n_c}, \quad (3)$$

где N – общее число обучающих примеров; C – количество классов; n_c – число объектов класса c.

Дополнительно применялась аугментация данных, включающая горизонтальное отражение изображений, случайный поворот в диапазоне [−15°, +15°], случайное кадрирование с сохранением 80 – 100 % площади изображения, изменение яркости и контрастности в пределах ± 20 %, а также гауссово размытие.

Для миноритарных категорий III и IV дополнительно выполнялось увеличение объёма обучающей выборки путём генерации аугментированных изображений. Число обучающих примеров было увеличено в три раза. Аугментация применялась исключительно на этапе обучения, тогда как валидация и тестирование выполнялись на исходных изображениях без преобразований.

3. Экспериментальная оценка

Экспериментальные исследования проведены на выборке из 1722 изображений профилей пользователей социальной сети. Разметка выполнена двумя независимыми экспертами с последующим согласованием результатов. Коэффициент согласованности Коэна составил $\kappa = 0{,}89$, что свидетельствует о высокой согласованности экспертных оценок. Для обучения, валидации и тестирования использовалось стратифицированное разбиение данных в пропорции 70 %, 15 % и 15 % соответственно. Для всех экспериментов использовалось одинаковое стратифицированное разбиение обучающей, валидационной и тестовой выборок.

Общие результаты классификации приведены в табл. 2. Значения Accuracy, macro-F1, macro-Precision и macro-Recall составили 0,85, 0,83, 0,84 и 0,82 соответственно.

Табл. 2. Результаты по категориям

Категория	Precision	Recall	F1
I. Индивидуальный портрет	0,91	0,93	0,92
II. Групповой портрет	0,87	0,89	0,88
III. Комбинированный портрет	0,78	0,71	0,74
IV. Изображение животного	0,86	0,80	0,83
V. Непортретное	0,79	0,76	0,77
VI. Пустое	0,98	0,99	0,98

Детальный анализ ошибок представлен в виде матрицы ошибок (рис. 2).

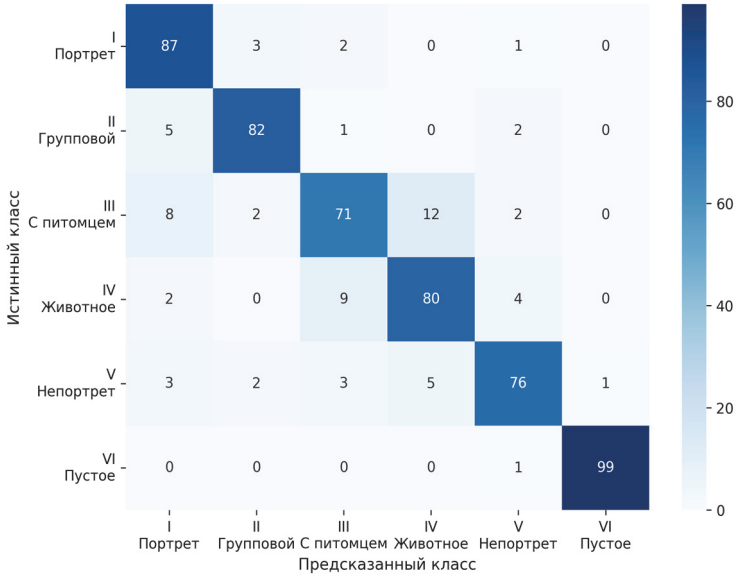


Рис. 2. Матрица ошибок классификации по функциональным категориям

Наибольшее число ошибок наблюдается между категориями III (комбинированный портрет) и IV (изображение животного): 12 % изображений категории III ошибочно классифицированы как IV, а 9 % изображений категории IV – как III. Данные категории различаются преимущественно относительной значимостью объектов в кадре, тогда как объектный состав изображений может совпадать.

Категория III также смешивается с категорией I (индивидуальный портрет): доля таких ошибок составляет 8 %. В большинстве случаев животное располагается на периферии кадра и занимает площадь, близкую к пороговому значению 15%, используемому при разметке.

Категория VI (пустое изображение) классифицируется практически без ошибок (Recall = 0,99). Стандартные изображения-заглушки имеют устойчивую визуальную структуру и практически не пересекаются с другими категориями.

Сравнение методов (табл. 3) показывает преимущество глубоких нейронных сетей по сравнению с классическими подходами. Точность метода SVM с признаками HOG составляет 0,61. Для архитектур VGG16, EfficientNet-B0 и ResNet50 данный показатель находится в диапазоне 0,81–0,85. Наилучшие результаты получены для модели ResNet50 (Accuracy = 0,85, F1 = 0,83).

Табл. 3. Сравнение методов

Метод	Accuracy	F1
SVM, HOG	0,61	0,58
VGG16	0,81	0,79
EfficientNet-B0	0,84	0,82
ResNet50	0,85	0,83

Для сравнения с нейросетевой моделью дополнительно рассмотрен подход, основанный на явной детекции объектов. В качестве базового метода использовался детектор YOLOv8, предобученный на наборе данных COCO. Детекция выполнялась с порогом уверенности 0,25 и порогом IoU 0,45. Для каждого изображения определялись количество обнаруженных людей и животных, а также относительные площади соответствующих ограничивающих рамок.

На основе результатов детекции выполнялась классификация изображений в соответствии с критериями, приведёнными в табл. 1. Категория I присваивалась при обнаружении одного человека, категория II – при обнаружении двух и более людей. Категория III соответствовала случаям одновременного присутствия человека и животного, занимающего не менее 15 % площади изображения. Категория IV присваивалась изображениям, в которых животное являлось доминирующим объектом сцены. Категория V включала изображения без обнаруженных людей и животных. Категория VI определялась по наличию стандартного изображения-заглушки платформы.

Оценка выполнялась на той же тестовой выборке, что и для нейросетевой модели. Результаты сравнения представлены в табл. 4.

Accuracy и F1 для подхода на основе детекции объектов составили 0,72 и 0,68 соответственно, тогда как для ResNet50 данные показатели достигли 0,85 и 0,83.

Табл. 4. Сравнение классификации по объектным признакам и признакам изображения

Метод	Accuracy	F1
YOLO с экспертными правилами	0,72	0,68
ResNet50	0,85	0,83

Среди 39 ошибок тестовой выборки выделены три основные группы.

Первая группа связана с пограничными соотношениями объектов и включает 26 ошибок (67 %). К ней относятся изображения, в которых отношение площадей объектов «человек – животное» близко к пороговому значению, используемому при разметке. Например, человек занимает 52 % площади кадра, а собака – 18 %. В подобных случаях изображение находится на границе между категориями, а объектный состав оказывается недостаточным для однозначного определения функциональной категории.

Вторая группа включает 8 ошибок (21 %) и связана с кадрированием изображений. На групповых фотографиях в кадре может полностью присутствовать только одно лицо, тогда как остальные люди представлены фрагментарно. В результате изображение ошибочно относится к категории индивидуальных портретов.

Третья группа содержит 5 ошибок (13 %) и связана с семантической неоднозначностью изображений. В неё входят рисунки, игрушки и другие визуальные объекты, имеющие сходство с изображениями животных. Подобные изображения нередко ошибочно относятся к категории животных.

Наибольшая доля ошибок связана с изображениями, для которых категория определяется соотношением объектов в кадре. Объектный состав изображения при этом оказывается недостаточным для однозначной классификации.

Далее рассмотрены характерные примеры изображений, относящихся к данной группе (рис. 3).

Поскольку исходные изображения содержат персональные данные пользователей социальной сети, на рис. 3 приведены синтезированные генеративной моделью изображения, воспроизводящие композиционные характеристики реальных примеров выборки: состав объектов, соотношение занимаемых ими площадей и взаимное расположение.



Рис. 3. Изображения с пограничным соотношением объектов

Пример 1. Изображение молодой женщины с кошкой на руках. Площадь лица – 48 %, площадь кошки – 17 %. Истинная категория: III (комбинированный портрет). Предсказание модели: I (индивидуальный портрет). Соотношение площадей близко к пороговому значению, при этом кошка частично перекрыта руками и занимает относительно небольшую часть изображения.

Пример 2. Изображение человека с крупной собакой на поводке. Площадь человека – 35 %, площадь собаки – 22 %. Истинная категория: III. Предсказание: IV (изображение животного). Животное занимает значительную часть кадра и располагается на переднем плане, что приводит к отнесению изображения к категории IV.

Пример 3. Групповое фото трёх человек, у одного из которых на руках находится небольшая собака. Истинная категория: II (групповой портрет). Предсказание: III (комбинированный портрет). Несмотря на наличие нескольких людей в кадре, присутствие животного приводит к ошибочному отнесению изображения к категории III.

Для визуализации семантического разрыва извлечены признаки предпоследнего слоя сети размерности 2048 и выполнено их отображение в двумерное пространство методом t-SNE. Результат представлен на рис. 4.

Визуализация признакового пространства методом t-SNE показывает различия в распределении функциональных категорий. Категории I и II формируют компактные и хорошо разделённые кластеры, поскольку различаются объектным составом сцены. Категории III и IV, напротив, частично перекрываются в признаковом

пространстве. Обе категории содержат изображения животных, поэтому их признаки оказываются близкими. Основное различие между ними связано не с набором объектов, а с их относительной значимостью в композиции изображения.

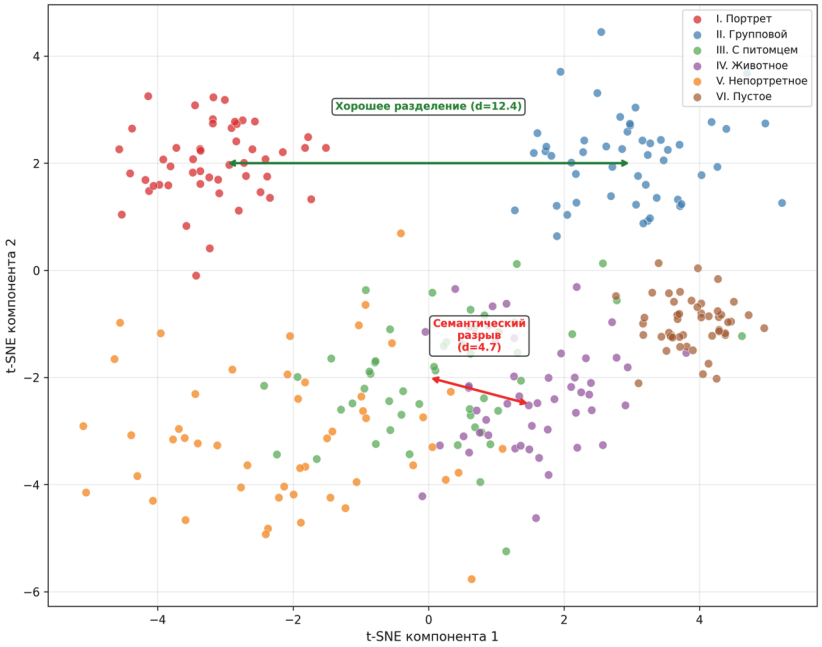


Рис. 4. Визуализация признакового пространства

Для количественной оценки разделимости категорий вычислялись центроиды кластеров в двумерном пространстве t-SNE. Межклассовое расстояние определялось как евклидово расстояние между соответствующими центроидами.

Межклассовое расстояние между центроидами кластеров для категорий I–II составляет 12,4, тогда как для категорий III–IV – 4,7. Таким образом, категории, различающиеся отношениями между объектами, разделяются в признаковом пространстве существенно хуже, чем категории, различающиеся объектным составом.

Для оценки вклада отдельных компонентов метода выполнено последовательное исключение элементов модели, результаты которого приведены в табл. 5.

Табл. 5. Вклад компонентов

Конфигурация	**Accuracy**	**F1**	**Δ**
Полная модель	0,85	0,83	–
Только первый этап обучения	0,78	0,72	-0,07
Без аугментации миноритарных классов	0,81	0,74	-0,04
Без взвешивания классов	0,82	0,76	-0,03
VGG16	0,81	0,79	-0,04
Обучение с нуля (без ImageNet)	0,62	0,55	-0,23

Наибольшее снижение качества наблюдается при отказе от переноса обучения. Обучение модели ResNet50 с нуля приводит к уменьшению точности с 0,85 до 0,62. При объёме выборки 1722 изображения использование предварительно обученной модели существенно влияет на итоговое качество классификации.

Исключение второго этапа обучения, связанного с дообучением блоков conv4_x–conv5_x, снижает точность до 0,78. Полученный результат показывает, что адаптация высокоуровневых признаков к функциональной задаче вносит заметный вклад в качество классификации.

Отказ от аугментации миноритарных категорий уменьшает точность до 0,81, а исключение взвешенной функции потерь – до 0,82. Таким образом, оба механизма компенсации дисбаланса классов оказывают сопоставимое влияние на результат.

Замена ResNet50 на архитектуру VGG16 также приводит к снижению качества классификации до 0,81. Наилучшие результаты получены для архитектуры ResNet50, использующей остаточные соединения и двухэтапную стратегию переноса обучения.

4. Обсуждение

Несмотря на достигнутую точность 85 %, большинство ошибок связано с категориями, различающимися не набором объектов, а их ролью в композиции сцены. Среди 39 ошибочных классификаций 26 случаев (67 %) относятся к изображениям с пограничными соотношениями объектов.

При использовании детектора YOLO и экспертных правил значение Accuracy составило 0,72 вместо 0,85 для ResNet50. Наибольшие различия наблюдаются для категорий III и IV, различающихся соотношением объектов в кадре.

Отказ от предварительно обученных весов ImageNet снижает точность с 0,85 до 0,62, а исключение этапа дообучения блоков conv4_x-conv5_x – до 0,78. Без дополнительной адаптации качество классификации заметно снижается.

Полученные результаты позволяют явно охарактеризовать, как изменение постановки задачи влияет на применимость методов. Для категорий, различающихся объектным составом (I, II, V, VI), переход от объектной разметки к функциональной практически не сказывается на качестве: объектные признаки остаются информативными, что подтверждается значениями F1 от 0,77 до 0,98 и большим межклассовым расстоянием в признаковом пространстве (12,4 для пары I-II). Для категорий, различающихся отношениями между объектами при совпадающем объектном составе (III, IV), наблюдается систематическое снижение качества для всех рассмотренных методов: F1 категории III (0,74) минимален по таксономии, межклассовое расстояние сокращается до 4,7, а подход на основе явной детекции объектов теряет наибольшую долю качества именно на этих категориях. Таким образом, семантический разрыв проявляется как локализованная деградация на подмножестве категорий, разметка которых зависит от компоненты R(x).

Данная интерпретация уточняет и границы применимости стандартного переноса обучения. Дообучение высокоуровневых слоёв частично компенсирует разрыв (прирост точности с 0,78 до 0,85), поскольку свёрточные признаки неявно кодируют пространственную информацию о сцене. Однако сохраняющееся перекрытие кластеров категорий III и IV показывает, что неявного кодирования отношений недостаточно: для дальнейшего повышения качества требуются модели, использующие явное представление структуры сцены, например графы сцены [13 – 15].

К ограничениям проведённого исследования относится объём выборки (1722 изображения) и использование одного предметного домена: изображений профилей пользователей социальных сетей. Проверка полученных выводов на других типах данных требует дополнительного исследования.

Дальнейшее развитие работы может быть связано с использованием механизмов внимания и методов явного моделирования структуры сцены. Особый интерес представляет учёт пространственных отношений между объектами, включая их относительные размеры, взаимное расположение и степень перекрытия. Другим направлением является построение моделей, использующих результаты предварительной детекции объектов и их конфигурацию при принятии решения о функциональной категории изображения.

Заключение

В работе исследована задача классификации изображений по функциональному назначению на примере изображений профилей пользователей социальных сетей. Предложена таксономия из шести функциональных категорий и формальная постановка задачи, учитывающая объектный состав сцены, отношения между объектами и контекстные признаки.

Экспериментальные результаты показали, что объектный состав изображения не всегда достаточен для определения его функциональной категории. Сравнение нейросетевой модели с подходом, основанным на детекции объектов и экспертных правилах, выявило снижение качества классификации при использовании только объектной информации.

Полученные результаты могут быть использованы при разработке методов анализа изображений, ориентированных на учёт структуры сцены и отношений между объектами. Перспективным направлением дальнейших исследований является построение моделей, использующих явное представление пространственных отношений между объектами.

Благодарности

Исследование выполнено за счет средств государственного задания Министерства науки и высшего образования Российской Федерации, шифр проекта FSSS-2026-0003.

References

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22(12): 1349-1380. DOI: 10.1109/34.895972.

[2] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 2008; 40(2): 5. DOI: 10.1145/1348246.1348248.

[3] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018; 40(6): 1452-1464. DOI: 10.1109/TPAMI.2017.2723009.

[4] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009: 248-255. DOI: 10.1109/CVPR.2009.5206848.

[5] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report. Toronto, Canada: University of Toronto; 2009.

[6] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014. Lecture Notes in Computer Science. Vol 8693. Springer; 2014: 740-755. DOI: 10.1007/978-3-319-10602-1_48.
[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2016: 770-778. DOI: 10.1109/CVPR.2016.90.
[8] Andriyanov NA, Dementiev VE, Tashlinsky AG. Object detection in images: From Bayes and Neyman-Pearson criteria to neural network-based EfficientDet detectors [In Russian]. Computer Optics 2022; 46(1): 139-159. DOI: 10.18287/2412-6179-CO-922.
[9] Sokolova AD, Savchenko AV, Nikolenko SI. Rare data search in the problem of face recognition in images [In Russian]. Computer Optics 2022; 46(5): 801-807. DOI: 10.18287/2412-6179-CO-1061.
[10] Egorov AD, Reznik MS. Interpretation of animal actions from their image in near real time [In Russian]. Computer Optics 2023; 47(2): 278-286. DOI: 10.18287/2412-6179-CO-1138.
[11] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. Vol 27. MIT Press; 2014: 3320-3328.
[12] Kornblith S, Shlens J, Le QV. Do better ImageNet models transfer better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2019: 2661-2671. DOI: 10.1109/CVPR.2019.00277.
[13] Johnson J, Krishna R, Stark M, Li LJ, Shamma DA, Bernstein MS, Fei-Fei L. Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2015: 3668-3678. DOI: 10.1109/CVPR.2015.7298990.
[14] Lu C, Krishna R, Bernstein M, Fei-Fei L. Visual relationship detection with language priors. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science. Vol 9905. Springer; 2016: 852-869. DOI: 10.1007/978-3-319-46448-0_51.
[15] Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 2017; 123(1): 32-73. DOI: 10.1007/s11263-016-0981-7.
[16] Jia M, Wu Z, Reiter A, Cardie C, Belongie S, Lim SN. Intentonomy: A dataset and study towards human intent understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2021: 12986-12996.
[17] Liu L, Preotiuc-Pietro D, Samani ZR, Moghaddam ME, Ungar L. Analyzing personality through social media profile picture choice. In: Proceedings of the Tenth International AAAI Conference on Web and Social Media. AAAI Press; 2016: 211-220.
[18] Lang P, inventor; Imaging Therapeutics Inc, assignee. Methods and devices for quantitative analysis of X-ray images. U.S. Patent 7,050,534. August 20, 2002.

Сведения об авторах

Жданова Александра Николаевна, 1986 года рождения, в 2012 году защитила диссертацию на соискание учёной степени кандидата технических наук по специальности 05.13.10 «Управление в социальных и экономических системах». В настоящее время является доцентом кафедры программных систем Самарского национального исследовательского университета имени академика С.П. Королёва. Область научных интересов: интеллектуальный анализ данных, машинное обучение и глубокое обучение, анализ и интерпретация мультимодальных данных, нейросимвольные методы, графы знаний и объяснимый искусственный интеллект. E-mail: *aleksandra.zhd@yandex.ru*

Куприянов Александр Викторович, 1978 года рождения, в 2004 году получил учёную степень кандидата технических наук, а в 2013 году – доктора технических наук. В настоящее время является директором Института информатики и кибернетики Самарского национального исследовательского университета имени академика С.П. Королёва, заведующим кафедрой технической кибернетики и ведущим научным сотрудником Института систем обработки изображений НИЦ «Курчатовский институт». Область научных интересов: обработка цифровых сигналов и изображений, распознавание образов и искусственный интеллект, анализ и интерпретация наноразмерных изображений, биомедицинская визуализация и анализ изображений. E-mail: *akupr@ipsiras.ru*

Поступила в редакцию 25 июня 2026 г. *Окончательный вариант – 10 июля 2026 г.*

Semantic gap analysis in image classification by functional purpose

A.N. Zhdanova [1,], A.V. Kupriyanov [1,2]*
[1] Samara National Research University, Moskovskoye Shosse 34, Samara, 443086, Russia;
[2] Image Processing Systems Institute, NRC "Kurchatov Institute",
Molodogvardeiskaya Str. 151, Samara, 443001, Russia;
** Corresponding Author*

Abstract

The paper addresses the problem of image classification by functional purpose, which differs from traditional object classification in that the category is determined by the functional role of the image and depends on the composition of scene objects, their spatial arrangement, and contextual information. The study investigates the semantic gap between the object feature space formed by convolutional neural networks pretrained on object-centered datasets and the space of functional categories. A taxonomy of functional image categories is proposed, together with formal classification criteria based on the relative importance of objects and their spatial relationships.

Experimental evaluation was conducted on a dataset of 1,722 real social media profile images annotated by experts (Cohen's kappa coefficient $\kappa = 0.89$). Using a ResNet50 architecture with a two-stage transfer learning strategy, classification performance with an accuracy of 0.85 and an F1 score of 0.83 was achieved. For comparison, an object-based approach using YOLOv8 object detection combined with expert-defined rules was also evaluated, yielding an Accuracy of 0.72. Analysis of classification errors showed that 67% of misclassifications were associated with borderline cases in which the object composition alone was insufficient to determine the functional category unambiguously.

Feature space analysis revealed reduced separability of categories distinguished by relationships between objects compared with categories distinguished by object composition. The obtained results indicate the limitations of directly transferring object-level features to functional image classification tasks and highlight the importance of incorporating object relationships into image analysis models.

Keywords: image classification; functional categories, semantic gap, transfer learning, convolutional neural networks, feature space.

Citation: Zhdanova AN, Kupriyanov AV. Semantic gap analysis in image classification by functional purpose. Computer Optics 2026; 50(5): 2079. doi: 10.18287/COJ2079.

Acknowledgements: The research was carried out within the state assignment theme FSSS-2026-0003.

About authors

Aleksandra Nikolaevna Zhdanova (b. 1986) received the Candidate of Engineering Sciences degree in 2012 in the specialty 05.13.10 – Control in Social and Economic Systems. She is currently an Associate Professor at the Department of Software Systems, Samara National Research University (Samara University). Her research interests include data mining, machine learning and deep learning, multimodal data analysis and interpretation, neuro-symbolic methods, knowledge graphs, and explainable artificial intelligence. E-mail: *aleksandra.zhd@yandex.ru*

Alexander Victorovich Kupriyanov (b. 1978) received the Candidate of Engineering Sciences degree in 2004 and the Doctor of Engineering Sciences degree in 2013. He is currently Director of the Institute of Informatics and Cybernetics and Head of the Department of Technical Cybernetics at Samara National Research University, and Leading Researcher at the Image Processing Systems Institute, NRC "Kurchatov Institute". His research interests include digital signal and image processing, pattern recognition and artificial intelligence, nanoscale image analysis and understanding, and biomedical image analysis. E-mail: *akupr@ipsiras.ru*

Received on June 25, 2026. *The final version – July 10, 2026.*