Сеть потока-движения и глубины Network для монокулярной стереосистемы и за ее пределами

Кайсюань Ван и Шаоцзе Шэнь

***Аннотация*—Мы предлагаем метод1, основанный на обучении , который решает монокулярную стереосистему и может быть расширен для объединения информации о глубине из нескольких целевых кадров. Учитывая два неограниченных изображения с a монокулярной камеры с известной внутренней калибровкой, наша сеть оценивает относительные положения камеры и карту глубины исходного изображения. The Основной вклад of the предлагаемого метода состоит в трехкратном увеличении. Во-первых, сеть предназначена для статических сцен, которые совместно оценивают оптический поток и движение камеры. При совместной оценке flow пространство поиска оптического потока space is постепенно сокращается , что приводит к an эффективной и точной оценке потока. Во-вторых, предлагается новый слой триангуляции proposed для кодирования the расчетного оптического потока и camera движения камеры , избегая при этом общих численных проблем, вызванных эпиполярностью. В-третьих, помимо оценки глубины с двумя видами, мы дополнительно расширяем вышеупомянутые сети, чтобы объединить информацию о глубине из нескольких целевых изображений и оценить карту глубины исходного изображения. Чтобы принести дальнейшую пользу исследовательскому сообществу, мы внедряем инструменты для создания реалистичных наборов данных "структура в движении", чтобы можно было хорошо обучить и оценить глубокие сети. Предлагаемый метод сравнивается с предыдущими методами и позволяет получить самые современные результаты за меньшее время. Изображения из реальных приложений и Google Планета Земля используются для демонстрации the generalization способности of the метода к обобщению.**

***Index Термины индекса*—SLAM, визуальное обучение, воздушные системы: восприятие и автономия.**

1. IИ НТПРОИЗВОДСТВО

**D**

Благодаря богатой информации в изображениях, структура из движения (SfM) имеет of жизненно важное значение в компьютерном зрении

и робототехника. Учитывая a набор of неограниченных изображений, SfM стремится to оценить карты глубины и относительные позы камеры. Традиционные системы, например, COLMAP [1], [2], сначала оценивают относительные положения камер, находя соответствия разреженных характерных точек , а затем используют the оцененное camera положение камеры для расчета плотных карт глубины. Извлеченные разреженные объекты игнорируют другую информацию на изображениях, такую как линии, и не способствуют последующей оценке глубины. Основные параметры сцены , такие как структуры и формы объектов, также трудно интегрировать в the конвейер of традиционных методов.

Чтобы лучше использовать image информацию об изображениях и использовать контекстные априоры, было предложено множество методов [3]–[5] для решения задач монокулярного стерео (УЛП с двумя видами) с использованием сверточных нейронных сетей (CNNs). DeMoN [3] является a новаторской работой , которая впервые

Рукопись получена September 10 сентября 2019 г.; принята February 2 февраля 2020 г. Дата of

публикация 21 февраля 2020 года; дата выхода текущей версии 9 марта 2020 года. Это письмо было рекомендовано для публикации заместителем редактора Л. Пауллом и редактором С. Бенке после оценки of the reviewers’ комментариев рецензентов. Эта работа была поддержана Программой стипендий PhD в Гонконге. *(Автор-корреспондент: Кайсюань Ван.)*

Кайсюань Ван и Шаоцзе Шэнь работают в ЕЭК, HKUST, Гонконг, Китай (электронная почта: [kwangap@ust.hk;](mailto:kwangap@ust.hk) [eeshaojie@ust.hk).](mailto:eeshaojie@ust.hk)

1https://github.com/HKUST-Aerial-Robotics/Flow-Motion-Depth. Цифровой Object идентификатор объекта 10.1109/LRA.2020.2975750

оценивает оптический поток, а затем разлагает его на карту глубины и позу камеры. Оптический поток, карты глубины и позы камеры are затем последовательно уточняются by a цепочкой of сетей кодировщика-декодера для обработки больших углов обзора. LS-Net [4] использует прогнозируемую карту глубины и позу камеры в качестве инициализации, чтобы итеративно minimize минимизировать photometric reprojection ошибку фотометрического through перепроектирования с помощью решателя на основе обучения. В отличие от LS-Net, где шаги обновления are вычисляются by a сетью, BA-Net [5] предлагает a bundle adjustment уровень настройки пакета для прогнозирования the коэффициента factor затухания алгоритма Левенберга - Марквардта [6] и вычисляет обновление. Для дальнейшего сокращения пространства оптимизации BA-Net также параметризует depth карту глубины как a линейную комбинацию of 128 single-view однопросмотровых прогнозируемых базовых карт. Использование информации вся изображения, тем выше методы создания надежной камерой, позы и ровную глубину картах. Хотя эти методы достижения впечатляющих результатов срав- нению с традиционными методами, их нужно несколько итераций (например, 15 повторений в ЛС-продажи и БА-продажи) , чтобы сходиться, и большинство методов (например, LS-продажи и Демон) рассчитать карту глубины, используя только один целевой кадр.

В этом письме мы повышаем эффективность и точность современного уровня за счет включения знаний о предметной области и дальнейшего расширения метода для объединения информации с различной глубиной. Первый вклад нашей работы заключается в совместной оценке оптического потока и позы камеры. Мы наблюдаем, что в монокулярных стереопроблемах оптический поток между многовидовыми изображениями вызван эго-движением движущейся камеры в статических сценах , так что оптический поток ограничен вдоль эпиполярных линий. Совместными усилиями, учитывая оптический поток и позы камеры, пространство поиска пикселей может постепенно сокращаться, повышая как эффективность , так и точность. Предложен новый triangulation слой триангуляции is proposed для кодирования расчетного оптического потока и движения камеры без численных проблем, вызванных неограниченными перемещениями камеры . The Закодированная информация из the triangulation слоя триангуляции is используется для оценки the depth карты глубины the исходного изображения. Во многих приложениях исходное изображение наблюдается несколькими целевыми изображениями. Помимо проблемы с двумя видами, мы дополнительно расширяем сети , чтобы объединить depth информацию о глубине из нескольких целевых кадров. Информация о глубине из разных пар изображений объединяется слоями среднего объединения и затем используется для прогнозирования карты глубины. На рис. 1 показан рабочий процесс of нашего метода оценки the оптического потока, позы камеры и карты глубины с учетом нескольких изображений. Используя многовидовые наблюдения, можно создавать надежные и точные карты глубины can be .

Для обучения и оценки методов УЛП на основе обучения требуется множество of изображений с ground truth позами камер наземной poses съемки и depth картами глубины. Существующие наборы данных, например SUN3D [7] и Scenes11 [3], содержат либо изображения низкого качества с камер RGB-D, либо нереалистичные синтетические изображения. Для обучения и оценки наших предлагаемых сетей мы разрабатываем инструменты, которые могут генерировать неограниченное количество высококачественных реалистичных изображений с ground truth картами глубины достоверности maps и

2377-3766 © 2020 IEEE. Личное использование is разрешено, но для переиздания/распространения требуется IEEE разрешение IEEE.

См . h[ttps://www.ieee.org/публикации/права/inde](http://www.ieee.org/publications/rights/index.html)x.html для получения дополнительной информации.

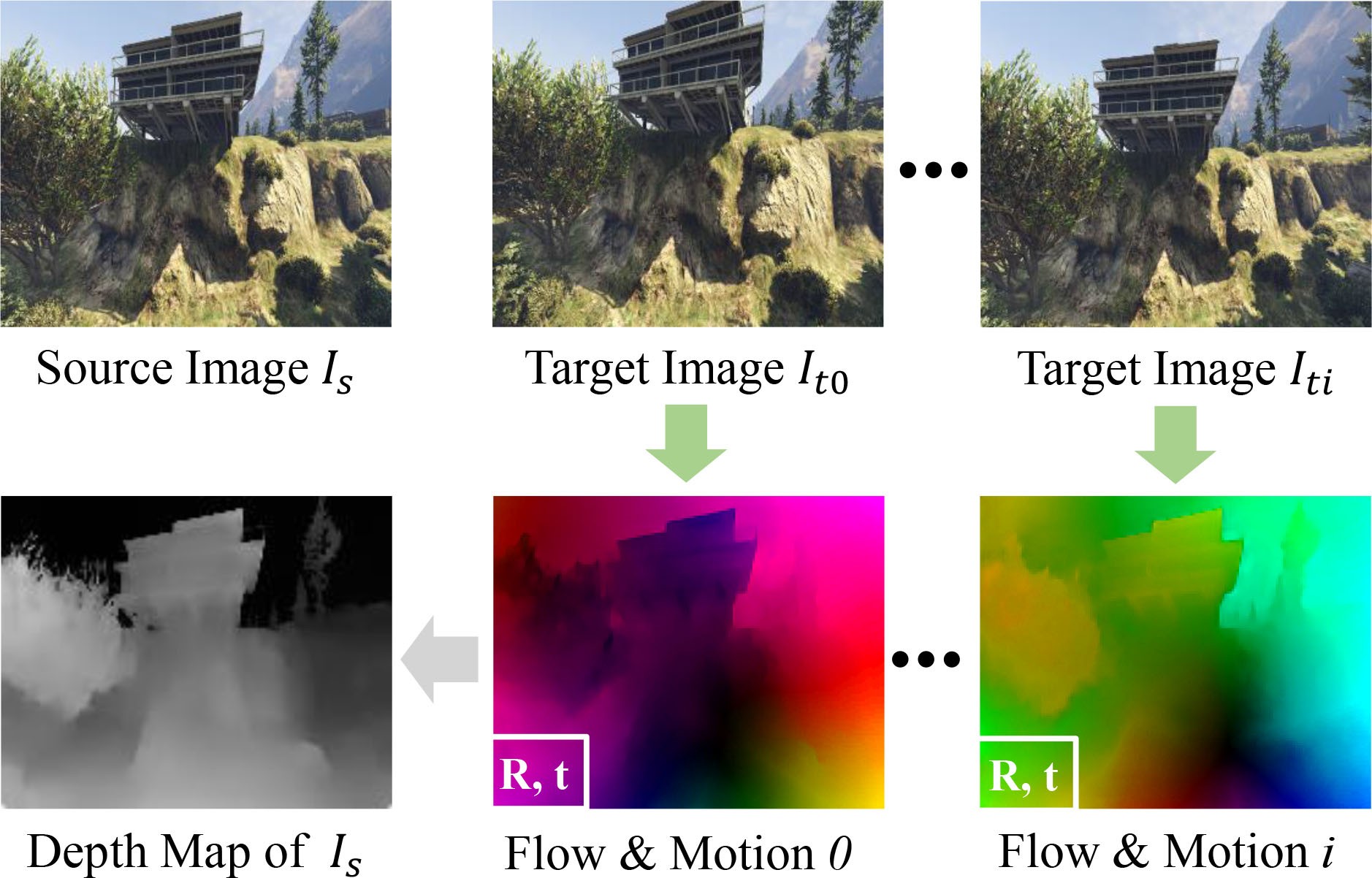


Рис. 1. Иллюстрация предлагаемого способа. Учитывая несколько изображений с движущейся монокулярной камеры, сеть движения потока (зеленая на рисунке) сначала оценивает the оптический поток и camera положение камеры между the исходным изображением и каждым целевым изображением. Оцененный поток и движение дополнительно объединяются сетью глубин (серый на the рисунке) для вычисления the depth карты глубины the исходного изображения.

камера позирует из the игры Grand Theft Auto V (GTA5). В интересах сообщества компьютерного зрения мы выпускаем инструменты и сгенерированные наборы данных с открытым исходным кодом.

To Подводя the итог of the , можно сказать, что в письме are the содержится следующее:

. A Сеть , которая совместно оценивает оптический поток и камеру

позы, заданные изображениями с двумя видами. При расчетных позах камеры оптический поток ограничен эпиполярными линиями таким образом, что поток можно регулировать, а пространство поиска сокращается.

. A Новый triangulation слой триангуляции , который кодирует the оценочную

оптический поток и положение камеры таковы, что сеть глубины может триангулировать глубину каждого пикселя без численных проблем.

. The Глубинная сеть is дополнительно расширяется , чтобы объединить depth информацию о глубине.

сопряжение (например, поток и движение) из нескольких пар изображений. Объединяя несколько наблюдений, image можно be estimated более точно и надежно оценить глубину исходного изображения .

. Open source Инструменты с открытым исходным кодом для настройки неограниченного реалистичного syn-

thetic полученные изображения с различными дневными, внутренними параметрами etc. The и т.д. Извлеченные изображения serve служат a дополнительным набором данных для обучения и оценки learning-based методов УЛП, основанных на обучении. methods.

1. РОКРЫЛЕННЫЙ ЖОРК

В этом разделе мы описываем связанную с этим работу с использованием нейронных сетей для оценки the camera поз камеры и depth карт глубины given для двух или более изображений.

ДеМоН [3] -это a новаторская работа , которая совместно оценивает depth карты глубины и camera позы камеры с учетом two-view изображений с двумя видами. Чтобы эффективно использовать наблюдения с двумя видами, ДеМоН адаптирует потоковые сети [8] , чтобы сначала оценить оптический поток между двумя изображениями, а затем разложить поток на позы камеры и карты глубины. Чтобы еще больше улучшить the качество, ДеМоН итеративно уточняет the оптический поток, позу камеры и карту глубины, используя две сети кодер - декодер , и, наконец, преобразует карту глубины в более высокое разрешение.

CodeSLAM [9] и BA-Net [5] параметризуют depth карты глубины как компактные представления, так что и движение камеры, и depth карта глубины могут быть solved явно решены с помощью классических optimization методов оптимизации . CodeSLAM использует an auto-encoder and автокодер и декодер для представления the depth карты глубины как функции function of the соответствующего изображения и

неизвестный код. Неизвестный код может быть решен совместно с позой камеры путем минимизации фотометрической и геометрической ошибок. Благодаря from the гибкости of the классической оптимизации CodeSLAM может одновременно оценивать несколько карт глубины и позы камеры. Чтобы сделать представление глубины подходящим для SfM задач SfM, BA-Net внедряет the bundle настройку пакета в качестве a отдельного слоя в сеть, и весь процесс можно обучать от начала до конца . В отличие от CodeSLAM и BA-Net, LS-Net [4] обучает CNN как решатель наименьших квадратов solver для обновления camera поз камеры и depth значений глубины. Начиная с инициализированных карт глубины и поз камеры, these для сходимости этих методов требуется несколько итераций to .

Для решения многовидовых задач с известными внешними и внутренними параметрами было предложено множество подходов, основанных на обучении . DeepMVS [10] генерирует объемы затрат с использованием изученных карт функций, а затем оценивает карту несоответствий путем объединения нескольких объемов затрат. MVDepthNet [11], DPSNet [12] и MVSNet [13], [14] решают одну и ту же reconstruction задачу реконструкции , но различаются в the расчете объемов затрат volumes и the struc- структуре сетей. Расчет объема затрат, который явно измеряет сходство между пикселями,создает высококачественные карты глубины. В предлагаемом методе мы применяем аналогичную стратегию , которая использует оценочные camera позы камеры для точного 1D pixel соответствия 1D пикселей.

В этом письме мы предлагаем a метод , который is отличается от всех упомянутых выше стереометрических методов монокуляра. Основное различие заключается в том, что наш метод не итеративно уточняет оценку, а скорее генерирует результаты, используя только один прямой проход в сети движения потока и сети глубины. Ключом к the повышению эффективности и качества является the совместная оценка of как оптического потока, так и движения камеры. Высококачественный оптический поток напрямую устанавливает точные плотные пиксельные соответствия между изображениями, обеспечивая точную триангуляцию глубины. Кроме того, the предлагаемый метод может быть расширен для оценки the depth карты глубины исходного изображения путем объединения информации из нескольких целевых изображений.

1. NРАБОТА НАД AАРХИТЕКТУРОЙ

Как показано на фиг. 2, предлагаемый способ состоит из двух сетей: одна потока-движения Сеть и одна глубина сети. Учитывая исходное изображение *ЯЗ* и изображение цели *ят* статической сцены, поток-движение сетевых оценки оптического потока между двумя изображениями и тем относительно камеры представляют в более грубой к тонкой манере. С камерой, позы оценка, пространство поиска оптического потока может быть постепенно снижена по epipolar линии. Кроме того, проблема диафрагмы в оптическом потоке может быть уменьшена ограничением эпиполярной линии. С помощью расчетного оптического потока и движения камеры можно непосредственно триангулировать значение глубины каждого пикселя. Однако шаг триангуляции не является численно стабильным вокруг эпиполярного [15]. Вместо этого мы предлагаем слой триангуляции для кодирования информации об оцененном оптическом потоке и позах камеры. Слой обрабатывается глубинной сетью для оценки карты глубины *Is*. Сеть глубины также может быть расширена для объединения информации из нескольких целевых изображений. Когда исходное изображение наблюдается несколькими целевыми изображениями, карта глубины исходного изображения может быть решена путем объединения информации из *всех* пар источник-цель .

В the следующих разделах мы сначала объясним the конструкцию of the

сеть потока-движения, глубинная сеть, которая обрабатывает проблемы УлП с двумя кадрами. В разделе III-C сеть глубины дополнительно расширена для объединения информации о нескольких глубинах и оценки the depth карты глубины the исходного изображения.

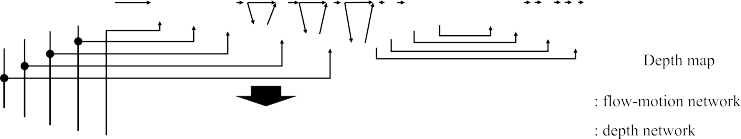
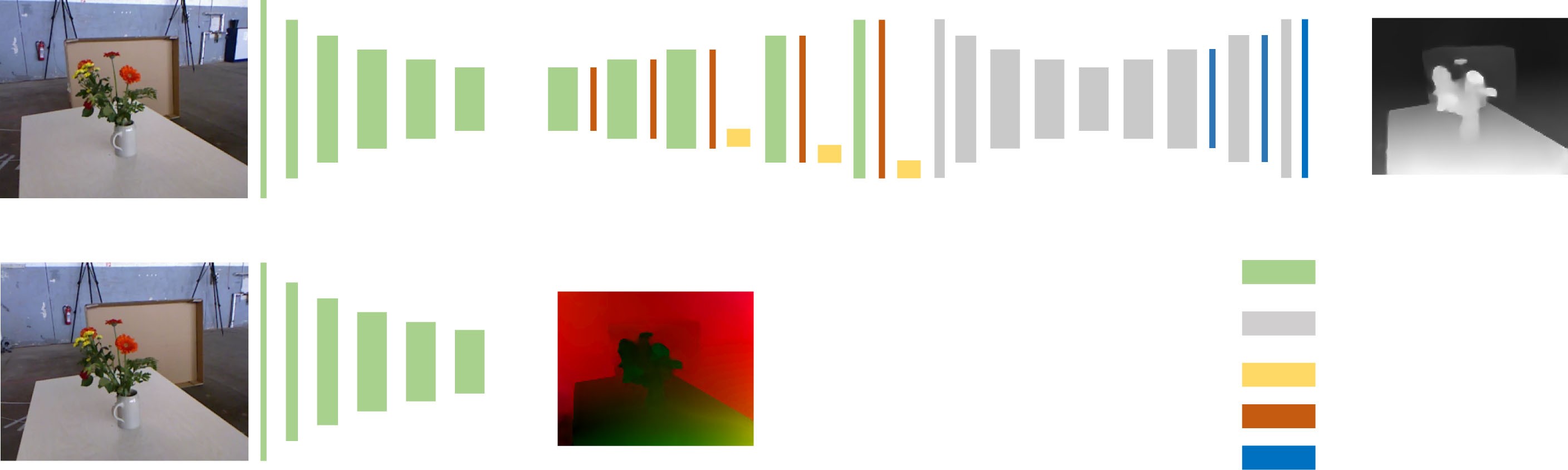


Рис. 2. The Архитектура of the предлагаемой flow-motion сети движения потока и depth сети глубины. Здесь only the two-view architecture для простоты показана только архитектура с двумя видами shown for . The Расширение для объединения multiple depth информации о нескольких глубинах is показано на рис. 5. The flow-motion Сеть движения потока совместно оценивает the оптический поток и camera позы камеры, а сеть глубины network триангулирует the глубину of каждого пикселя в the исходном изображении. Хотя обе сети are основаны на простых архитектурах кодер-декодер , the предлагаемая совместная оценка (Раздел III-A) и уровень триангуляции (Раздел III-B) обеспечивает высококачественную и эффективную оценку.

1. *Flow-Motion Сеть Движения Потока*

A Ряд of работ [8], [16]–[18] have показали the успех of использования CNNs для оценки плотного оптического потока между двумя изображениями. The Предлагаемая flow-motion сеть движения потока имеет similar структуру, аналогичную to the современной PWC-сети [18] , но is адаптирована для статических сцен

отношения,

**x** + **w**(**x**) *T*

1

*F* **x**

= 0*,* (2)

1

и совместно оценивает camera позы камеры.

Для обеспечения устойчивости к изменениям освещения и угла обзора входные ages are изображения преобразуются в L-level пирамиды объектов уровня L pyramids с помощью a простого CNN. The feature Карта объектов на the *i*-м уровне, **f** *i*, is processed обрабатывается тремя

*x*

*z*

где *F* = *K*−*T* **t** *RK*−1 - the фундаментальная матрица. С помощью

оценочная camera поза камеры, the upsampled увеличенный оптический поток of каждого пикселя

×

**w***i*+1(**x**) можно be регулировать , проецируя the соответствующую точку на the эпиполярную линию,

*вверх*

simplпростой е конвоlutionaл ь н о й слойс то порождаете тhe е нехт левел еeaturе *х*×*е*2 − *у*×*е е*

*y*

*x*

*y*

— *e e*

pyramid уровней пирамиды are , при **этом f** 0 является the исходным 3-канальным

отображение **f** *i*+1 с the size уменьшением размера на 2. В этой работе *L* = 6

*вверх,р*

*e*2 + *e*2

*y*×*e*2 − *x*×*e e*

— *e e*

изображение. **f** *i* и **f** *i* are используются для обозначения the *i*карт объектов i-го уровня feature maps

**w***i*+1 (**x**) = 1

— **x***,* (3)

*x y x*

*x y y z*

*с т*

из *Is* и *It*, соответственно.

The оптического потока **w** is оценивается от грубого до тонкого для обработки большого смещения пикселей. На *i*-м уровне оптический поток **w***i*+1 с the *i+*1-го уровня is сначала билинейно преобразуется в **w***i*+1 как

*вверх*

инициализация **w***i*. Объем затрат **c***i* строится с использованием **f** *i* и **f** *i*. Каждый элемент в объеме затрат представляет собой сходство характеристик между a пикселем **x***s* в **f** *i* и a пикселем **x***t* в **f** *i*,

*s*

*t*

*с т*

**c***i*(**x** *,* **x** ) = 1 *i T i*

гдеe [*ex, ey, ez*]*T* = *F* [**x***,* 1]*T* anи d [*x,*× *y*×]*T* = **x** + **w***i*+1(**x**).

, так the как соответствующие пиксели are ограничены эпиполярным epipolar

*вверх*

линии, it is не обязательно to совпадать с пикселями , расположенными далеко от the линий. Кроме того, проблема диафрагмы, при которой соответствие пикселей не может быть определено из-за to the неоднозначных совпадений, может быть уменьшена путем включения ограничения эпиполярной линии. Однако эпиполярные линии, которые определяются оценочными позами камеры, могут not be accurate быть недостаточно точными to , чтобы исключить все пиксели с the линий. Здесь мы постепенно уменьшаем the search пространство поиска с грубой пирамиды

*s t* (**f***s*(**x***s*))

*N*

*i*

**f***t* (**x***t*)*,* (1)

уровни до тонких уровней. На the *i*-м уровне the соответствующие пиксели of

и *Ni* -это размерность объекта **f** *i*. Из-за перехода от грубого к тонкому способу only a subset of pixels in **f** *i* is needed для расчета объема затрат требуется только подмножество пикселей в f i calculate the cost . Объем затрат volume **c** , upsampled увеличенный оптический поток **w***вверх* , и

*s*

*i*

*t*

*i*+1

*y*

*s*

пиксель **x***s* is параметризуется как

**x***t* ∈ **x***s* + **w***вверх,r* (**x***s*)+

*i*+1

*h*(*e ,* −*e*

)*T* + *v*(*e ,e* )*T*

*x*

*x*

*y*

.

*e*2 + *e*2 .

**f** *i* are используются для прогнозирования the оптического потока **w***i* с помощью the DenseNet [19]

структуры.

Приведенная выше cost volume конструкция объема затрат и оценка оптического потока потока повторяются are повторяются от грубого к тонкому до тех оптический пор, пока оптический поток

максимум

максимум

*h* ∈ [−*hi*

*, hi*

]*,v* ∈ [−*vi*

*x*

*, vi*

максимум

максимум

*y*

] *,* (4)

оценивается желаемое разрешение is . В этой работе мы адаптируемся и

где *hя*

обозначает диапазон поиска вдоль эпиполярных линий и

улучшите the вышеуказанные процессы , включив the статическую сцену *vi*

максимум

*i*

предварительная и совместная оценка the camera позы камеры.

На разных уровнях пирамиды несколько сверточных слоев и линейных слоев используются для прогнозирования положения исходного кадра with относительно to the целевого кадра. The Поза состоит из матрицы вращения matrix *R* и вектора перемещения **t**. Хотя простые CNN генерируют точные оценки позы (как показано в экспериментах), также могут быть использованы более продвинутые сети (например, DFE [20]). При расчетном движении камеры и откалиброванном внутреннем *K*, вектор потока каждого пикселя может быть упорядочен вдоль соответствующей эпиполярной линии, а пространство поиска пикселей в объеме затрат volume может быть narrowed сужено.

В статических средах пиксель **x** в the исходном изображении и его optical flow вектор оптического потока **w**(**x**) к the целевому изображению имеют the следующие значения

макс -это диапазон поиска range по вертикали to от строк. В общей сложности (2*ч*макс + 1)(2*вi* + 1) пиксели are сопоставляются для каждого пикселя на the *i*-м уровне.

Рис. 3 иллюстрирует the разницу между the суммой volume затрат в PWC-сети и предлагаемой сетью движения потока. С учетом предшествующей статической сцены и предполагаемого движения optical можно регулировать предполагаемый оптический поток be , а the размер объема the затрат volume is уменьшается, что приводит к эффективной оценке.

максимум

1. *Глубинная Сеть*

Учитывая расчетный оптический поток **x** и положение камеры *R*, **t**, глубину пикселя depth *d* можно be легко триангулировать , решив,

**w**(**x**)+ **x** = *λ*(*KRK*−1[**x***,* 1]*T* · *d* + *K***t**)*,* (5)

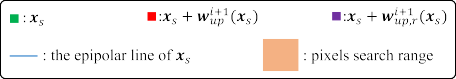
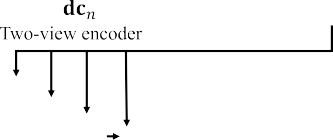
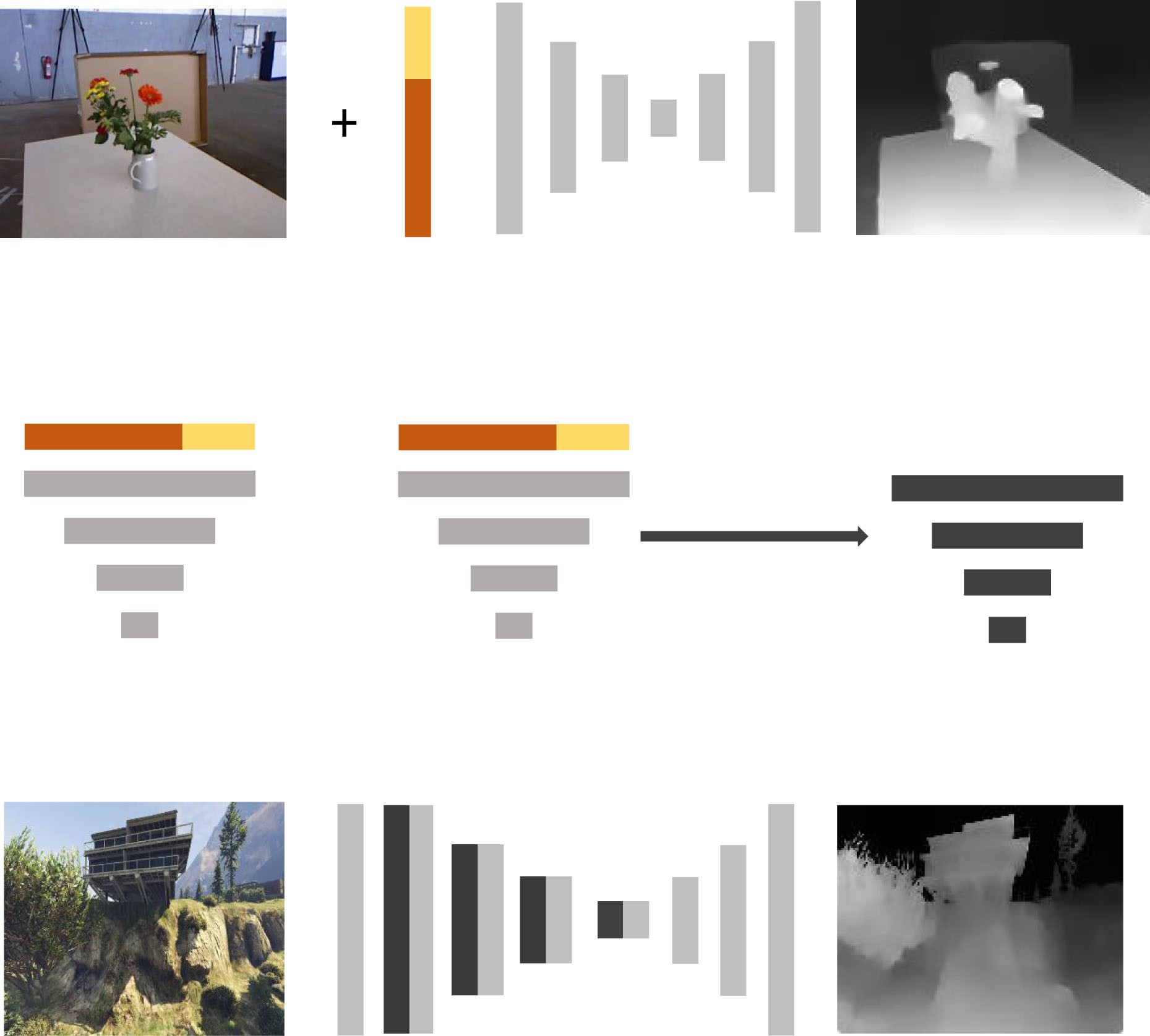




Рис. 3. Различия между расчетом объема затрат в PWC-Net (слева) и the предлагаемым flow-motion network (справа). Fили каждый pixel **x***s* in **f** *i*, PWC-Net соответствует a fixed set из pixels (окрашен в оранжевый цвет) aкруглый **x***s* + **w***i*+1(**x***s*) для создания объема затрат . С the другой стороны, the предлагаемая flow-motion сеть движения потока



*s*

*вверх*

во-первых,т рег ос теин и я л flо в**ш и***i*+1(**х***с*) into **w***i*в о в и+1 (**х***с*) anи д мащес пихэлс

*вверх*

вокруг the эпиполярной линии.

*вверх,р*



Рис. 5. Расширение сетки глубины для объединения информации о нескольких глубинах. (a) Сеть оценки глубины с двумя видами. (b) Расширение слияния на несколько глубин. Сеть *two- view кодировщиков с двумя видами network* кодирует the depth информацию о глубине of каждой image пары изображений в

глубина cоды **dc***i*. Mконечные коды are объединены в **dc**× и the *multiview fusion в многоуровневую сеть* takeс использованием tak e s **d c× t для**× to estimatоценки e depth карты.

Сеть глубины network is an -это сеть кодировщика-декодера network , которая использует триангуляционный слой **tri**, исходное изображение *Is*, оцененный оптический поток **w** и the последний слой of the flow-motion сети движения потока в качестве входных данных для оценки the depth карты глубины the исходного изображения.

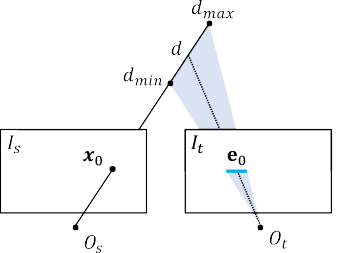
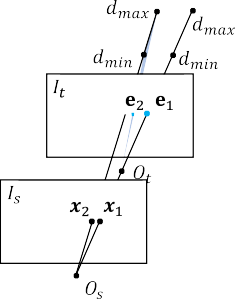
 

Рис. 4. Пример to , показывающий the численную стабильность при триангуляции. *Os* и *Ot* - the оптические центры of *Is* и *It*, соответственно. *d*max и *dmin* - the максимальная и минимальная глубина of the сцены. **е***я* находится в с, соответствующая epipolar линия из пЯхЭль **х***я*. (а) в стерео конфигурации, Тего глубина может быть надежно рассчитана на нахождение в соответствующие точки На **е**0. (б) в неограниченной монокуляр стерео проблемы, в epipolar строки **Е**1 О **х**1 (*в epipolar точкой*) вырождается в точку, так в глубину является ненаблюдаемым. Для пикселей вблизи the эпиполярной точки, таких как **x**2, the эпиполярная линия **e**2 is очень короткая, и в результате возникает помехоустойчивость.

где *λ*([*x, y, z*]*T* ) = [*x/z, y/z*]*T* - the dehomogenization функция дегомогенизации. Однако на этом этапе триангуляции существуют два недостатка step. : во-первых, the глубина is решается независимо для каждого пикселя, поэтому общая гладкость и параметры сцены игнорируются. Во - вторых, пиксели вокруг эпиполяра (проекция оптического центра целевого кадра на исходное изображение) не могут быть надежно триангулированы. На рис. 4 показаны the потенциальные численные проблемы при различных camera движениях камеры.

Для решения the вышеуказанных проблем ДеМоН использует сети для уточнения

1. *Слияние глубины многовидового просмотра Fusion*

В реальных приложениях (например, навигация роботов) глубина исходного изображения может быть определена с помощью нескольких целевых изображений. Здесь мы расширяем предлагаемую стереосеть с монокуляром с двумя видами works , чтобы объединить multiview информацию о нескольких видах. По сравнению с two-view парами изображений с двумя видами, многовидовые изображения дают больше информации о структуре окружающей среды, поэтому объединенные карты глубины могут быть более надежными и точными. Однако объединение информации о глубине из многовидовых изображений нетривиально из-за произвольного числа image пар изображений и различных depth масштабов глубины.

На рис. 5 показано, как расширяется сетка глубины с двумя видами. two-view Сетка глубины с двумя видами net , представленная в разделе III-B is , разделена на две части: кодер с двумя видами и слияние нескольких видов. Первая часть независимо кодирует the триангуляционный слой **tri** of каждой image пары изображений в коды глубины с несколькими разрешениями **dc**. Коды глубины из нескольких пар изображений {**dc**0, ..., **dc***N* 1} объединяются pooling слоями среднего объединения. The Объединенный код of каждого пикселя **dc** (**x**) is рассчитывается как,

— ×

*N* -11

Σ

карты триангулированной глубины maps (с заданной глубиной set to 0). Здесь вместо of уточнения the triangulated depth карт триангулированной глубины мы предлагаем an

**постоянный ток**×(**x**) = 1

*N*

**постоянный ток***i*(**x**)*.* (7)

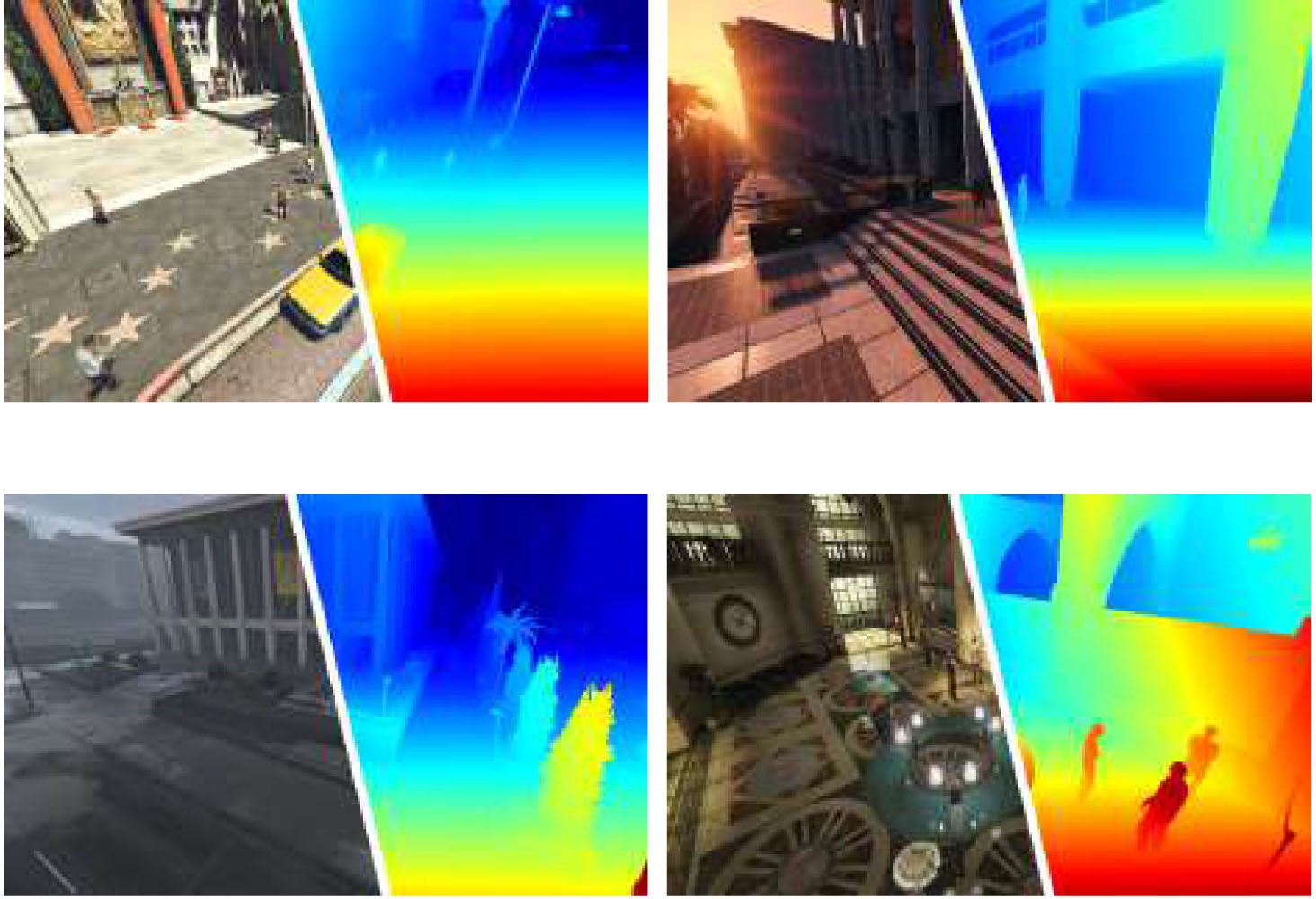
*i*=0

восьмиканальный уровень , который кодирует всю the информацию для triangu- триангуляции. Слой называется слоем триангуляции **tri**, и для каждого пикселя **x**,

**tri**(**x**) = [**w**(**x**)+ **x***,KRK*−1[**x***,* 1]*T ,K***t**]*T Т.* (6)

Использование слоев объединения для объединения информации использовалось во многих многовидовых стереофонических работах (например, DeepMVS [10]). В отличие от этих работ, мы используем несколько слоев объединения для объединения кодов глубины с разным разрешением, чтобы сохранить как глобальную информацию , так и мелкие детали are . Сеть слияния network

takes tта к е с т е ч е в ы е с л о в а н и е с л о в а, чтобы оценить соответствующую карту глубины.e fused depth code **dc**× and the source image *Is* to estimate the corresponding depth map.



1. NЭТЮДЫ DЭТЮДЫ

*A. Оптический поток и Camera движение камеры*

Пространство поиска для расчета объема затрат постепенно сокращается с грубого до мелкого. Сеть движения потока оценивает оптический поток от уровня 5 до уровня 1. С 5-го уровня на 1-й уровень шаги поиска *h*max и *v*max устанавливаются равными 4*,* 4*,* 4*,* 4*,* 3 и 4*,* 4*,* 4*,* 2*,* 1 , соответственно. На 1-м уровне сопоставляется только 21 пиксель (в PWC-Net используется 81 пиксель). The optical flow Потеря оптического потока is определяется как,

{ } { }

*Lflow* = Σ Σ ǁ**w***l*(**x**) − **w**ˆ *l*(**x**)ǁ2*,* (8) гдеe **w**ˆ *l* is the соответствуетg ground truth optical flow at tон

5

*l*=1

x

*l*-й уровень.

Вращение камеры rotation **r** is параметризуется как the трехмерный вектор вращения: **r** = *θ***v**, где *θ*-угол поворота, а **v**- ось вращения. Подобно демону, перевод камеры **t** is нормализуется как a единичный вектор из-за to the ненаблюдаемого масштаба. Поскольку оптический поток при грубых разрешениях не может обеспечить точное соответствие пикселей, движение камеры оценивается от level 3 до level 1. With the ground truth При использовании камеры с круглым mразрешением ˆ**r** anи d ˆ**t**, the motion потеря движения составляет,



Рис. 6. Образцы из the набора данных GTA-SfM dataset , включая различные погодные условия, время и сцены. Гибкость в изменении среды и настроек камеры повышает the удобство использования the набора данных в deep learning исследованиях с глубоким обучением.

1. DАТАСЕЦ
2. *DeMoN Набор данных демонов*

ДеМоН предлагает набор наборов данных для обучения и оценки глубоких сетей. Набор данных содержит изображения из нескольких источников, таких как камеры RGB-D [7], [24], результаты multiview SfM [1], [2], [25], [26], и синтетические изображения [3]. В общей сложности набор данных DeMoN содержит 57 тысяч пар изображений для обучения и 354 пары для тестирования.

Хотя набор данных DeMoN широко использовался в предыдущих работах [3] - [5], он содержит несколько ограничений. Во-первых, глубина

3

Σ

*Lдвижение* = ǁ**r***l* − ˆ**r**ǁ2

*l*=1

1. *Depth Оценка Глубины*

3

+ ǁ**t***l* − ˆ**t**ǁ2*.* (9)

Σ

*l*=1

карты с камер RGB-D cameras are не синхронизируются с the цветными изображениями и обеспечивают измерение глубины менее 10 метров . Во-вторых, большинство поз камеры на изображениях реального мира рассчитываются методами, основанными на оптимизации, на которые могут be повлиять by image шумы изображения или outlier особенности выбросов. Наконец, the

Несколько карт глубины оцениваются глубина сеть в разных разрешениях (с уровня 3 на уровень 1). Мы принимаем на глубину параметризация из собственных *и соавт.* [21], что выход сети является то войдите глубина: бревно(*д*) *р*. Из-за того, чтобы на шкале неоднозначности в УЛП проблемы, масштабно-инвариантного глубину ошибки для каждой точки **х** , это вычисляется по формуле,

∈

*дл* (**х**) = - ло -г(*гл*)(**х**)+ *αл* − лог(*д* ^ *л*)(**х**) (10) где *Д* € находится на первом правду глубина карте, и *αл* =

*e*

Σ log(*d* )(**x**) − log(*d* )(**x**) масштабирует по оценочную глубину

1 ˆ*l l N* x

maps. Помимо the l1 нормы l1 ошибки of the глубины error *Ld*, мы также следуем

Демон для наказания ошибок высокой частоты errors с использованием ошибки the градиента error *Lg*,

3

визуализированные синтетические изображения в the наборе are not данных нереалистичны. все эти аспекты ограничивают the производительность of the обученных сетей.

*B. Набор данных GTA-SfM Dataset*

Чтобы преодолеть ограничения в наборе данных демона, мы предлагаем набор данных GTA - SfM в качестве дополнения. Набор данных взят из GTA-V, игры с открытым миром с крупномасштабными city моделями городов. Благодаря to the активному сообществу мы разрабатываем инструменты для извлечения неограниченного количества реалистичных изображений с картами глубины и позами камеры. Извлеченные карты глубины обеспечивают измерение глубины для всех объектов на изображениях, включая тонкие структуры или отражающие поверхности. Мы извлекли 71 k пар of изображений для обучения и 5 k пар для тестирования. Наборы данных для обучения и тестирования не имеют общих сцен. Отличается от the DeMoN набора данных демона, один источник

*Ld* = Σ Σ ǁ*dl* (**x**)ǁ

*e*

*берХу*

*,* (11)

изображение может содержать несколько целевых изображений, таким образом the multiview depth fusion , можно протестировать многовизорное слияние глубин be .

*l*=1 х

*Lg* = Σ Σ .∇*xdl* (**x**). + .∇*ydl* (**x**). *,* (12)

*e*

*e*

3

*l*=1

x

A Аналогичный набор данных, MVS-SYNTH, is выпущен компанией Deep-

MVS [10] с использованием graphics debugging средств отладки графики. По сравнению с MVS-синтезатором, инструменты GTA-SfM tools могут свободно настраивать the camera FOV камеры,

погода и дневное время таковы, что разнообразие и удобство использования набора данных

где ǁ· ǁ*берХу* - the обратный Хубер [22], [23]:

are совершенствуются. Кроме того, траектория камеры trajectory помечается вручную annotated , чтобы камеры перемещались с большим перемещением и поворотом. Рис. 6

ǁ*x*ǁ*берХу*

= |*x*| если |*x*|≤ 1

*x*2 , если |*x*| *>>* 1*.*

(13)

показывает образцы из the предлагаемого набора данных.

Используя норму берХу, большие ошибки глубины наказываются нормой L2, а небольшие ошибки глубины также могут быть эффективно оптимизированы нормой the L1 .

1. E ЭКСПЕРИМЕНТЫXPERIMENTS

В этом разделе мы подробно оцениваем the производительность of the предлагаемой сети движения потока network и сети глубины . We Сначала мы

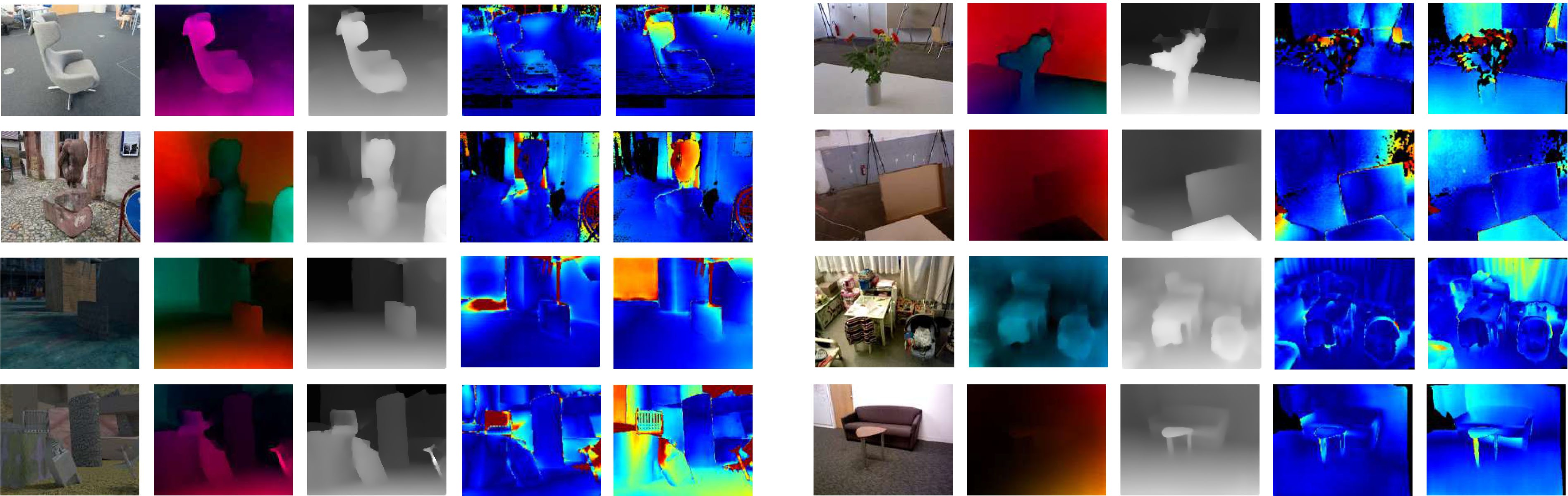


Рис. 7. Качественные результаты по базе данных демонов. Слева направо: исходное изображение, расчетный оптический поток, карта расчетной глубины, карта ошибок L1-rel и ошибка L1-rel карты расчетной глубины. Карта ошибок имеет цветовую маркировку JET. Как показано, наш метод генерирует высококачественные карты оптического потока и глубины. С помощью предлагаемого triangulation слоя триангуляции the depth карты глубины имеют меньше ошибок L1-rel .

сравните the предложенную сеть с the предыдущими работами [3]–[5] по парам изображений с двумя видами с использованием набора данных DeMoN. Затем производительность слияния глубин оценивается с использованием предлагаемого набора данных GTA - SfM. Эффективность предложенной совместной оценки движения потока и триангуляционного слоя **tri** также продемонстрирована в исследовании абляции. Далее мы демонстрируем способность of the метода к обобщению с использованием изображений реального мира images и изображений Google Earth .

1. *Evaluation Показатели Оценки*

Для оценки предполагаемого движения камеры и карт глубины используются различные показатели . мы следуем методу оценки method , используемому в DeMoN. Погрешность поворота определяется относительным углом между расчетным поворотом камеры и поворотом земной поверхности. Из-за неоднозначности масштаба в задачах УлП ошибка перевода определяется углом между нормализованными векторами перевода. Для оценки глубины *d* is сначала оптимально масштабируется расчетная глубина d scaled [3], затем вычисляются три показателя глубины are ,

L1-inv(*d, d*ˆ) = 1 Σ .1*/d*(**x**) − 1*/d*ˆ(**x**). *,* (14)

*N*

ТАБЛИЦА I

CСРАВНЕНИЕ ПО TWO-VIEW PЗАДАЧАМ

x



sc-inv(*d, d*ˆ) = 1 Σ *z*(**x**)2 − 1 Σ *z*(**x**) 2*,* (15)

*N*

*N* 2

x

x

L1-rel(*d, d*ˆ) = 1 Σ .*d*(**x**) − *d*ˆ(**x**). ,*d*ˆ(**x**)*,* (16)

*N*

x

где *z*(**x**) = log(*d*(**x**)) log(*d*ˆ(**x**)), а *N* - номер пикселя .

−

1. *Two-View Оценка с двух Точек Зрения*

Мы обучаем сеть движения потока и сеть глубины , используя *только* набор данных демона для справедливого сравнения. Обратите внимание, что демон обучается с большим набором данных, включая другие синтетические изображения. В экспериментах размер изображений изменяется до 320 256. Сеть движения потока network была обучена для 750 k шагов с the Adam помощью оптимизатора Adam [27]. С помощью обученной сети движения потока сеть глубины обучается для 260 к шагов. В соответствии с размером модели размер the мини-пакета size is устанавливается to равным 16 для the flow-motion сети движения потока и 24 для сети триангуляции . Оба learning-based метода основаны на обучении

×

(ДеМоН, LS-Сеть и BA-Сеть) и классический метод сравниваются в эксперименте. Классический метод предложен и оценен в демоне, который решает позы камеры с помощью нормализованного 8-точечного алгоритма [28] (с использованием функций SIFT) с последующей минимизацией ошибок перепроектирования. Карты глубины оцениваются с использованием plane sweep стереофонического и полуглобального сопоставления с разверткой плоскости [29].

Таблица I и рис. 7 показаны результаты сравнения глубины и движения. Благодаря совместной оценке потока и движения предлагаемый метод в большинстве случаев обеспечивает наилучшую оценку движения камеры most of the . The Предлагаемая глубинная сеть также обеспечивает стабильно лучшую производительность по сравнению с DeMoN. Com- По сравнению с BA - Сетью , которая итеративно уточняет the результаты (всего 95 мс ), наш метод генерирует стабильно лучшие позы камеры и конкурентоспособные depth карты глубины без каких-либо итераций (47 мс в

ТАБЛИЦА II

DEPTH MВсе, ЧТО Мне НУЖНО ЗНАТЬОERRORGTA- ССамого началаM D.

ТАБЛИЦА III

ЭФФЕКТИВНОСТЬFFECTIVENESS OF THE JМАЗИ FLOW-MДЛЯ СТИМУЛЯЦИИ НИЗКОЙ ЧУВСТВИТЕЛЬНОСТИSTIMATION



ТАБЛИЦА IV



EFFECTIVENESS OF THE TRIANGULATIONЭф ФЕКТИВНОСТЬ ТРИАНГУЛЯЦИИ Л АЙЕРAYER

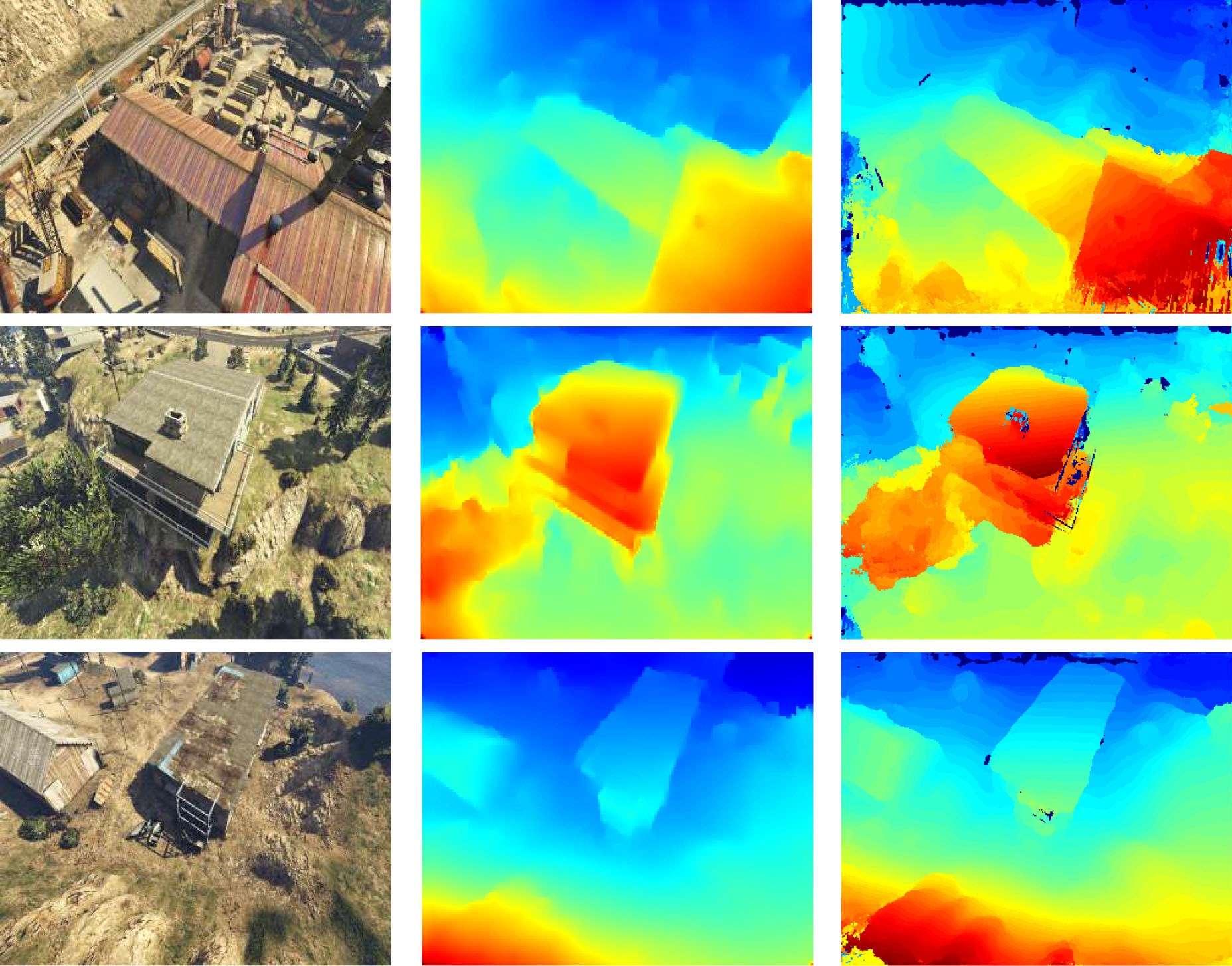


Рис. 8. Quality Сравнение качества of сгенерированных depth карт глубины по the предлагаемому методу и DEEPMVS. Каждое исходное изображение просматривается 6 целевыми изображениями, и Deep - MVS is provided снабжен ground truth позами камеры наземной съемки.

итого). Как показано на рис. 7, благодаря слою триангуляции, который кодирует геометрическую информацию, как близкие, так и удаленные объекты are реконструируются точно.

1. *Depth Fusion Оценка Глубины Слияния*

Поскольку набор данных DeMoN предоставляет только пары изображений с двумя видами , мы используем предлагаемый набор данных GTA-SfM для обучения и оценки производительности the слияния глубины fusion нескольких видов. We Сначала мы обучаем the flow-motion сеть движения потока с использованием two-view пар изображений с двумя видами pairs для 210 к шагов, а затем обучаем расширенную сеть слияния с несколькими видами для 130 к. steps. Размеры кода sizes для depth слияния по глубине is установлены на 128, 128, 128, 64 и 64 от грубого до тонкого соответственно.

Сначала мы оцениваем качество карт предполагаемой глубины, используя различное количество of целевых изображений. Мы также сравниваем сеть глубины с DeepMVS [10], которая также обучается с использованием изображений из GTA5. DeepMVS принимает изображения с ground truth позами камеры наземной правды poses в качестве входных данных, и наш метод использует только изображения. Для каждого количества целевых изображений мы *случайным* образом отбираем 300 пар и вычисляем среднюю depth ошибку глубины. В таблице II показано the depth качество глубины при различном количестве целевых изображений. Очевидно, что качество глубины улучшается при more наблюдении большего количества изображений , что свидетельствует об эффективности многовидового слияния и соответствует опыту классических методов УЛП. Мы также визуализируем оцененные карты глубины для качественного сравнения на рис. 8. Наш метод оценивает гладкие и подробные карты глубины, а DeepMVS оценивает дискретные карты глубины с выбросами.

1. *Ablation Исследование Абляции*

Здесь мы изучаем эффективность вклада: совместную оценку движения потока и слоя триангуляции с использованием одного и того же набора данных DeMoN.

Рис. 9. Сгенерируйте the предложенный метод по real-world изображениям реального мира.

**Совместная оценка:** Для оценки the важности of the epipolar ограничения эпиполярной линии и уменьшения пространства поиска мы удаляем оценку положения камеры на средних уровнях, и движение камеры оценивается с использованием окончательной оценки потока. Без ограничения эпиполярной линии на каждом уровне выполняется поиск 81 пикселя (такой же, как у PWC-Net) each . Как показано в таблице III, the совместная оценка улучшает *как* the оптический поток , так и оценку положения камеры pose .

**Triangulation Слой триангуляции:** The triangulation Слой is триангуляции предлагается для кодирования расчетного оптического потока и движения камеры без какой-либо числовой нестабильности. Чтобы продемонстрировать the эффективность, мы заменяем the triangulation слой триангуляции with a непосредственно триангулированной depth картой глубины. Подобно демону [3], значения NaN устанавливаются равными 0. Мы также проверяем the производительность оценки the глубины estimation с использованием оптического потока и позы камеры без совместной оптимизации. Сравнение показано в таблице IV. В the таблице мы также оцениваем the производительность прямой триангуляции глубины по оптическому потоку в первом ряду. С the помощью предложенного **tri**, depth сеть глубины может лучше использовать оцененный оптический поток и camera позы камеры.

1. *Generalization Способность к обобщению*

Чтобы проверить the generalization способность of the предлагаемого метода к обобщению, мы далее используем метод для оценки карт глубины изображений из разных источников. На рис. 9 показаны расчетные depth карты глубины изображений , полученных с помощью DJI Phantom 4 (наружная камера) или портативной камеры (внутренняя камера). На рис. 10 показаны карты предполагаемой глубины изображений с Google Earth. Карта глубины каждого исходного изображения состоит из 5 или 6 целевых изображений. Поскольку предлагаемый метод сначала создает высококачественные соответствия пикселей, а затем триангулирует глубину каждого пикселя, он может эффективно использовать многовидовые наблюдения и generalizes хорошо обобщается на другие изображения. Более подробная are информация содержится в the дополнительном материале.



Рис. 10. Сгенерируйте the предложенный метод для Google Earth изображений Google Earth.

1. CЗАКЛЮЧЕНИЕМ И FРАЗВИТИЕМ WORK

В этом письме мы предлагаем a flow-motion сеть движения потока и a depth сеть глубины, которая может оценить движение камеры и карту глубины с учетом нескольких стереоизображений движения. Обе сети тщательно разработаны, чтобы использовать многовидовые геометрические ограничения между оптическим потоком, движением камеры и картами глубины. Мы дополнительно расширяем the depth сеть глубины , чтобы объединить multiple depth информацию о нескольких глубинах в карту глубины. Для расширения доступных наборов данных предлагается инструмент с открытым исходным кодом для извлечения неограниченного количества реалистичных изображений с truth позами камеры наземной poses съемки и depth картами глубины.

RРЕЗУЛЬТАТЫ

* 1. Дж . Л. Шонбергер и Дж.-М. Фрам, “Пересмотр структуры из движения revisited,” ”, в

*Proc. IEEE Conf. Вычисление. Vision Распознавание образов Видения.*, 2016, с. 4104-4113.

* 1. J. Шонбергер, Э. Чжэн, М. Поллефейс и Дж.-М. Фрам, “Выбор вида selection по пикселям для неструктурированного многовидового стерео” , в *Proc. Eur. Conf. Comput. Vision*, 2016, стр. 501-518.
  2. Б. Умменхофер *et и др.*, “Демон: сеть глубины и движения network для обучения монокулярной стереосистеме", в *Proc. IEEE Conf. Вычислить. Распознавание Зрительного Паттерна.*, 2017, с. 5622-5631.
  3. R. Кларк, М. Блош, Дж. Чарновский, С. Лойтенеггер и А. Дж. Дэвисон, “Обучение to решению нелинейных наименьших квадратов для монокулярной стереосистемы” , в *Proc. Eur. Conf. Comput. Vision*, 2018, стр. 284-299.
  4. C. Тан и П. Тан, “BA-Сеть: Dense bundle сеть настройки плотных пучков network,” ", в *Proc. Int. Conf. Узнайте. Представления*, 2019 год.
  5. Дж. Ночедал и С. Дж.Райт, *Численная оптимизация*. Берлин, Германия: Спрингер, 2006.
  6. J. Сяо, А. Оуэнс и А. Торральба, “SUN3D: база данных больших пространств , реконструированных с использованием sfm и меток объектов”, в *Proc. IEEE Евро. Конф. Вычисл. Видение*, 2013, с. 1625-1632.
  7. A. Досовицкий, П. Фишер, Э. Илг, П. Хауссер, К. Хазирбас и В. Голков, “Потоковая сеть: изучение оптического потока с помощью сверточных сетей”, в *Proc. IEEE Int. Conf. Вычисл. Видение*, 2015, стр. 2758-2766.
  8. M. Блош, Дж.Чарновски, Р. Кларк, С. Лойтенеггер и А. Дж. Дэвисон, “CodeSLAM изучает компактное, оптимизируемое представление для плотного визуального шлема”, в *Proc. IEEE Conf. Вычислить. Распознавание Зрительного Паттерна.*, 2018, с. 2568.
  9. P. Хуан, К. Мацен, Дж. Копф, Н. Ахуджа и Дж. Хуан, “DeepMVS: изучение многовидового стереопсиса”, в *Proc. IEEE Conf. Вычислить. Картина Видения Распознается.*, 2018, с. 2821-2830.
  10. К. Ван и С. Шен, “MVDepthNet: Real-time multiview depth нейронная сеть для многовидовой оценки глубины в реальном времени”, в *Proc. IEEE Int. Конф. 3D Видение (3DV)*, 2018, с. 248-257.
  11. S. Im, Х. Чон, С. Лин и И. С. Квеон, “DPSNet: стереосистема сквозной deep plane развертки в глубокой плоскости stereo,” ”, в *Proc. Int. Conf. Узнайте. Представления*, 2019.
  12. Y. Яо, З. Ло, С. Ли, Т. Фан и Л. Куан, “MVSNet: Depth вывод о глубине для неструктурированного многовидового стерео”, в *Проц. Евро. Конф. Вычислить. Видение*, 2018, с. 767-783.
  13. Y. Яо, З. Ло, С. Ли, Т. Шен, Т. Фан и Л. Куан, “Рекуррентная MVSNet для вывода стереоизображения с несколькими видами высокого разрешения на глубину”, в *IEEE Conf. Вычисл. Vision Распознавание образов видения.*, 2019, с. 5520-5529.
  14. R. Хартли и А. Зиссерман, *Геометрия множественного представления в компьютерном зрении*. Кембридж, Массачусетс, США: Cambridge Univ. Press, 2004.
  15. E. Ilg, Н. Майер, Т. Сайкия, М. Кейпер, А. Досовицкий и Т. Брокс, “FlowNet 2.0: Эволюция оценки оптического потока с помощью глубоких сетей ”, в *Proc. IEEE Conf. Вычислить. Распознавание Зрительного Паттерна.*, 2017, с. 1647-1655.
  16. A. Ранджан и М. Дж.Блэк, “Оценка оптического потока с использованием пространственной пирамидной сети”, в *Proc. IEEE Conf. Вычислить. Распознавание Зрительного Паттерна.*, 2017, с. 2720-2729.
  17. D. Сан, Х. Ян, М. Лю и Дж. Каутц, “PWC-Net: CNNs для оптического потока с использованием пирамиды, деформации и cost объема затрат” , в *Proc. IEEE Conf. Вычисление Vision распознавания образов. Recognit.*, 2018, с. 8934-8943.
  18. G. Хуан, З. Лю, Л. ван дер Маатен и К. К. Вайнбергер, “Плотно связанные сверточные сети” , в *Proc. IEEE Conf. Вычисл. Vision Распознавание образов видения.*, 2017, с. 2261-2269.
  19. R. Ранфтл и В. Колтун, “Глубокая фундаментальная матричная оценка", в *Proc. Eur. Conf. Comput. Vision*, 2018, стр. 284-299.
  20. Д. Эйген, К. Пухрш и Р. Фергус, “ map Прогнозирование карты глубины по a одному изображению с использованием многомасштабной глубокой сети", в *Proc. Нейронный Инф. Процесс. Система.*, 2014, с. 2366-2374.
  21. Я. Лайна, С. Руппрехт, В. Белагианнис, Ф. Томбари и Н. Наваб, “ depth Прогнозирование более глубокой глубины с использованием полностью сверточных остаточных сетей” , в *Proc. Int. Conf. 3D Vision*, 2016, стр. 239-248.
  22. Л. Цвальд и С. Ламберт-Лакруа, “Штраф БерХу и сгруппированный эффект" , 2012, *arXiv:1207.6868*.
  23. J. Штурм, Н.Энгельхард, Ф. Эндрес, У. Бургард и Д. Кремерс, “ Эталон для оценки систем RGB - D SLAM”, в *Proc. IEEE/RSJ Int. Conf. Conf. Система интеллектуальных роботов Syst.*, 2012, стр. 573-580. 573–580.
  24. S. Фурманн, Ф. Ланггут и М. Гезеле, “MVE многовидовая struction среда для реконструкции" , в *Proc. Eurographics Workshop График семинара по еврографии. Культурное наследие*, 2014, с. 11-18.
  25. B. Умменхофер и Т. Брокс, “Глобальная, плотная многомасштабная реконструкция для a миллиарда точек” , в *Proc. IEEE Int. Conf. Comput. Vision*, 2015, стр. 1341-1349.
  26. Д. П. Кингма и Дж . Л. Ба, “АДАМ: A метод for стохастической оптимизации" , в *Proc. Int. Conf. Learn. Представления*, 2015.
  27. Р. И. Хартли, “В защиту алгоритма из the восьми точек algorithm,” *”, IEEE Trans. Образец Анал. Mach. Маш.*интеллекта, том 19, № 6, стр. 580-593, июнь 1997 года.
  28. H. Хиршмюллер, “Стереообработка с помощью полуглобального сопоставления и взаимной информации” *, IEEE Trans. Образец Анал. Mach. Маш.*, vol. интеллекта, том 30, вып. 2, стр. 328-341, февраль 2008 г.