# Обучение алгоритма семантической сегментации изображений на выборке с разнообразными типами аннотаций

Р. В. Шаповалов, Д. П. Ветров, А. А. Осокин, П. Коли

В традиционной постановке задача семантической сегментации изображений использует обучающую выборку изображений, размеченных попиксельно. Получение такой разметки требует значительных человеческих усилий. Предлагается метод обучения семантической сегментации, позволяющий использовать менее подробную информацию, получение которой на практике требует меньше усилий, например, плотные рамки вокруг объектов на изображении или множество уникальных меток изображения.

**Ключевые слова:** машинное обучение, структурный метод опорных векторов, функция потерь, семантическая сегментация изображений.

## 1. Введение

Многоклассовая семантическая сегментация изображений заключается в том, чтобы каждому пикселю изображения сопоставить метку категории из заранее определённого множества (результат сегментации изображения на рис. 1а приведён на рис. 1б). Семантическая сегментация — одна из фундаментальных задач компьютерного зрения, поскольку к ней сводятся другие важные задачи. Если известны маски всех объектов на изображении, то становятся тривиальными задачи, например, детектирования объектов определённой категории или их подсчёта. С другой стороны, получение семантической сегментации в явном виде требуется в прикладных задачах, таких как автономная навигация автомобилей [1], оценка позы человека [2] или восстановление трёхмерной структуры сцены [3].

При обучении алгоритма сегментации сложность представляет разметка изображений обучающей выборки — она требует значительных человеческих усилий. В отличие от полной (*сильной*) разметки, гораздо проще получить *слабую аннотацию* изображения, под которой мы понимаем некоторую статистику от полной разметки. Примерами слабых аннотаций служат метки изображения, которые отражают присутствие или отсутствие категорий; метки площади, которые содержат число пикселей каждой категории на изображении; набор плотных рамок для объектов, присутствующих в разметке; а также набор семян — подмножеств координат пикселей, принадлежащих объектам (рис. 1). Например, в наборе данных PASCAL VOC 2012 только 2913 из 11540 (25%) изображений размечены полностью, для остальных известны только плотные рамки некоторых категорий объектов. Даже если процесс разметки можно контролировать, имеет смысл использовать различные типы аннотаций, поскольку они лучше характеризуют различные семантические категории. Например, категории-объекты (такие как *знак*, *корова*, *автомобиль*) хорошо описываются рамками, а категории-фон (*небо*, *трава*, *вода*), которые обычно занимают значительную часть изображения, — метками изображения.

В литературе описаны методы, которые используют слабые аннотации для обучения семантической сегментации, но большинство из них используют только метки изображения. Например, Вежневец и др. [4, 5] используют вероятностную графическую модель над набором изображений, чтобы распространять информацию о предполагаемой разметке между изображениями. В этой статье мы представляем метод для обучения семантической сегментации по смеси сильно- и слабоаннотированных изображений. Метод позволяет учитывать разные типы слабой аннотации, даже в рамках одного изображения.

Работа базируется на недавних исследованиях по использованию метода опорных векторов с латентными переменными (*latent-variable structural support vector machines, LV-SSVM*) для задач обучения со слабым наблюдением [6, 7, 8]. В отличие от них, наш метод использует специализированные функции потерь, которые измеряют рассогласованность разметки, предсказанной алгоритмом, с верной (возможно, слабой) аннотацией данного изображения. Мы определяем эти функции потерь так, чтобы они оценивали матожидание расстояния Хэм-

(а) Изображение.


(б) Полная разметка.

(в) Аннотация с помощью рамки.

(г) Аннотация с помощью семени.

небо самолёт дерево трава

(д) Аннотация метками изображения.

Рис. 1. Различные типы аннотаций для изображения из набора данных MSRC.

минга от разметки, предсказанной алгоритмом, до разметок, удовлетворяющих слабой аннотации изображения. Благодаря такому определению, функции, специализированные для разных типов аннотаций, определены в одном масштабе. Таким образом, наш метод содержит только один гиперпараметр, который регулирует относительный вклад полностью размеченных и слабоаннотированных данных. Он необходим, поскольку последние обычно менее информативны. Мы эмпирически покажем, как балансирование этого параметра может улучшить качество сегментации.

Для того чтобы обучить LV-SSVM с использованием различных типов аннотаций, необходимо определить специализированные функ-

ции потерь. Для введённых функций потерь необходимо описать алгоритмы *вывода, дополненного функцией потерь* и *вывода, согласованного с аннотацией*. Первый алгоритм выводит разметку изображения, высоко ранжируемую текущей моделью, но при этом сильно отличающуюся от верной аннотации, а второй выводит разметку, высоко ранжируемую текущей моделью, при этом согласующуюся с верной аннотацией (для слабых аннотаций существует множество разметок, согласующихся с ними). Мы покажем, как решать эти оптимизационные задачи для различных функций потерь, используя эффективные комбинаторные алгоритмы, основанные на разрезах в графах.

**Связь с предыдущими исследованиями.** Наша работа тесно связана со статьёй Кумара и др. [7], которые использовали пошаговый метод обучения семантической сегментации по изображениям с различными типами аннотаций. Их метод сначала обучает LV-SSVM, использующий функции потерь, определённые для частичных разметок (один из видов слабых аннотаций). При этом вывод, дополненный функцией потерь производится с помощью алгоритма итеративного обновления мод условных распределений ($ICM$), который сходится лишь к локальному минимуму, поэтому требует качественной инициализации. С помощью обученной модели выводятся частичные разметки для слабоаннотированных изображений, согласованные с их аннотациями, заданными в виде рамок или меток изображения. Модель затем дообучается, при этом выведенные частичные разметки рассматриваются как верные частичные разметки для этой части выборки. В отличие от Кумара и др. [7] мы используем специализированные функции потерь для различных типов аннотаций, которые минимизируются одновременно при обучении. Таким образом, наш метод не нуждается в «загрузочных» обучающих данных вроде частично размеченных изображений. Кроме того, мы определяем функции потерь так, что возможно использовать эффективные алгоритмы вывода, вместо использования эвристики ICM. Также мы описываем несколько другие типы слабых аннотаций.

Функции потерь, используемые нами, не всегда допускают вывод, дополненный функцией потерь, который декомпозируется на индивидуальные переменные. Аналогичные задачи решаются также в

недавних исследованиях по обучению на сильной разметке с недекомпозируемыми функциями потерь [9, 10]. Плетшер и Коли [9] используют функции потерь со слагаемыми высокого порядка, которые штрафуют разницу в площади целевой категории для бинарных разметок. Они показывают, как использовать разрезы на графах для эффективного вывода, дополненного функцией потерь. Тарлоу и Цемель [10] используют вывод с помощью передачи сообщений при обучении структурного метода опорных векторов для трёх различных функций потерь, штрафующих: отклонение от разметки, зависящее от площади (*PASCAL VOC loss*), недостаточную полноту рамки и сильное нарушение примерной границы сегментации.

**Новизна работы заключается в следующем:**

- мы предлагаем метод обучения семантической сегментации изображений, основанный на LV-SSVM, который минимизирует различные функции потерь, специализированные для различных видов аннотаций;
- мы определяем функции потерь для трёх популярных типов аннотаций (помимо полной разметки изображения) и их комбинаций: меток изображения, плотных рамок и семян объектов;
- мы предлагаем эффективные алгоритмы вывода, необходимые для обучения LV-SSVM с введёнными функциями потерь.

## 2. Структурный метод опорных векторов с латентными переменными

### 2.1. Структурное обучение для семантической сегментации

В этой подсекции мы формально определим задачу структурного обучения на основе максимизации отступа и покажем, как она применяется к задаче семантической сегментации изображений.

Пусть $\mathcal{X}$ — некоторое пространство наблюдаемых признаков, а $\mathcal{Y}$ — пространство ненаблюдаемых ответов. Как правило, в структурном обучении они описывают сложные объекты, например, целые изображения, и обладают большой размерностью.

**Определение 1.** *Дискриминантная функция* $F : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$ — функция, отражающая, насколько хорошо ответ соответствует признакам. В данной статье предполагается, что она линейно зависит от $d$-мерного вектора параметров (весов) $\mathbf{w}$: $F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^\intercal \Psi(\mathbf{x}, \mathbf{y})$, где вектор $\Psi(\mathbf{x}, \mathbf{y})$ обозначает так называемые обобщённые признаки объекта $\mathbf{x} \in \mathcal{X}$ и разметки $\mathbf{y} \in \mathcal{Y}$. $\Psi(\mathbf{x}, \mathbf{y})$ определяется в соответствии с предметной областью, а веса $\mathbf{w}$ настраиваются по обучающим данным.

**Определение 2.** *Структурным классификатором* $\mathsf{H} : \mathcal{X} \to \mathcal{Y}$ назовём функционал, представимый в виде $\mathsf{H}(\mathbf{x}; \mathbf{w}) = \arg\max\limits_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$.

Задача структурного обучения заключается в том, чтобы настроить наиболее подходящие параметры $\mathbf{w}$ функционала $\mathsf{H}$ на заданной обучающей выборке: $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, $\mathbf{x}_n \in \mathcal{X}_n, \mathbf{y}_n \in \mathcal{Y}_n$. Одной из наиболее часто используемых формализаций является структурное обучение на основе максимизации отступа (также известное как структурный метод опорных векторов, *SSVM*) [11, 12, 13].

**Оптимизационная задача 1 (структурный метод опорных векторов).**

$$\min_{\mathbf{w}, \boldsymbol{\xi} \geqslant \mathbf{0}} \ \frac{1}{2}\mathbf{w}^\intercal \mathbf{w} + \frac{C}{N}\sum_{n=1}^{N} \xi_n, \tag{2.1}$$

$$F(\mathbf{x}_n, \mathbf{y}_n; \mathbf{w}) \geqslant \max_{\bar{\mathbf{y}} \in \mathcal{Y}_n} \left(F(\mathbf{x}_n, \bar{\mathbf{y}}; \mathbf{w}) + \Delta(\bar{\mathbf{y}}, \mathbf{y}_n)\right) - \xi_n, \quad \forall n \in \{1, \ldots, N\}. \tag{2.2}$$

Здесь $\Delta(\bar{\mathbf{y}}, \mathbf{y}_n)$ — *функция потерь*, задающая степень удалённости некоторого ответа $\bar{\mathbf{y}} \in \mathcal{Y}_n$ от верного ответа $\mathbf{y}_n \in \mathcal{Y}_n$, $\xi_n$ — неотрицательные дополнительные переменные, а $C$ — гиперпараметр, регулирующий относительный вклад функции потерь и регуляризатора. При обучении подбираются такие параметры, что функция $F(\mathbf{x}_n, \cdot)$ максимальна на ответах $\bar{\mathbf{y}}$, близких к верному, и тем меньше, чем ответ дальше от верного.

**Определение 3.** Задача максимизации, возникающая в (2.2), называется *выводом, дополненным функцией потерь* (*loss-augmented inference*).

Покажем теперь, как структурное обучение применяется к семантической сегментации изображений. Мы предполагаем, что на изображении задано разбиение пикселей на *суперпиксели* $\mathcal{V}$ — группы соседних пикселей, сходных по цвету и текстуре.

**Определение 4.** Рассмотрим дискретное изображение высоты $H$ и ширины $W$. Разбиением на суперпиксели назовём функционал $\mathsf{S} : \{1,\dots,H\} \times \{1,\dots,W\} \to \mathcal{V}$, относящий каждый пиксель к одному из суперпикселей.

Для описания пространств $\mathcal{X}, \mathcal{Y}$ и функции $F$ необходимы дополнительные построения. Рассмотрим неориентированный граф $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Его вершины $\mathcal{V}$ отождествим с суперпикселями, а рёбрами $\mathcal{E}$ соединим суперпиксели, у которых есть общая граница. Обозначим $\mathbf{x}_i \in \mathbb{R}^d$ вектор признаков суперпикселя $i \in \mathcal{V}$, $\mathbf{x}_{ij} \in \mathbb{R}^e$ — вектор признаков, описывающий сходство соседних суперпикселей $i$ и $j$, а $\mathbf{x} = \bigoplus_{i\in\mathcal{V}} \mathbf{x}_i \oplus \bigoplus_{(i,j)\in\mathcal{E}} \mathbf{x}_{ij}$ — их конкатенацию. Кроме того, каждому суперпикселю $i$ сопоставлена переменная $y_i$, которая принимает значение одной из меток категорий из множества $\mathcal{K} = \{1,\dots,K\}$. Пространство $\mathcal{X}$ содержит всевозможные признаки изображения $\mathbf{x}$, а пространство $\mathcal{Y}$ — всевозможные разметки $\mathbf{y} = \{y_i\}_{i\in\mathcal{V}}$. Допустимость разметки и признаков для конкретного изображения определяется только числом суперпикселей. Мы можем определить дискриминантную функцию $F$ следующим образом:[1]

$$F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\intercal \Psi(\mathbf{x},\mathbf{y}) = \sum_{i\in\mathcal{V}} \sum_{k=1}^{K} [y_i = k](\mathbf{x}_i^\intercal \mathbf{w}_k^{\mathrm{u}}) + \sum_{(i,j)\in\mathcal{E}} [y_i = y_j](\mathbf{x}_{ij}^\intercal \mathbf{w}^{\mathrm{p}}), \tag{2.3}$$

где $\mathbf{w} = \bigoplus_{k=1}^{K} \mathbf{w}_k^{\mathrm{u}} \oplus \mathbf{w}^{\mathrm{p}}$ — вектор параметров модели, $\mathbf{w}_k^{\mathrm{u}} \in \mathbb{R}^d$, $\mathbf{w}^{\mathrm{p}} \in \mathbb{R}^e$. Слагаемые в первой и второй суммах называются унарными и парными потенциалами, соответственно. Мы полагаем парные веса $\mathbf{w}^{\mathrm{p}}$ и парные признаки $\mathbf{x}_{ij}$ неотрицательными числами, и таким образом получаем *ассоциативную* дискриминантную функцию [11]. В этом случае задача вычисления $\mathsf{H}(\mathbf{x})$, то есть максимизации $F(\mathbf{x},\mathbf{y};\mathbf{w})$ по $\mathbf{y}$, хотя и является NP-трудной, может быть эф-

[1] Здесь используются скобки Айверсона: $[e] = 1$, если логическое выражение $e$ верно, и $[e] = 0$ иначе

фективно решена приближённо, например с помощью алгоритма $\alpha$-расширения [14].

В задаче сегментации в качестве функции потерь часто используется расстояние Хэмминга (число неправильно распознанных пикселей):

$$\Delta(\bar{\mathbf{y}}, \mathbf{y}_n) = \sum_{i \in \mathcal{V}_n} c_i^n [\bar{y}_i \neq y_i^n], \text{[2]} \tag{2.4}$$

где $c_i^n$ — площадь $i$-го суперпикселя $n$-го изображения. Эта функция потерь декомпозируется по переменным. Это значит, что вывод, дополненный функцией потерь, вычислительно не сложнее, чем максимизация дискриминантной функции $F(\mathbf{x}, \mathbf{y}; \mathbf{w})$ и так же может быть выполнен с помощью $\alpha$-расширения. Известны также некоторые частные случаи функций потерь высоких порядков (то есть не декомпозирующихся на функции от переменных или их пар), которые допускают эффективный приближённый вывод [9, 10, 15].

Оптимизационная задача 1 выпукла и может быть решена, например, методом секущей плоскости [12, 13]. В этом методе ограничения (2.2) заменяются экспоненциально большим числом линейных ограничений, далее на каждой итерации допустимый политоп приближается с помощью добавления к нему самого нарушаемого ограничения, которое находится с помощью вывода, дополненного функцией потерь в (2.2).

## 2.2. Обучение со слабыми аннотациями

Рассмотрим случай, когда помимо $N$ полностью размеченных изображений, обучающая выборка содержит $M$ слабо аннотированных: $\{(\mathbf{x}_m, \mathbf{z}_m)\}_{m=N+1}^{N+M}$. Мы предполагаем, что слабая аннотация изображения однозначно определяет подмножество полных разметок, совместных с ней ($\zeta(\mathbf{z}_m) \subseteq \mathcal{Y}_m$), и таким образом менее информативна, чем неизвестная полная разметка $\mathbf{y}_m$. Примерами слабых

[2] На практике в разметке суперпикселя может встретиться несколько меток (такие суперпиксели называют *гетерогенными*). В этом случае функция потерь также равна числу неверно распознанных пикселей. Чтобы не загромождать нотацию, мы рассматриваем только гомогенные суперпиксели. Вывод тривиально обобщается на гетерогенный случай

аннотаций в задаче семантической сегментации служат: 1) плотные рамки сегментов данной метки; 2) значение некоторой глобальной статистики (площадь, средняя интенсивность, число связных компонент и т. д.) для сегментов данной метки; 3) подмножество пикселей, принадлежащих данной метке (семена). Мы далее обобщаем стандартную формулировку SSVM на случай присутствия в обучающей выборке полностью размеченных и слабоаннотированных данных.

**Оптимизационная задача 2 (обобщённый SSVM).**

$$\min_{\mathbf{w},\boldsymbol{\xi}\geqslant\mathbf{0},\boldsymbol{\eta}\geqslant\mathbf{0}} \frac{1}{2}\mathbf{w}^{\intercal}\mathbf{w} + \frac{C}{N+M}\left(\sum_{n=1}^{N}\xi_n + \alpha\sum_{m=1}^{M}\eta_m\right), \tag{2.5}$$

$$F(\mathbf{x}_n,\mathbf{y}_n;\mathbf{w}) \geqslant \max_{\bar{\mathbf{y}}\in\mathcal{Y}_n}\left(F(\mathbf{x}_n,\bar{\mathbf{y}};\mathbf{w}) + \Delta(\bar{\mathbf{y}},\mathbf{y}_n)\right) - \xi_n, \quad \forall n, \tag{2.6}$$

$$\max_{\mathbf{y}\in\zeta(\mathbf{z}_m)} F(\mathbf{x}_m,\mathbf{y};\mathbf{w}) \geqslant \max_{\bar{\mathbf{y}}\in\mathcal{Y}_m}\left(F(\mathbf{x}_m,\bar{\mathbf{y}};\mathbf{w}) + \mathrm{K}(\bar{\mathbf{y}},\mathbf{z}_m)\right) - \eta_m, \forall m. \tag{2.7}$$

Здесь $\mathrm{K}(\bar{\mathbf{y}},\mathbf{z}_m)$ — *слабая функция потерь*, задающая степень несогласованности некоторого ответа $\bar{\mathbf{y}} \in \mathcal{Y}_m$ со слабой аннотацией $\mathbf{z}_m$, $\eta_m$ — неотрицательные дополнительные переменные.

**Определение 5.** Задача максимизации $\max_{\mathbf{y}\in\zeta(\mathbf{z})} F(\mathbf{x}_n,\mathbf{y};\mathbf{w})$ на множестве, ограниченном аннотацией $\mathbf{z}$, возникающая в левой части (2.7), называется *выводом, согласованным с аннотацией* (*annotation-consistent inference*).

Заметим, что при $M = 0$ оптимизационная задача 2 сводится к стандартной постановке SSVM, а при $N = 0$ это частный случай SSVM с латентными переменными (*LV-SSVM*) [16]. Заметим также, что полная разметка $\mathbf{y}_n$ является вырожденным случаем слабой аннотации, где $\zeta(\mathbf{z}_n) = \{\mathbf{y}_n\}$. Таким образом, оптимизационная задача 2 эквивалентна LV-SSVM, с тем исключением что она содержит балансирующий коэффициент $\alpha$.

Оптимизационная задача (2.5)–(2.7) невыпукла. Следуя Йу и Йоахимсу [16], мы используем специфическую структуру задачи — сумму выпуклой и вогнутой функции. Это позволяет применить вогнуто-выпуклую процедуру (*concave-convex procedure, CCCP*) [17] для её приближённого решения. При этом помимо вывода, дополненного функцией потерь в (2.6), необходимо также эффективно выполнять вывод, дополненный слабой функцией потерь в (2.7), а также вывод,

согласованный с аннотацией в левой части (2.7). Последние две задачи зависят от используемого типа аннотаций. В следующей секции описаны конкретные алгоритмы для трёх типов аннотаций.

## 3. Использование различных типов слабых аннотаций

Чтобы использовать конкретный вид слабой аннотации при обучении, необходимо определить функцию потерь для данного типа аннотации, которая допускает эффективный вывод, дополненный функций потерь и вывод, согласованный с аннотацией. Первый должен быть очень эффективным, поскольку он вызывается на каждой итерации обучения, и, как правило, является основным источником вычислительной сложности. Мы определим их для трёх типов слабых аннотаций и покажем, как их комбинировать.

### 3.1. Метки изображения

**Определение 6.** Назовём *сильной функцией потерь по меткам изображения* следующую функцию:

$$\Delta_{\text{il}}(\bar{\mathbf{y}}, \mathbf{y}) = \sum_{i \in \mathcal{V}} c_i[\nexists j \in \mathcal{V} : y_j = \bar{y}_i \vee \nexists j \in \mathcal{V} : \bar{y}_j = y_i]. \qquad (3.1)$$

Эта функция штрафует суперпиксели, помеченные метками, которых нет в $\mathbf{y}$, а также суперпиксели, верные метки которых не присутствуют в $\bar{\mathbf{y}}$.

**Определение 7.** *Метками изображения* называется множество $\mathbf{z} \subset \mathcal{K}$ меток категорий, присутствующих на изображении. Пусть $\mathbf{y}$ — разметка изображения, тогда уникальные метки изображения $\mathbf{z} = \{y_i \mid i \in \mathcal{V}\}$ (рис. 1д).

**Определение 8.** Пусть $\mathbf{z}$ — метки изображения. Назовём *слабой функцией потерь по меткам изображения* следующую функцию, параметризованную числами $S_k$, для $k \in \mathbf{z}$:

$$\text{K}_{\text{il}}(\bar{\mathbf{y}}, \mathbf{z}) = \text{K}_{\text{il}}(\bar{\mathbf{y}}, \mathbf{z}; S_k) = \sum_{k \notin \mathbf{z}} \sum_{i \in \mathcal{V}} c_i[\bar{y}_i = k] + \sum_{k \in \mathbf{z}} S_k \prod_{i \in \mathcal{V}} [\bar{y}_i \neq k]. \quad (3.2)$$

**Теорема 1.** *Пусть* $\mathbf{z}$ *— множество меток категорий, присутствующих в* $\bar{\mathbf{y}}$*, а* $S_k$ *— число пикселей в каждой из них. Тогда слабая функция потерь по меткам изображения является верхней оценкой сильной с мультипликативным коэффициентом не более 2:*

$$\frac{1}{2}\mathrm{K}_{il}(\bar{\mathbf{y}}, \mathbf{z}) \leqslant \Delta_{il}(\bar{\mathbf{y}}, \mathbf{y}) \leqslant \mathrm{K}_{il}(\bar{\mathbf{y}}, \mathbf{z}). \tag{3.3}$$

**Доказательство.** Преобразуем $\mathrm{K}_{\mathrm{il}}$, учитывая определение $\mathbf{z}$:

$$\begin{aligned}\mathrm{K}_{\mathrm{il}}(\bar{\mathbf{y}}, \mathbf{z}) = \sum_{i\in\mathcal{V}} c_i[\bar{y}_i \notin \mathbf{z}] + \sum_{k\in\mathbf{z}}\sum_{i\in\mathcal{V}} c_i[y_i = k]\prod_{j\in\mathcal{V}}[\bar{y}_j \neq k] = \\ \sum_{i\in\mathcal{V}} c_i[\nexists j \in \mathcal{V} : y_j = \bar{y}_i] + \sum_{i\in\mathcal{V}} c_i[\nexists j \in \mathcal{V} : \bar{y}_j = y_i] = \\ \sum_{i\in\mathcal{V}} c_i\left([\nexists j \in \mathcal{V} : y_j = \bar{y}_i] + [\nexists j \in \mathcal{V} : \bar{y}_j = y_i]\right). \end{aligned} \tag{3.4}$$

Верность (3.3) следует из того факта, что для любых $a \in \{0, 1\}$, $b \in \{0, 1\}$ верно $\frac{1}{2}(a + b) \leqslant \max\{a, b\} \leqslant a + b$, что может быть проверено непосредственно.

На практике значение коэффициентов $S_k$ в определении слабой функции потерь неизвестно. Обозначим число пикселей изображения $S = \sum_{i\in\mathcal{V}} c_i$. Будем считать, что эта величина распределена мультиномиально над допустимыми метками классов: $\{S_k\}_{k\in\mathbf{z}} \sim \mathcal{M}(\mathbf{q}, S)$.

**Теорема 2.** *Пусть* $\hat{S}_k = q_k S$*, тогда* $\mathrm{K}_{il}(\bar{\mathbf{y}}, \mathbf{z}; \hat{S}_k) = \mathbb{E}\mathrm{K}_{il}(\bar{\mathbf{y}}, \mathbf{z}; S_k)$*, где матожидание берётся по распределению* $\{S_k\}_{k\in\mathbf{z}} \sim \mathcal{M}(\mathbf{q}, S)$*, то есть* $\hat{S}_k$ *обеспечивает несмещённую оценку слабой функции потерь.*

**Доказательство.**

$$\begin{aligned}\mathbb{E}\mathrm{K}_{\mathrm{il}}(\bar{\mathbf{y}}, \mathbf{z}; S_k) = \sum_{k\notin\mathbf{z}}\sum_{i\in\mathcal{V}} c_i[\bar{y}_i = k] + \mathbb{E}\sum_{k\in\mathbf{z}} S_k \prod_{i\in\mathcal{V}}[\bar{y}_i \neq k] = \\ = \sum_{k\notin\mathbf{z}}\sum_{i\in\mathcal{V}} c_i[\bar{y}_i = k] + \sum_{k\in\mathbf{z}} \mathbb{E}S_k \prod_{i\in\mathcal{V}}[\bar{y}_i \neq k] = \\ = \sum_{k\notin\mathbf{z}}\sum_{i\in\mathcal{V}} c_i[\bar{y}_i = k] + \sum_{k\in\mathbf{z}} \hat{S}_k \prod_{i\in\mathcal{V}}[\bar{y}_i \neq k]. \end{aligned} \tag{3.5}$$

Параметры распределения $\mathbf{q}$ могут быть оценены по полностью размеченной части выборки. Однако на практике размеченных изображений мало, и такая оценка получается неустойчивой. Поэтому мы предполагаем равномерные $\mathbf{q}$. Итак, мы используем следующую слабую функцию потерь по меткам изображений:

$$\mathrm{K}_{\mathrm{il}}(\mathbf{y}, \mathbf{z}) = \sum_{k \notin \mathbf{z}} \sum_{i \in \mathcal{V}} c_i [y_i = k] + \sum_{k \in \mathbf{z}} \frac{S}{|\mathbf{z}|} \prod_{i \in \mathcal{V}} [y_i \neq k]. \tag{3.6}$$

При заданной слабой функции потерь $\mathrm{K}_{\mathrm{il}}$ необходимо продемонстрировать алгоритмы для задач вывода в (2.7). Для вывода, согласованного с аннотацией $\max_{\mathbf{y} \in \zeta(\mathbf{z}_m)} F(\mathbf{x}_m, \mathbf{y}; \mathbf{w})$ мы используем $\alpha$-расширение только над метками из $\mathbf{z}_m$. Это может привести к несогласованной разметке — некоторые метки из $\mathbf{z}_m$ могут отсутствовать в $\mathbf{y}$. Предлагается использовать следующую эвристику для того, чтобы сделать найденную разметку удовлетворяющей ограничению. Для каждой метки $k$, такой что $k \in \mathbf{z}_m$ и $k \notin \mathbf{y}$, находится суперпиксель $i' = \arg\max_{i \in \mathcal{V}} F(\mathbf{x}_n, T(\mathbf{y}, i, k); \mathbf{w})$, где $T(\mathbf{y}, i, k)$ — разметка, отличающаяся от $\mathbf{y}$ только тем, что в $i$-й позиции находится $k$. В качестве новой разметки выбирается $T(\mathbf{y}, i', k)$. На практике применение этой эвристики не даёт значимого улучшения по сравнению с использованием несогласованных разметок.

Вывод, дополненный потерями, теперь не декомпозируется на унарные и парные потенциалы. Преобразуем функционал следующим образом:

$$\begin{aligned} &\max_{\bar{\mathbf{y}} \in \mathcal{Y}_m} \left(F(\mathbf{x}_m, \bar{\mathbf{y}}; \mathbf{w}) + \mathrm{K}_{\mathrm{il}}(\bar{\mathbf{y}}, \mathbf{z}_m)\right) = \\ &= \max_{\bar{\mathbf{y}} \in \mathcal{Y}_m} \left( F(\mathbf{x}_m, \bar{\mathbf{y}}; \mathbf{w}) + \sum_{k \notin \mathbf{z}} \sum_{i \in \mathcal{V}} c_i [\bar{y}_i = k] - \sum_{k \in \mathbf{z}} \frac{S}{|\mathbf{z}|} [\exists i : \bar{y}_i = k] \right) + \text{const}. \end{aligned} \tag{3.7}$$

Последняя максимизация соответствует выводу с штрафами за использование меток, для чего может использоваться эффективный алгоритм на основе $\alpha$-расширения [15].

### 3.2. Плотные рамки

Дальнейшие типы аннотации оперируют понятием *объектов* реального мира, таких как конкретный автомобиль или человек. На изображениях им соответствуют *образы объектов* — множества пикселей, получившихся проектированием этого объекта в пространство изображения. Не все категории в задаче семантической сегментации соответствуют объектам — такие категории как *трава*, *небо* являются фоновыми, поэтому для них не подходят соответствующие типы аннотации. С формальной точки зрения, будем считать образом объекта связную область пикселей изображения одной категории.

**Определение 9.** *Рамкой*, аннотирующей объект категории $k$, называется структура $z$, задающая прямоугольник на изображении, включающий в себя образ этого объекта. Для $z$ определены функции $\mathit{label}(z)$, а также $\mathit{left}(z), \mathit{right}(z), \mathit{top}(z), \mathit{bottom}(z)$, определяющие границы прямоугольника. Пусть $\mathbf{y}$ — разметка изображения, а $\mathcal{P}'_k$ — некоторое подмножество пикселей, получивших метку $k$: $\mathcal{P}'_k \subset \{\mathbf{p} \mid y_{\mathsf{S}(\mathbf{p})} = k\}$. Рамка $z$ описывает множество $\mathcal{P}'_k$, если $\mathcal{P}'_k \subset [\mathit{left}(z), \mathit{right}(z)] \times [\mathit{top}(z), \mathit{bottom}(z)]$, а также $\mathit{label}(z) = k$ (см. рис. 1в).

**Определение 10.** Пусть задано число $r \in [0, 0.5)$. Будем называть рамку $z$ *$r$-плотной* по отношению к множеству пикселей $\mathcal{P}'_k$, если выполняются следующие предположения о пересечении множеств:

$$\mathcal{P}'_k \cap \big( [\mathit{left}(z), \mathit{left}(z) + r(\mathit{right}(z) - \mathit{left}(z))] \times \times [\mathit{top}(z), \mathit{bottom}(z)] \big) \neq \varnothing, \tag{3.8}$$

$$\mathcal{P}'_k \cap \big( [\mathit{right}(z) - r(\mathit{right}(z) - \mathit{left}(z)), \mathit{right}(z)] \times \times [\mathit{top}(z), \mathit{bottom}(z)] \big) \neq \varnothing, \tag{3.9}$$

$$\mathcal{P}'_k \cap \big( [\mathit{left}(z), \mathit{right}(z)] \times \times [\mathit{top}(z), \mathit{top}(z) + r(\mathit{bottom}(z) - \mathit{top}(z))] \big) \neq \varnothing, \tag{3.10}$$

$$\mathcal{P}'_k \cap \big( [\mathit{left}(z), \mathit{right}(z)] \times \times [\mathit{bottom}(z) - r(\mathit{bottom}(z) - \mathit{top}(z)), \mathit{bottom}(z)] \big) \neq \varnothing. \tag{3.11}$$

Будем обозначать это отношение следующим образом: $z \sqsupseteq_r \mathcal{P}'_k$.

Согласно этому определению, расстояние от множества $\mathcal{P}'_k$ до каждой из сторон рамки не превосходит некоторого порога, зависящего от измерений рамки. Согласно исследованиям типичных анно-

таций, производимых пользователями, большинство рамок оказываются $r$-плотными с $r = 0.06$ [18]. Поэтому в дальнейшем под плотной рамкой мы будем понимать 0.06-плотную рамку.

**Определение 11.** *Аннотацией плотными рамками* категорий $\mathcal{K}' \subset \mathcal{K}$ на некотором изображении называют множество рамок, плотных по отношению к образам каждого из объектов категорий из $\mathcal{K}'$. Пусть $\mathbf{y}$ — разметка изображения, и для каждой категории $k \in \mathcal{K}'$ задано покрытие $\{\mathcal{P}_k^i\}_t$ множества пикселей, отнесённых к этой категории: $\bigcup_t \mathcal{P}_k^t = \{\mathbf{p} \mid y_{\mathsf{S}(\mathbf{p})} = k\}$, причём все $\mathcal{P}_k^t$ представляют собой связные множества. Тогда аннотация плотными рамками — это множество $\mathbf{z}^{\text{bb}} = \{z_k^t\}_{t,k}$, таких что $z_k^t \sqsupseteq_r \mathcal{P}_k^t,\ \ \mathit{label}(z_k^t) = k,\ \forall t, \forall k \in \mathcal{K}'$.

Заметим, что аннотация плотными рамками определяется по полной разметке неоднозначно из-за неединственности покрытия $\{\mathcal{P}_k^t\}_i$ и определения $r$-плотной рамки при $r > 0$.

Объекты на изображении удобно аннотировать плотными рамками. С другой стороны, сегменты фоновых категорий не соответствуют объектам, аморфны и часто их плотная рамка близка к границам изображения, поэтому рамки добавили бы мало информации к метке изображения. Далее в этом разделе рассматриваются аннотации, которые состоят одновременно из рамок и меток изображения. Например, для изображения могут быть заданы рамки для автомобилей и пешеходов, а также известно, что дополнительно присутствуют пиксели зданий, дороги, неба. Будем предполагать, что в рамках конкретного изображения категория может быть задана либо рамками, либо меткой изображения, хотя тип аннотаций для категории может меняться от изображения к изображению (см. в разделе 4.3 пример, демонстрирующий когда это может быть полезно).

**Определение 12 (*слабая функция потерь при наличии рамок*).** Пусть слабая аннотация изображения $\mathbf{z}$ задана парой $(\mathbf{z}^{\text{il}}, \mathbf{z}^{\text{bb}})$ метки изображения и множества рамочных аннотаций. Разобьём множество меток $\mathcal{K}$ на три подмножества в соответствии со слабой аннотацией $\mathbf{z}$: метки, которые определены рамками ($\mathbf{k}_b = \bigcup_{z \in \mathbf{z}^{\text{bb}}} \mathit{label}(z)$), метки, которые присутствуют в других местах ($\mathbf{k}_p = \mathbf{z}^{\text{il}}$) и метки, которые отсутствуют на изображении ($\mathbf{k}_a = \mathcal{K} \setminus (\mathbf{k}_b \cup \mathbf{k}_p)$). Множество суперпикселей $\mathcal{V}$ также разбивается: $\mathbf{v}_k = \left\{ i \in \mathcal{V} : \exists \mathbf{p} \in \bigcup_{z \in \mathbf{z}^{\text{bb}} : \mathit{label}(z) = k} \mathit{box}(z) : i = \mathsf{S}(\mathbf{p}) \right\}$ — объединение супер-

пикселей, находящихся хотя бы частично в рамках с меткой $k \in \mathbf{k}_b$, и $\mathbf{v}_0 = \mathcal{V} \setminus \bigcup_{k \in \mathbf{k}_b} \mathbf{v}_k$. Тогда объединённая слабая функция потерь выглядит так:

$$\mathrm{K}_{\text{il-bb}}(\mathbf{y}, \mathbf{z}) = \sum_{k \in \mathbf{k}_a} \sum_{i \in \mathcal{V}} c_i \left[ y_i = k \right] + \sum_{k \in \mathbf{k}_p} \sigma_k \prod_{i \in \mathcal{V}} \left[ y_i \neq k \right] +$$

$$\beta \sum_{z \in \mathbf{z}^{\text{bb}}} \left( \sum_{p=top(z)}^{bottom(z)} \nu_p^z \prod_{q=left(z)}^{right(z)} \left[ y_{\mathsf{S}(p,q)} \neq label(z) \right] + \sum_{q=left(z)}^{right(z)} \omega_q^z \prod_{p=top(z)}^{bottom(z)} \left[ y_{\mathsf{S}(p,q)} \neq label(z) \right] \right)$$

$$+ \sum_{k \in \mathbf{k}_b} \sum_{i \in \mathbf{v}_0} c_i \left[ y_i = k \right]. \quad (3.12)$$

Первые два слагаемых несут такой же смысл, как в (3.6). Третье слагаемое штрафует *пустые* строки и столбцы внутри рамок, то есть те, которые не содержат ни одного пикселя, выведенного как метка рамки (см. рис. 2). Последнее слагаемое штрафует метки рамок вне соответствующих рамок. Оценим параметры этой функции, предполагая, что половина каждой из рамок занята объектом соответствующей категории.

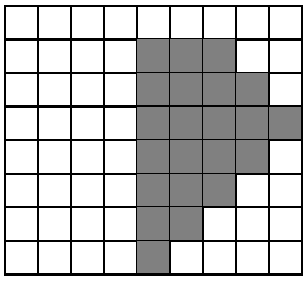

Рис. 2. Пример разметки внутри рамки. Клетки соответствуют пикселям. Серые клетки помечены меткой, равной метке рамки, белые — остальными метками. Разметка не является плотной, так как верхняя строка и четыре левых столбца — пустые.

**Теорема 3.** *Предположим, что в неизвестной разметке изображения каждый пиксель внутри рамки $z_i$ независимо принимает метку $label(z_i)$ с вероятностью* 0.5*, иначе принимает одну из меток в $\mathbf{k}_p$. Предположим снова, что количество пикселей для меток из $\mathbf{k}_p$*

*распределено мультиномиально с равномерными параметрами. Тогда, если рамки не пересекаются, при следующих параметрах оценка функции* $\mathrm{K}_{il\text{-}bb}$ *является несмещённой:* $\nu_p^z = (right(z) - left(z))/2$, $\omega_q^z = (bottom(z) - top(z))/2$, $\sigma_k = (S + \sum_{i \in \mathbf{v}_0} c_i)/2|\mathbf{z}^{il}|$, $\beta = 1$.

**Доказательство.** Пусть $S^{z_i}$ — количество пикселей внутри рамки $z_i$, принадлежащих категории $label(z_i)$. По предположению теоремы оно распределено по биномиальному закону: $S^{z_i} \sim \mathcal{B}(0.5, |box(z_i)|)$. Математическое ожидание этой величины равно $\hat{S}^{z_i} = |box(z_i)|/2$. Пусть $S^{\mathrm{il}}$ — число пикселей изображения, относящихся к категориям из $\mathbf{k}_p$. Зная $S^{z_i}$, можно оценить $S^{\mathrm{il}} = S - \sum_{z_i \in \mathbf{z}^{\mathrm{bb}}} S^{z_i}$. Рассуждая аналогично доказательству теоремы 2, получим оценку $\hat{\sigma}_k = \mathbb{E}S^{\mathrm{il}}/|\mathbf{z}^{\mathrm{il}}|$, которая позволяет несмещённо оценить $\mathrm{K}_{\mathrm{il\text{-}bb}}$. Поскольку $S^{\mathrm{il}}$ линейно зависит от $S^{z_i}$, можно заменить последнее на его оценку. Отсюда $\hat{\sigma}_k = (S - \sum_{z_i \in \mathbf{z}^{\mathrm{bb}}} |box(z_i)|/2)/|\mathbf{z}^{\mathrm{il}}| = (S - \sum_{k \in \mathbf{k}_b} \sum_{i \in \mathbf{v}_k} c_i/2)/|\mathbf{z}^{\mathrm{il}}| = (S + \sum_{i \in \mathbf{v}_0} c_i)/2|\mathbf{z}^{\mathrm{il}}|$.

Покажем несмещённость оценки, задаваемой третьим слагаемым на примере штрафа за пустые строки; для столбцов доказательство аналогично. Пусть $\hat{\nu}_p^z$ — математическое ожидание числа пикселей категории $label(z)$ в строке $p$. Согласно модели, $\hat{\nu}_p^z = (right(z) - left(z))/2$. Рассмотрим строки, в которых не найдено ни одного пикселя категории $label(z)$. Математическое ожидание ошибки на них равно $\hat{\nu}_p^z$. Строки, в которых выведен хотя бы один пиксель категории $label(z)$, не штрафуются. Таким образом, при $\nu_p^z = \hat{\nu}_p^z$, третье слагаемое даёт несмещённую оценку на число неправильно классифицированных пикселей категории $label(z)$ в пустых строках рамки $box(z)$.

Ещё более точную оценку можно получив, явно учтя в модели неравномерность распределения пикселей внутри рамки $z$, для которых метка равна $label(z)$. Коэффициенты $\nu_p^z$ и $\omega_q^z$ позволяют варьировать штраф за пустые строки и столбцы соответственно, в зависимости от их расположения в рамке. При достаточном количестве полностью размеченных изображений можно обучить специфичные для категорий профили $\nu^z$ и $\omega^z$.

В предыдущей подсекции мы показали, как обрабатывать первые два слагаемых в выводе, дополненном функцией потерь — первое декомпозируется на унарные потенциалы, а второе представляет собой штраф за наличие метки. Последнее слагаемое также декомпозиру-

ется на унарные потенциалы. Третье слагаемое — сумма потенциалов высокого порядка. Для каждой рамки $z$ каждая её строка и каждый столбец порождает потенциал над вершинами, соответствующими суперпикселям, которые пересекает эта строка/столбец. Мы также назначаем штраф за присутствие метки $label(z)$ на соответствующих вершинах, но не на всём графе, также модифицируя процедуру $\alpha$-расширения [15].

При выводе, согласованном с рамочной аннотацией, необходимо вывести разметку, в которой только суперпиксели внутри рамок могут получать метки соответствующих объектных категорий, причём, в соответствии с определением $r$-плотной рамки, сегменты объектов должны быть связными и примыкать к рамке плотно, с допуском не более $r$ от соответствующего измерения (напомним, что мы используем постоянное значение $r = 6\%$). Ограничение на метки вне рамок легко удовлетворяется при выводе: можно подавить нежелательные метки вне рамок, установив бесконечные унарные потенциалы.

Чтобы обеспечить плотность рамок, мы используем вариацию алгоритма *акцентирования* (*pinpointing*) [18], модифицированного для работы с многоклассовой сегментацией. Это эвристический алгоритм, гарантирующий, что разметка будет обеспечивать плотность рамок, однако не гарантируется оптимальность в классе таких разметок. Сначала вывод выполняется без ограничений на плотность. Затем, пока все ограничения не выполнены, одна из вершин меняет унарный потенциал, и выполняется шаг расширения. В нашей реализации выбирается вершина, соответствующую суперпикселю с наименьшим относительным потенциалом за $label(z)$ из тех, что ещё не получили эту метку. Этой вершине назначается бесконечный потенциал за метку $label(z)$, чтобы гарантировать, что метка вершины поменяется. Процедура конечна, если ни один суперпиксель не пересекает рамки разных меток, поскольку на каждой итерации хотя бы один суперпиксель внутри некоторой $box(z)$ меняет метку на $label(z)$.

Эксперименты показали, что при использовании такого типа аннотаций важна инициализация латентных переменных при обучении LV-SSVM. Наилучший результат имел место, когда изначально *все* суперпиксели внутри $box(z)$ получили метку $label(z)$.

Заметим, что Кумар и др. [7] использовали другой критерий для вывода, согласованного с аннотацией — они штрафуют пустые стро-

ки и столбцы внутри рамки (точная противоположность того, что наш алгоритм делает при выводе, дополненном рамочной функцией потерь). Эта эвристика не гарантирует плотность полученных сегментов внутри рамок.

### 3.3. Семена объектов

**Определение 13.** *Семенем*, аннотирующим объект категории $\dot{k}$, называется пара $z = (\mathbf{p}, \dot{k})$, задающая пиксель изображения, принадлежащий образу этого объекта. Пусть $\mathbf{y}$ — разметка изображения, а $\mathcal{P}'_k$ — некоторое подмножество пикселей, получивших метку $k$: $\mathcal{P}'_k \subset \{\mathbf{p} \mid y_{\mathsf{S}(\mathbf{p})} = k\}$. Семя $z = (\mathbf{p}, \dot{k})$ описывает множество $\mathcal{P}'_k$, если $\mathbf{p} \in \mathcal{P}'_k$, а также $\dot{k} = k$ (рис. 1г).

**Определение 14.** *Аннотацией семенами* категорий $\mathcal{K}' \subset \mathcal{K}$ на некотором изображении называют множество семян, принадлежащих образам каждого из объектов категорий из $\mathcal{K}'$. Пусть $\mathbf{y}$ — разметка изображения, и для каждой категории $k \in \mathcal{K}'$ задано покрытие $\{\mathcal{P}^i_k\}_i$ множества пикселей, отнесённых к этой категории: $\bigcup_i \mathcal{P}^i_k = \{\mathbf{p} \mid y_{v(\mathbf{p})} = k\}$, причём все $\mathcal{P}^i_k$ представляют собой связные множества. Тогда аннотация семенами — это множество $\mathbf{z}^{\text{os}} = \{z^i_k = (\mathbf{p}_{k,i}, k)\}_{i,k}$, таких что $\mathbf{p}_{k,i} \in \mathcal{P}^i_k,\ \forall i, \forall k \in \mathcal{K}'$.

Аннотация семенами определяется по полной разметке неоднозначно, однако предполагается, что семя находится в центре образа объекта. При выводе, согласованном с аннотацией, требуется, чтобы семена получили метку указанной категории. Ассоциативные парные потенциалы обычно распространяют эту метку на соседние суперпиксели.

**Определение 15 (*слабая функция потерь при наличии семян*).** Пусть слабая аннотация изображения $\mathbf{z}$ задана парой метки изображения и аннотации семенами $(\mathbf{z}^{\text{il}}, \mathbf{z}^{\text{os}})$. Определим объединённую слабую функцию потерь так:

$$\mathrm{K}_{\text{il-os}}(\mathbf{y}, \mathbf{z}) = \sum_{k \in \mathbf{k}_a} \sum_{i \in \mathcal{V}} c_i \left[y_i = k\right] + \sum_{k \in \mathbf{k}_p} \sigma_k \prod_{i \in \mathcal{V}} \left[y_i \neq k\right] + \\ + \beta \sum_{\substack{(\mathbf{p}', k') \\ \in \mathbf{z}^{\text{os}}}} \sum_{\mathbf{p} \in I} [y_{\mathsf{S}(\mathbf{p})} \neq k'] \exp\left(-\frac{\pi \|\mathbf{p} - \mathbf{p}'\|^2}{\tau_{k'}}\right). \quad (3.13)$$

Первые два слагаемых здесь несут тот же смысл, как в функции потерь для меток изображения. Третье слагаемое поощряет назначение метки семени в его окрестности. Покажем, как назначать параметры в этом случае.

**Теорема 4.** *Предположим, что в неизвестной разметке изображения число пикселей, отнесённых к меткам из* $\mathbf{z}^{il}$ *и* $\mathbf{z}^{os}$ *распределено мультиномиально с равномерными параметрами, и что для каждого семени* $z = (\mathbf{p}', k')$ *вероятность пикселя* $\mathbf{p}$ *принять метку* $k'$ *определяется гауссовым парзеновским окном:* $\exp\left(-\pi\|\mathbf{p} - \mathbf{p}'\|^2/\tau_{k'}\right)$. *Тогда при следующих параметрах оценка функции* $\mathrm{K}_{il\text{-}os}$ *является несмещённой:*

$$\begin{aligned} \tau_{k'} &= \frac{S}{(|\mathbf{z}^{il}| + \#Lab(\mathbf{z}^{os})) \cdot \#Obj(\mathbf{z}^{os}, k')}, \\ \sigma_k &= \frac{S}{|\mathbf{z}^{il}| + \#Lab(\mathbf{z}^{os})}, \quad \beta = 1, \end{aligned} \tag{3.14}$$

*если при этом семена находятся достаточно далеко друг от друга, а именно,* $\sum_{(\mathbf{p}',k')\in\mathbf{z}^{os}} \exp\left(-\pi\|\mathbf{p} - \mathbf{p}'\|^2/\tau_{k'}\right) \leqslant 1, \forall \mathbf{p} \in I$. *Здесь* $\#Lab(\mathbf{z}^{os})$ *— число различных меток в* $\mathbf{z}^{os}$, *а* $\#Obj(\mathbf{z}^{os}, k')$ *— число семян метки* $k'$ *в* $\mathbf{z}^{os}$.

**Доказательство.** Аналогично доказательству теоремы 2 можно получить оценку числа пикселей, отнесённых к каждой из категорий при мультиномиальном распределении: $\sigma_k = S/\left(|\mathbf{z}^{\mathrm{il}}| + \#Lab(\mathbf{z}^{\mathrm{os}})\right)$. Согласно условию теоремы, в окрестности семени $z = (\mathbf{p}', k')$ классификация пикселя $\mathbf{p}$ меткой, отличной от $k'$, влечёт в сильной функции потерь от неизвестной разметки слагаемое с математическим ожиданием $\exp\left(-\pi\|\mathbf{p} - \mathbf{p}'\|^2/\tau_{k'}\right)$. Исходя из линейности вхождения всех слагаемых, значение функции $\mathrm{K}_{\text{il-os}}$ равно математическому ожиданию функции потерь по неизвестным сильным разметкам. Остаётся определить масштаб парзеновского окна. При $\tau_{k'} = \sigma_{k'}/\#Obj(\mathbf{z}^{\mathrm{os}}, k')$, ожидаемое число меток в категориях из $\mathbf{z}^{\mathrm{os}}$ равно оценке на $\sigma_k$ (при условии достаточной удалённости семян):

$$\#Obj(\mathbf{z}^{\mathrm{os}}, k') \cdot \int_I \exp\left(-\frac{\pi\|\mathbf{p} - \mathbf{p}'\|^2}{\tau_{k'}}\right) d\mathbf{p} = \#Obj(\mathbf{z}^{\mathrm{os}}, k') \cdot \tau_{k'} = \sigma_{k'}. \tag{3.15}$$

Из равенства (3.15) получим искомую оценку $\tau_{k'}$.

Последний член функции потерь (3.13) декомпозируется на унарные потенциалы, так что вывод, дополненный функцией потерь, тривиален.

# 4. Эксперименты

## 4.1. Наборы данных, детали реализации, критерии качества

**Наборы данных.** Мы протестировали предложенный метод на двух наборах данных: MSRCv2[3] [19, 4] и SIFT-flow[4] [20, 21, 5]. Набор MSRC содержит 276 изображений в обучающей и 256 в тестовой выборке. Пиксели вручную отнесены каждый к одной из 23 категорий, хотя значительная их часть осталась неразмеченной. SIFT-flow содержит 2488 изображений в обучающей и 200 в тестовой выборке, они размечены с использованием 33 меток категорий.

**Структура модели и признаки.** Для набора MSRC суперпиксели получены с помощью авторской реализации детектора границ *gPb* [22]. Признаки унарных потенциалов следующие: гистограмма визуальных слов на основе дескриптора SIFT [23], построенная с помощью словаря из 512 слов, гистограмма цветов пикселей, построенная на словаре из 128 слов, гистограмма локаций на равномерной сетке $6 \times 6$. Объединённые векторы признаков нормализуются и отображаются в пространство более высокой размерности, где скалярное произведение приближает расстояние $\chi^2$ из оригинального пространства (размерность векторов признаков при этом утраивается) [24]. Признаки парных потенциалов состоят из 4 чисел: $\exp(-c_{ij}/10)$, $\exp(-c_{ij}/40)$, $\exp(-c_{ij}/100)$, 1. Здесь $c_{ij}$ — сила границы между суперпикселями, соответствующими вершинам $i$ и $j$, определённая детектором *gPb*.

Для набора SIFT-flow мы повторяем условия эксперимента Вежневца и др. [5]. Суперпиксели и признаки получены с помощью кода

[3] http://research.microsoft.com/en-us/projects/objectclassrecognition/
[4] http://people.csail.mit.edu/celiu/LabelTransfer/code.html

Тая и Лазебник [21]. Он использует графовую сегментацию Фельценсцвальба и Гуттенлохера [25] и затем вычисляет следующие признаки. Унарные потенциалы зависят от формы, положения, текстуры и пиксельной маски суперпикселей и их окрестностей: всего 3115 унарных признаков. Мы также преобразуем их, приближая ядро $\chi^2$, утраивая их размер [24]. Парные признаки вычисляются как расстояния над группами признаков суперпикселей ($\chi^2$-расстояния для гистограмм, евклидовы в противном случае), всего 26 парных признаков.

**Критерии качества.** Мы используем два объективных критерия качества сегментации, которые вычисляются по размеченной тестовой выборке: точность (*accuracy*) и средняя поклассовая полнота (*per-class recall*). Точность — это доля корректно распознанных пикселей тестовой выборки. Поклассовая полнота — это число корректно размеченных пикселей каждой категории, делённое на суммарную площадь категории в верной разметке, усреднённое по категориям. Следуя принятой практике [4, 26], мы исключили пиксели редких категорий (*лошадь* и *гора*) из подсчёта полноты для набора MSRC, однако учитываем метку *другое*, см. секцию 4.2. Аналогично мы не рассматриваем редкие категории (*корова*, *пустыня*, *луна*, *солнце*) при подсчёте полноты на наборе SIFT-flow.

### 4.2. Метки изображений

Мы автоматически получаем метки изображений из полной разметки, оставляя уникальные метки пикселей для каждого изображения. Изображение из набора MSRC обычно содержит один или несколько объектов конкретной целевой категории (например, *знак*, *корова*, *автомобиль*) на некотором фоне. Не любую фоновую категорию можно отнести к используемым 23 меткам, так что часть изображения может остаться неразмеченной. На практике некоторые изображения содержат только одну метку категории. В этом случае метка изображения однозначно определяет полную разметку. Чтобы избежать этого знания (нереалистичного при практическим использовании), мы моделируем дополнительную метку *другое*, к которой относятся все категории кроме обозначенных 23-х. Обычно разметки

имеют нечёткие границы, так что границы между сегментами различных меток также неразмечены (рис. 1б). Если мы будем относить их к категории *другое*, это может внести лишний шум в обучающую выборку. Поэтому необходимо использовать метку *другое* только для неразмеченных регионов, но не для границ. Мы используем следующий эвристический критерий для получения меток изображения: метка *другое* включается в список меток изображения тогда и только тогда, когда изображение содержит только одну метку или не менее 30% его пикселей неразмечены.

В нашей базовой постановке эксперимента имеется (возможно пустая) полностью размеченная часть обучающей выборки, при этом остальные изображения аннотированы метками изображений. Эти подмножества выбраны с помощью эвристического алгоритма так, чтобы пропорции меток в них отражали соответствующие пропорции во всей выборке.

Рис. 3а показывает точность и поклассовую полноту для сегментации тестовой выборки для различных размеров полностью размеченной части обучающей выборки, по сравнению с обучением на только сильно размеченной части выборки. В наиболее интересном случае, когда менее 20% обучающей выборки полностью размечены, слабо аннотированная подвыборка обеспечивает 10–15% увеличение и точности, и полноты. В случае полного отсутствия полных разметок, модель производит сегментацию с точностью 38% и полнотой 18%, что можно считать хорошим результатом для сегментации на 22 метки (полнота при случайной разметке составила бы 4.5%).

Когда в обучающей выборке одновременно присутствуют изображения с полной разметкой и со слабыми аннотациями, необходимо установить коэффициент $\alpha$ из (2.5). Рис. 3б показывает, что его оптимальное значение лежит ниже 1. Возможным объяснением этого факта является то, что слабо аннотированные изображения несут меньше информации, таким образом должны давать меньший вклад в целевую функцию. Для всех дальнейших экспериментов, где это применимо, мы используем $\alpha = 0.1$.

Поскольку наша реализация требовательна к ресурсам времени и памяти при обучении на наборе данных SIFT-flow (обучение длится до нескольких недель), мы не смогли провести настолько же подробный набор экспериментов. Вместо этого мы сравниваем обучение с

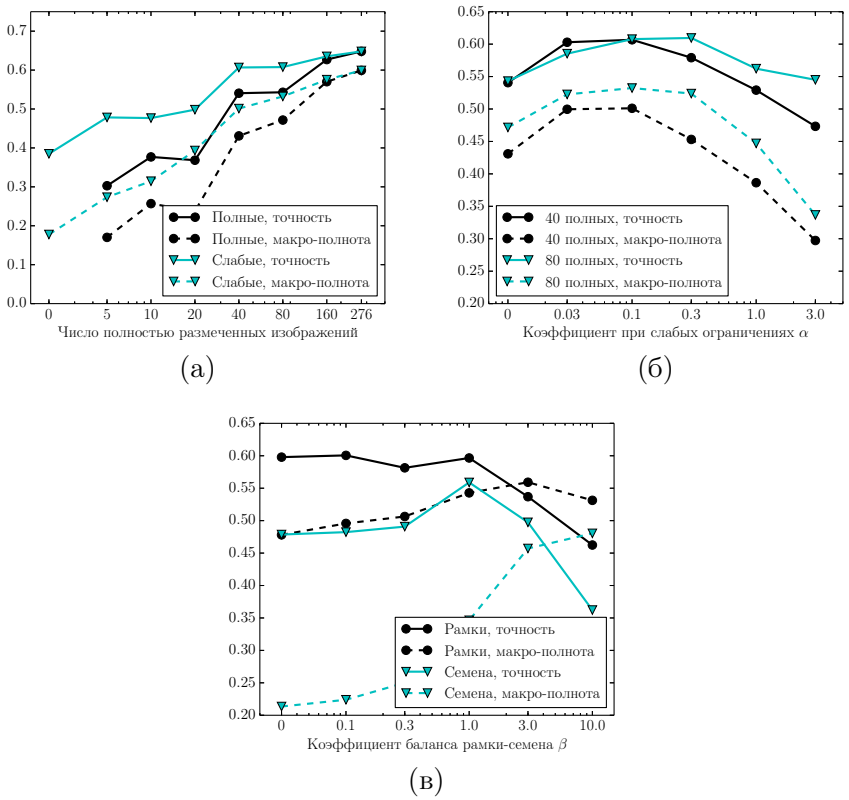


Рис. 3. Точность (сплошные линии) и поклассовая полнота (штриховые линии) при различных параметрах на наборе данных MSRC. (а) Изменение числа полностью размеченных изображений. Линии с круглыми маркерами показывают точность на тестовой выборке, если используются только полностью размеченные изображения, с треугольными — когда остальная часть обучающей выборки аннотирована метками изображений. (б) Изменение коэффициента слабой функции потерь $\alpha$. Линии с круглыми маркерами показывает точность сегментации, когда 40 изображений полностью размечены, с треугольными — когда 80 изображений; остальная часть обучающей выборки аннотирована метками изображений. (в) Изменение коэффициента функции потерь $\beta$ для плотных рамок (круглые маркеры) или семян объектов (треугольные маркеры). Все 276 изображений аннотированы метками изображений, а также все объекты аннотированы рамками или семенами, соответственно.

Таблица 1. Точность и средняя поклассовая полнота на наборе данных SIFT-flow. Первые две строки описывают обучение на подмножестве из 256 полностью размеченных изображений для моделей с парными потенциалами и без них, соответственно. Третья строка описывает обучение на наборе, где остальные 2232 изображения обучающей выборки аннотированы метками изображений. Последняя строка показывает результат обучения на полностью размеченной выборке из 2488 изображений.

| эксперимент | точн | полн |
|---|---|---|
| 256/256 полных, без парных связей (локальная) | 0.574 | 0.167 |
| 256/256 полных, иниц-я результатом локальной | 0.620 | 0.176 |
| 256/2488 полных, инициализация 256/256 | 0.674 | 0.208 |
| 2488/2488 полных | 0.696 | 0.246 |

полной разметкой со слабым обучением при фиксированной доле слабо аннотированных изображений, а именно при 256 полностью размеченных изображениях и 2232 — с метками изображений (Табл. 1). Эта слабообученная модель уступает обученной на полной разметке всего 2% по точности и 4% по полноте. Похожие результаты показала на этом наборе данных модель Вежневца и др. [5], которая также достигла полноты 21% при тех же признаках и суперпикселях. При этом их метод, хотя и добился того же результата, в отличие от нашего не используя даже 10% полностью размеченных изображений, является значительно более сложным: используется экстремально-рандомизированный хэширующий лес для нелинейного преобразования признаков, дополнительно обучаются априорные распределения «объектности» пикселей и категорий изображения, а также суперпиксели различных изображений соединяются в общую графическую вероятностную модель. Поскольку задача оптимизации, возникающая в SSVM с латентными переменными, невыпукла, алгоритм может остановиться в локальном минимуме или на плато целевой функции, так что хорошая инициализация желательна.

### 4.3. Добавление рамок и семян

Мы сгенерировали ещё два типа аннотаций для обучающих изображений набора MSRC. Как и в случае с метками изображений, мы генерируем аннотации по полной разметке. Плотные рамки и семе-

на объектов хорошо описывают объектные категории, но прибавляют мало информации для фоновых. Например, небо может занимать значительную часть изображения, так что его рамка не намного меньше всего изображения. Мы поделили список категорий на две части: фоновые, в том числе *трава*, *небо*, *гора*, *вода*, *дорога* и *другое*, и объектные, в которые вошли все остальные категории. Две категории, *здание* и *дерево*, проявляют двойственную природу — они могут отражать как основной объект на фотографии, так и задний фон (например, лес). Мы использовали следующую эвристику для каждого изображения: *здание* и *дерево* считаются фоном тогда и только тогда, когда помимо них на изображении есть другие объекты. Мы добавляем к меткам изображений обучающей выборки либо плотные рамки объектов, либо их семена. Для не-объектных категорий по-прежнему доступны только метки изображений. В качестве семян мы используем точки, наиболее удалённые от границ соответствующих объектов.

Таблица 2. Точность (первое число в каждой ячейке) и поклассовая полнота (второе число) на наборе MSRC, при обучении 1) только с полной разметкой, 2) если метки изображений (il) также доступны для оставшейся части выборки, 3) семена объектов (os) также доступны для оставшейся части выборки, 4) плотные рамки (bb) объектов доступны, 5) и семена, и плотные рамки доступны. Числа в последней колонке равны между собой, так как при полностью размеченной выборке слабая аннотация не добавляет информации.

| il | bb | os | 0/276 полных | 5/276 полных | 276 полных |
|---|---|---|---|---|---|
| − | − | − | n/a | 0.300/0.170 | 0.648/0.599 |
| + | − | − | 0.385/0.178 | 0.478/0.273 | 0.648/0.599 |
| + | − | + | 0.559/0.346 | 0.574/0.370 | 0.648/0.599 |
| + | + | − | 0.597/0.543 | 0.606/0.546 | 0.648/0.599 |
| + | + | + | 0.531/0.567 | 0.542/0.564 | 0.648/0.599 |

В таблице 2 собраны результаты эксперимента. Если полная разметка недоступна, и семена, и рамки значительно улучшают результат по сравнению с только метками изображений. Рамки особенно повышают поклассовую полноту — они помогают лучше обучать объектные категории, которые обычно занимают меньшую площадь, и соответственно дают низкий вклад в точность. В целом, обучение с

плотными рамками лишь на 5% уступает обучению с полной разметкой и по точности, и по полноте. Семена объектов дают меньший прирост качества, хотя их проще получать. Мы использовали значение $\beta = 1$ балансирующего коэффициента для типов слабой аннотации — вклад меток изображений и рамок (или семян) примерно одинаков (см. рис. 3в для подтверждения этой гипотезы).

## 5. Выводы

Представлен метод для обучения семантической сегментации изображений по различным типам аннотаций с помощью минимизации специализированных функций потерь для меток изображений, плотных рамок и семян объектов, в дополнение к полной разметке. Результаты показывают, что совместная аннотация, где фоновые категории заданы метками изображений, а объектные — плотными рамками, показывают лучшее качество сегментации тестовой выборки с учётом приложенных при аннотировании усилий.

## Список литературы

[1] Alvarez J. M., LeCun Y., Gevers T., Lopez A. M. Semantic Road Segmentation via Multi-scale Ensembles of Learned Features // ECCV. — 2012. — P. 586–595.

[2] Shotton J., Fitzgibbon A., Cook M., Sharp T., Finocchio M., Moore R., Kipman A., Blake A. Real-Time Human Pose Recognition in Parts from Single Depth Images // CVPR. — June 2011. — P. 1297–1304.

[3] Munoz D., Bagnell J. A., Vandapel N., Hebert M. Contextual classification with functional Max-Margin Markov Networks // CVPR. — June 2009. — P. 975–982.

[4] Vezhnevets A., Ferrari V., Buhmann J. M. Weakly Supervised Semantic Segmentation with a Multi-Image Model // ICCV. — 2011.

[5] Vezhnevets A., Ferrari V., Buhmann J. M. Weakly Supervised Structured Output Learning for Semantic Segmentation // CVPR. — 2012.

[6] Chang M.-W., Srikumar V., Goldwasser D., Roth D. Structured output learning with indirect supervision // ICML. — 2010.

[7] Kumar M. P., Turki H., Preston D., Koller D. Learning specific-class segmentation from diverse data // ICCV. — November 2011. — P. 1800–1807.

[8] Lou X., Hamprecht F. A. Structured Learning from Partial Annotations // ICML. — 2012.

[9] Pletscher P., Kohli P. Learning low-order models for enforcing high-order statistics // AISTATS. — 2012.

[10] Tarlow D., Zemel R. S. Structured Output Learning with High Order Loss Functions // AISTATS. — 2012.

[11] Taskar B., Chatalbashev V., Koller D. Learning associative Markov networks // ICML. — 2004. — P. 102–109.

[12] Tsochantaridis I., Joachims T., Hofmann T., Altun Y. Large margin methods for structured and interdependent output variables // JMLR. — 2006. 6. — P. 1453–1484.

[13] Joachims T., Finley T., Yu C. N. J. Cutting-plane training of structural SVMs // Machine Learning. — 2009. 77 (1). — P. 27–59.

[14] Boykov Yu., Veksler O., Zabih R. Fast approximate energy minimization via graph cuts // PAMI. — 2001. 23 (11). — P. 1222–1239.

[15] Delong A., Osokin A., Isack H. N., Boykov Yu. Fast Approximate Energy Minimization with Label Costs // IJCV. — July 2012. 96 (1). — P. 1–27.

[16] Yu C.-N. J., Joachims T. Learning structural SVMs with latent variables // ICML. — 2009.

[17] Yuille A. L., Rangarajan A. The concave-convex procedure (CCCP) // NIPS. — 2002.

[18] Lempitsky V., Kohli P., Rother C., Sharp T. Image segmentation with a bounding box prior // ICCV. — Sept. 2009. — P. 277–284.

[19] Shotton J., Winn J., Rother C., Criminisi A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation // ECCV. — 2006. — P. 1–14.

[20] Liu C., Yuen J., Torralba A. Nonparametric scene parsing: Label transfer via dense scene alignment // CVPR. — June 2009. — P. 1972–1979.

[21] Tighe J., Lazebnik S. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels // ECCV. — 2010.

[22] Arbeláez P., Maire M., Fowlkes C., Malik J. Contour detection and hierarchical image segmentation // PAMI. — May 2011. 33 (5). — P. 898–916.

[23] Lowe D. G. Distinctive Image Features from Scale-Invariant Keypoints // IJCV. — November 2004. 60 (2). — P. 91–110.

[24] Vedaldi A., Zisserman A. Efficient Additive Kernels via Explicit Feature Maps // CVPR. — July 2010.

[25] Felzenszwalb P. F., Huttenlocher D. P. Efficient Graph-Based Image Segmentation // IJCV. — September 2004. 59 (2). — P. 167–181.

[26] Shotton J., Johnson M., Cipolla R. Semantic texton forests for image categorization and segmentation // CVPR. — June 2008.