PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

DENSE SEMANTIC LABELING OF SUB-DECIMETER RESOLUTION IMAGES WITH CONVOLUTIONAL NEURAL NETWORKS (2016)

Michele Volpi
Devis Tuia - IEEE

CT3115 - Sistemas Urbanos Inteligentes

Pontificia Universidad Católica de Chile, Escuela de Ingeniería
Departamento de Ciencia de la Computación

CAMILA F. VERA VILLA

mayo 6, 2021

Estructura del paper

I. INTRODUCTION
II. CNN
III. CNN ARCHITECTURES CONSIDERED IN THE PAPER
IV. DATA AND EXPERIMENTAL SETUP
V. RESULTS
VI. DISCUSSION AND CONCLUSION

Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks

Michele Volpi, *Member, IEEE*, Devis Tuia, *Senior Member, IEEE*

***Abstract*—Semantic labeling (or pixel-level land-cover classification) in ultra high resolution imagery (< 10cm) requires statistical models able to learn high level concepts from spatial data, with large appearance variations. Convolutional Neural Networks (CNNs) achieve this goal by learning discriminatively a hierarchy of representations of increasing abstraction.**

In this paper we present a CNN-based system relying on an downsample-then-upsample architecture. Specifically, it first learns a rough spatial map of high-level representations by means of convolutions and then learns to upsample them back to the original resolution by deconvolutions. By doing so, the CNN learns to densely label every pixel at the original resolution of the image. This results in many advantages, including i) state-of-the-art numerical accuracy, ii) improved geometric accuracy of predictions and iii) high efficiency at inference time.

We test the proposed system on the Vaihingen and Potsdam sub-decimeter resolution datasets, involving semantic labeling of aerial images of 9cm and 5cm resolution, respectively. These datasets are composed by many large and fully annotated tiles allowing an unbiased evaluation of models making use of spatial information. We do so by comparing two standard CNN architectures to the proposed one: standard patch classification, prediction of local label patches by employing only convolutions and full patch labeling by employing deconvolutions. All the systems compare favorably or outperform a state-of-the-art baseline relying on superpixels and powerful appearance descriptors. The proposed full patch labeling CNN outperforms these models by a large margin, also showing a very appealing inference time.

***Index Terms*—Semantic labeling, Classification, Convolutional neural networks, Deconvolution networks, Deep learning, Sub-decimeter resolution, Aerial images.**

I. Introduction

SEMANTIC labeling is the task of assigning a semantic label (land-cover or land-use class) to every pixel of an image. When processing ultra-high resolution data, most of state-of-the-art methods rely on supervised classifiers trained on specifically hand-crafted feature sets (appearance descriptors), describing locally the image content. The extracted high-dimensional representation is assumed to contain enough information to cope with the ambiguities caused by the limited spectral information of the ultra-high resolution sensors.

In the pipeline described above, input images undergo a spatial feature extraction step implemented by specific operators on local portions of the image (patches, superpixels or regions, objects, etc.), so that particular *spatial* arrangements of colors are encoded into a high-dimensional representation. A supervised classifier is usually employed to learn a mapping from the appearance descriptors to the semantic label space, which in turn allows to assign a label to every region of a previously unseen test image. Common examples of spatial features are texture statistics, mathematical morphology and oriented gradients [1]. Other common approaches strategies rely on bag-of-visual-words [2]. This mid-level representation is based on a quantization of appearance descriptors such as gradients, orientations, texture (usually obtained with a clustering algorithm). This quantization is then pooled spatially into histograms of spatial occurrences of cluster labels, or bag-of-visual-words. For instance, in [3] bag-of-visual-words are used to classify image tiles and detect compound structures.

The drawback of these approaches is that the features depend on a specific (set of) feature extraction method, whose performance on the specific data is a-priori unknown. Moreover, most appearance descriptors depend on a set of free parameters, which are commonly set by user experience via experimental trial-and-error or cross-validation [1], [4]. Exhaustive and global optimization of such values is unfeasible in reasonable time, but the selection from random feature ensembles has shown to be an effective proxy [5]–[7]. In these cases, the filter families from which to chose from are predefined and the parameters of the system are selected heuristically by random search to minimize the error over the semantic labeling task. Although the selection of features is data-driven, the filters themselves are still not learned end-to-end from the data, thus potentially sub-optimal.

Deep learning deals with the development of systems trainable in an end-to-end fashion. End-to-end usually means learning jointly a series of feature extraction from raw input data to a final, task-specific, output. All deep learning systems implicitly learn representations optimizing the loss on top of the network, driving the training of the model's weights. They usually minimize a task-specific differentiable loss function for classification, regression, semantic labeling, super-resolution or depth estimation, and the network learns representations which minimize such loss. Most common deep learning algorithms are (stacked) autoencoders [8]–[10], restricted Boltzmann machines [11], [12] and deep belief networks [13], [14]. For a review of main approaches we refer to [14], [15]. In this paper we focus on Convolutional Neural Networks (CNNs) [16]. Differently from other approaches, CNNs were specifically designed for image classification tasks, i.e. assigning a single class label to an entire image / scene. Representations are obtained by learning a hierarchy of convolution filters from the raw image. All the weights of the convolutions are learned end-to-end to minimize the classification error of the model.

CNNs have become extremely successful in many modern, high-level, computer vision tasks, ranging from image classification to object detection, depth estimation and semantic

Manuscript received YYXX, 2016;
MV and DT are with the MultiModal Remote Sensing group, University of Zurich, Switzerland. Email: {michele.volpi,devis.tuia}@geo.uzh.ch, web: http://geo.uzh.ch/ , Phone: +4144 635 51 11, Fax: +4144 635 68 48.
Digital Object Identifier xxxx

Conceptos relevantes y contexto

Al procesar datos de ultra alta resolución

Métodos tradicionales

se basan en clasificadores supervisados entrenados en conjuntos hand-crafted, como descriptores de apariencia, que describen localmente el contenido de la imagen.

↓

El inconveniente: las características dependen de un método de extracción que comúnmente es según la experiencia del usuario a través de prueba y error experimental.

V/S

Deep learning - CNN

desarrollo end-to-end: aprende en conjunto de manera empírica una serie de extracciones de características desde los datos de input sin procesar hasta un output.

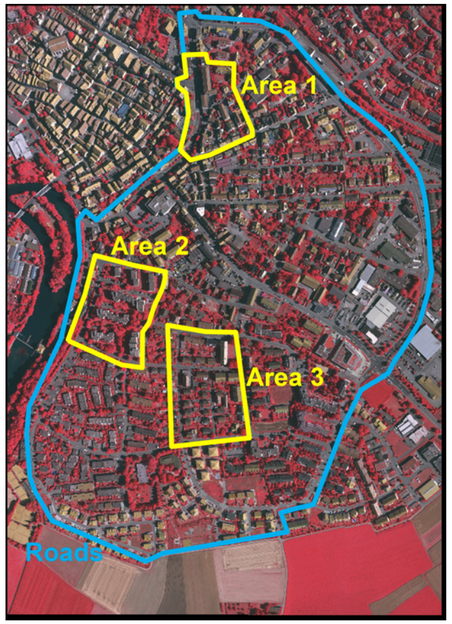
↓

Las CNN son muy usadas para tareas de visión computacional y también se han adaptado para problemas de etiquetado semántico.

Conceptos relevantes y contexto

DEEP LEARNING IN REMOTE SENSING

En sus inicios se centraban en lograr asignar una etiquetas semántica a parches grandes como "urbano", "urbano disperso" o "bosque".



¿Qué buscaban los autores?

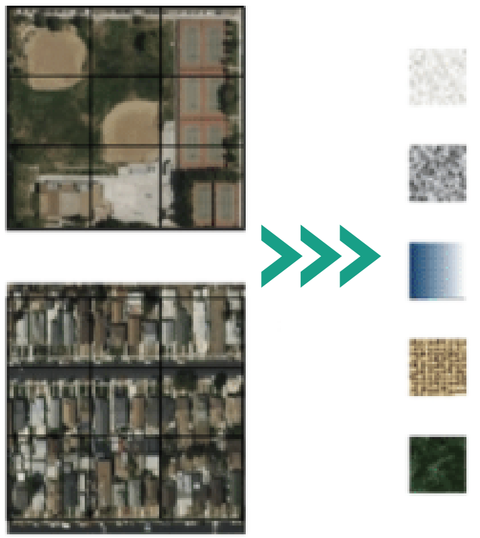
Entrenar un modelo que fuese capaz de etiquetar cada píxel presente en la imagen. Esto debe lograrse no solo aprendiendo la relación entre colores y etiquetas, sino principalmente aprendiendo y teniendo en cuenta las relaciones espaciales a diferentes escalas.

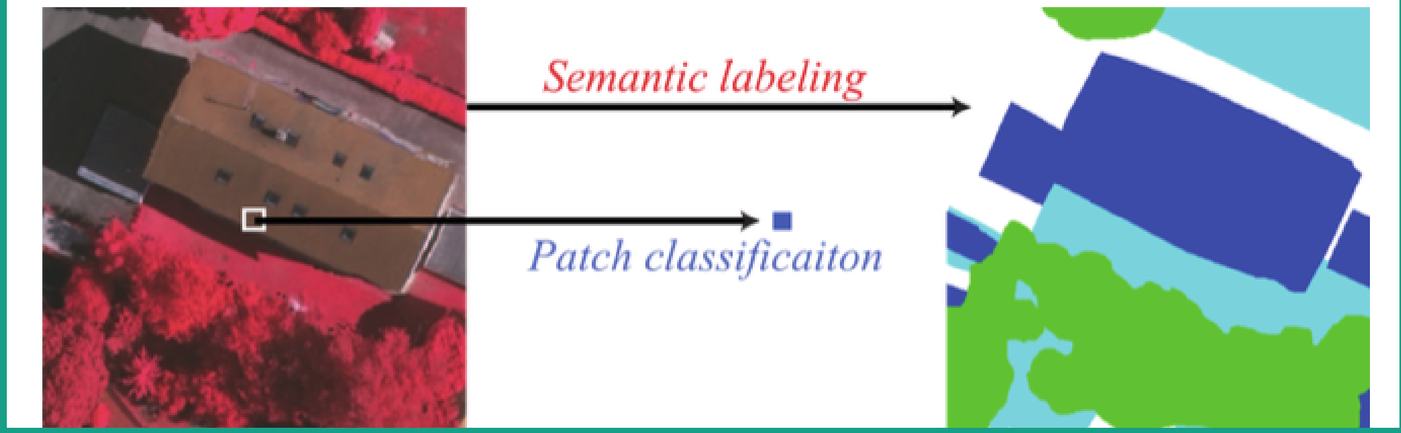
Semantic labeling

Conceptos relevantes y contexto

Etiquetado semántico (semantic labeling)

Es la tarea de asignar una etiqueta semántica (cobertura del suelo o clase de uso del suelo) a cada píxel de una imagen.



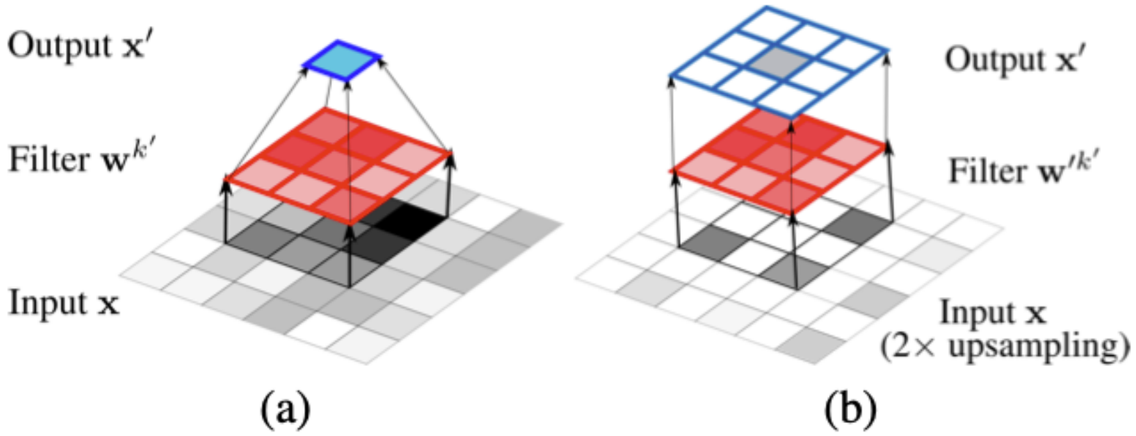


Pach classification: aprende una sola etiqueta por parche (la del píxel central)

Semantic labeling: aprende a predecir densamente las etiquetas semánticas para cada ubicación.

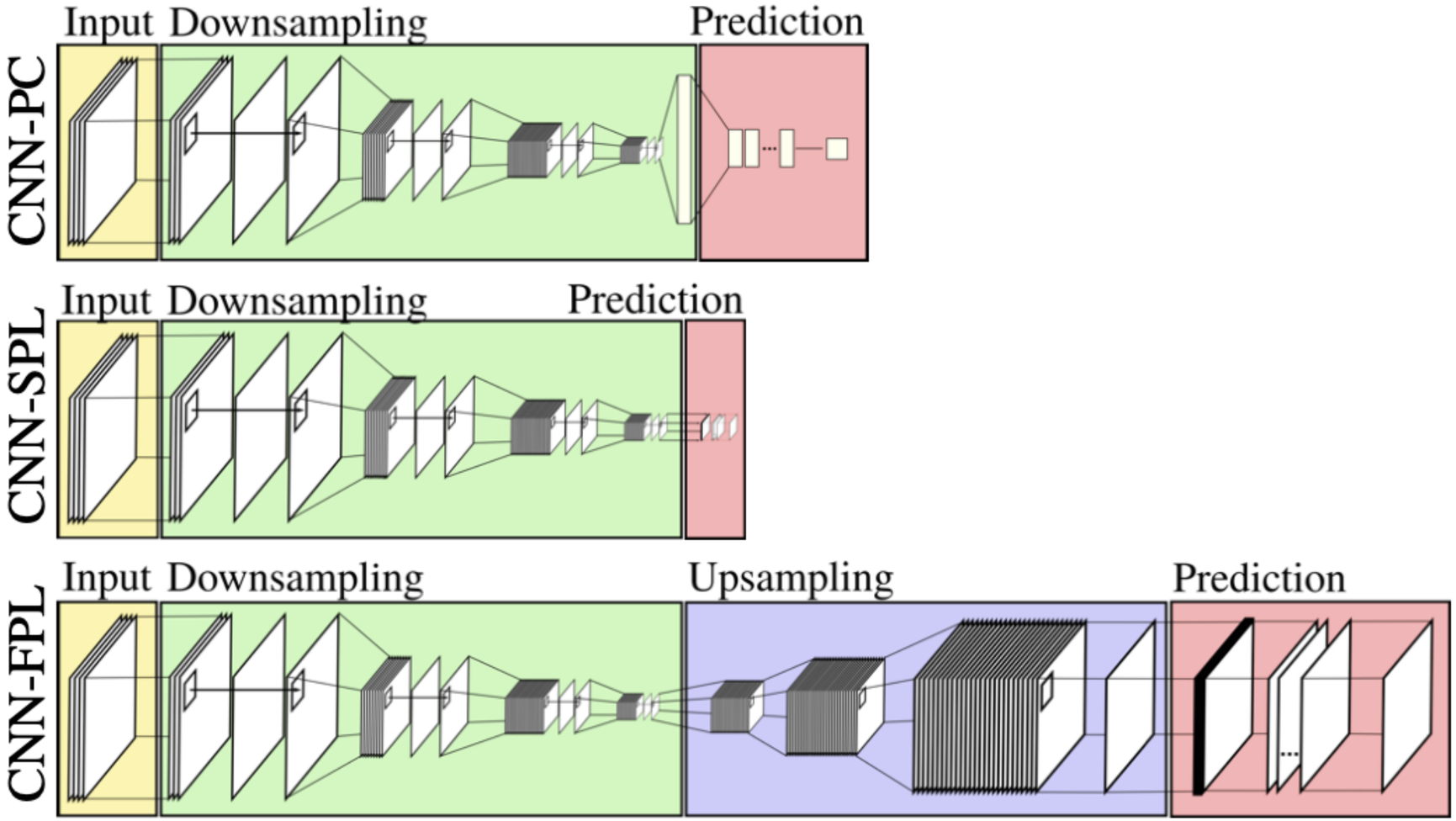
OBJETIVOS DE LOS AUTORES

- Entrenar un sistema **basado en CNN** para tareas de **etiquetado semántico denso** de manera supervisada.
- Trabajar con **imágenes aéreas con ultra-high spatial resolution** (gran cantidad de información geométrica)
- Introducir una red de **deconvolución** basada en parches para codificar primero las representaciones de la cobertura terrestre en un mapa aproximado (cuello de botella) y luego volver a muestrear las características al tamaño del parche de input original.
- Entrenar la red en forma de **parche** permite tratar con imágenes de cualquier tamaño, descomponiendo en subregiones.



Example of (a) convolution and (b) deconvolution.

3 ARQUITECTURAS

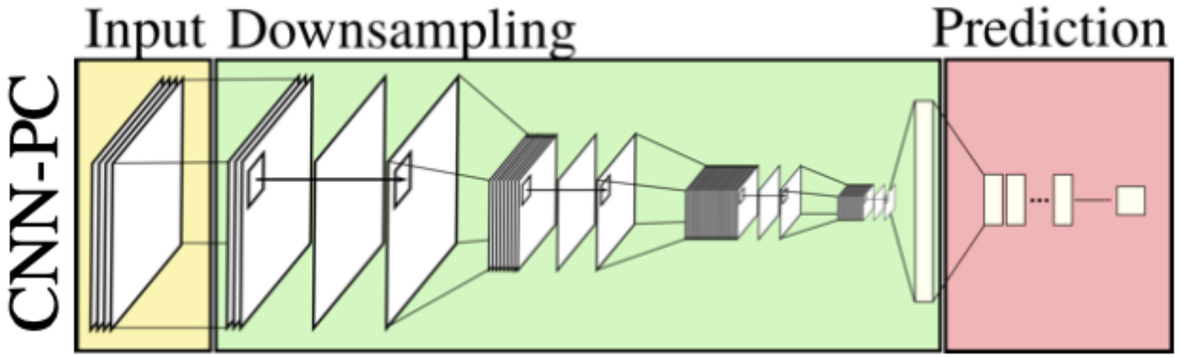


Para mitigar el overfitting:

- Dropout
- Batch normalization
- weight decay
- Data augmentation
  - Random sampling
  - Random transformations
  - Noise Injection

Esquema de las redes utilizadas para realizar el etiquetado semántico.

ARQUITECTURAS BASE: STANDARD PATCH-CLASSIFICATION SYSTEM



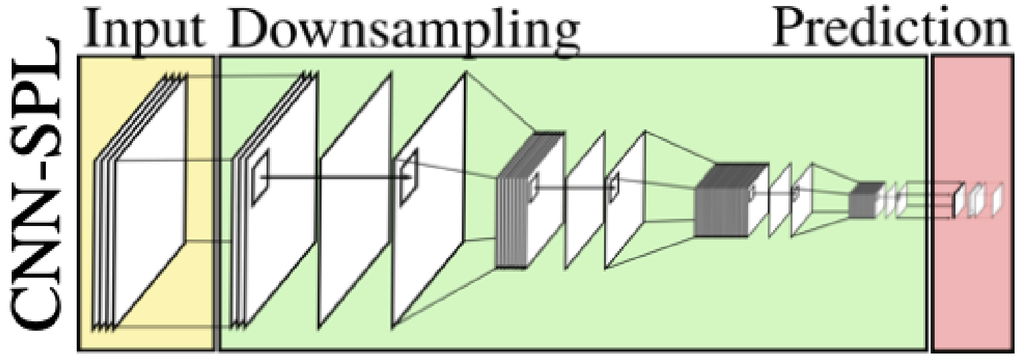
Toma como input un parche y predice la etiqueta del parche del **píxel central.**

Empleamos:

- 64 neuronas en la primera capa (filtros 7 × 7),
- 64 en la segunda (5 × 5),
- 128 en la tercera y finalmente (5 × 5) 256 en la cuarta capa.

La cuarta capa incluye Max-pooling y relu, y luego implementa una fully connected

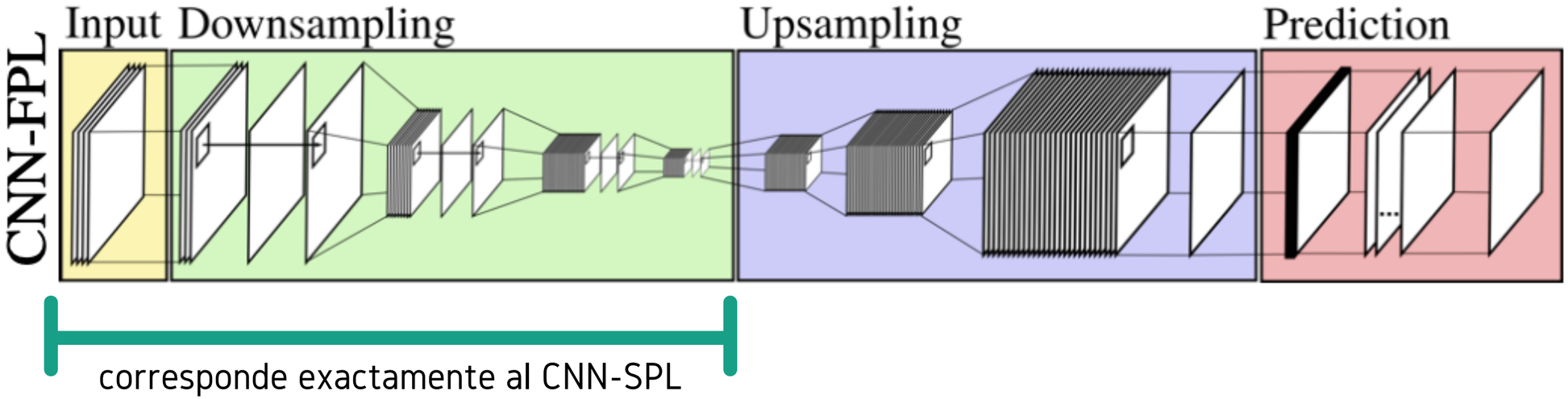
patch classification (CNN-PC)



Predice el etiquetado de un **subparche de tamaño 9 × 9** alrededor del centro del parche de input de 65 × 65. Esta red corresponde exactamente a la CNN-PC, pero se elimina el última Max-pooling y la capa fully connected de la CNN-PC se reemplaza por convoluciones 1 × 1, de modo que la salida ahora es un parche de etiquetas de la capa bottleneck.

subpatch labeling (CNN-SPL)

ARQUITECTURA PROPUESTA FINAL



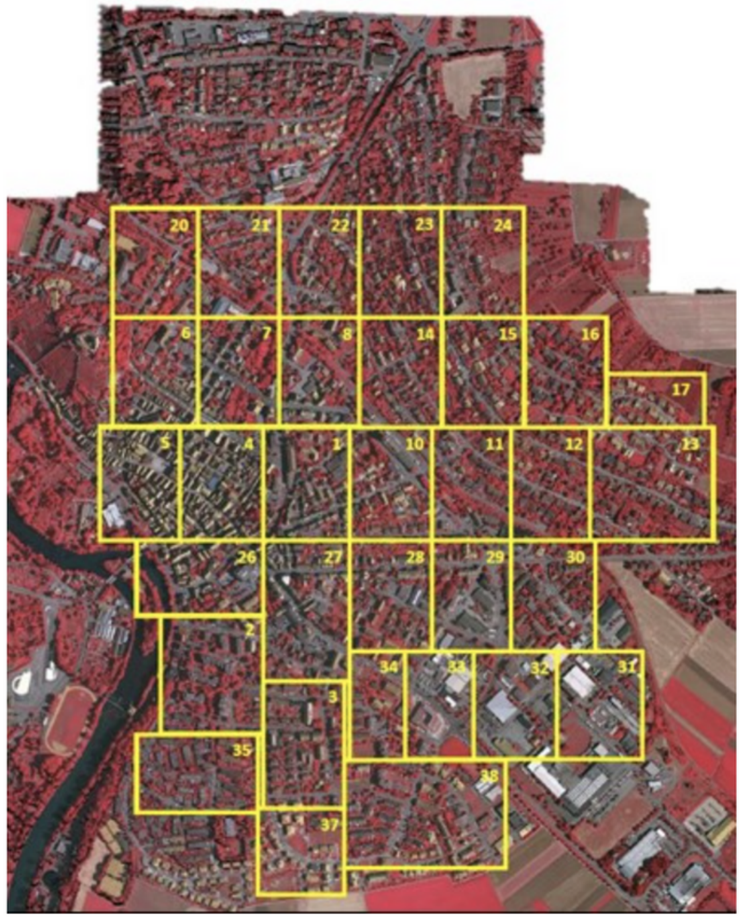
- Está compuesta por dos partes principales: un submuestreo y un bloque de deconvolución
- 3 capas de deconvolución para compensar 3 niveles de downsampling (poling máximo).
- El cuello de botella obliga a la red a aprender una representación espacial aproximada de las clases presentes en el parche.
- CNN-FPL se entrena igual que las arquitecturas anteriores. Sin embargo, el número de parámetros es significativamente mayor.
- Usaron **warm start** empleando pesos de CNN-PC para los bloques de downsampling.

full patch labeling by learned upsampling (CNN-FPL)

2 DATASETS: PARCHES DE INPUT DE 65x65 pixeles

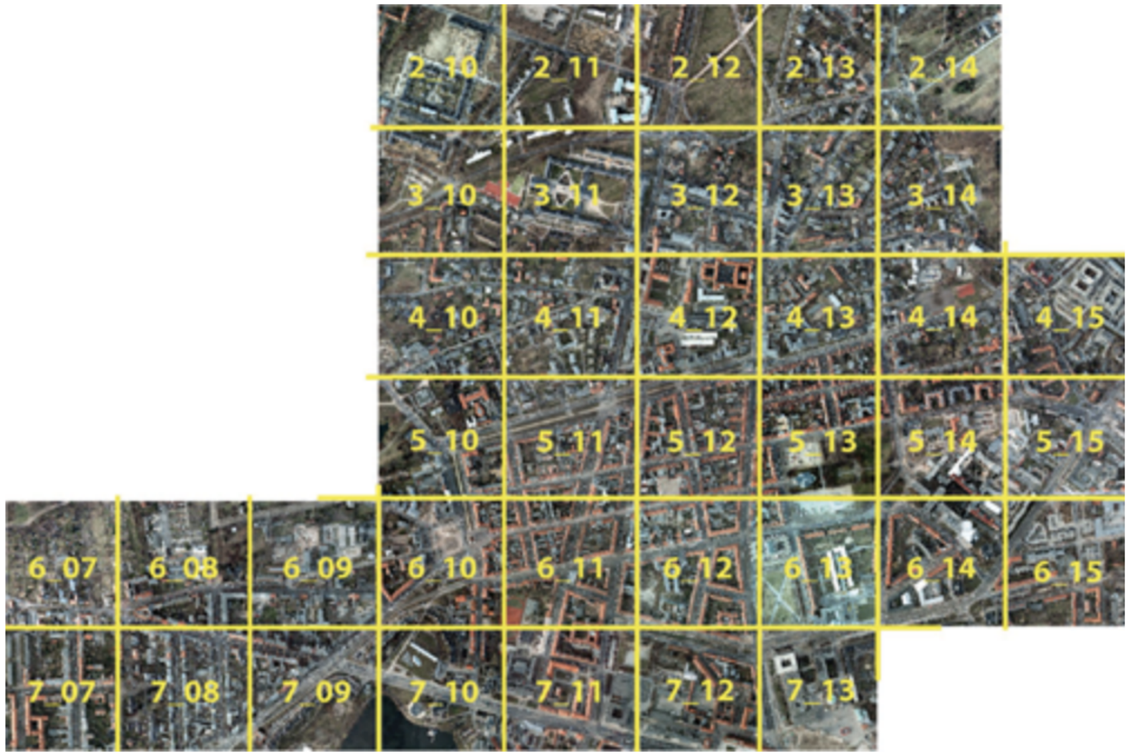
Vaihingen dataset

- Contiene 33 parches de diferentes tamaños
- Distancia de muestreo del suelo: 9cm



Potsdam dataset

- Contiene 38 parches de diferentes tamaños
- Distancia de muestreo del suelo: 5cm



MÉTRICAS

TABLE I
NUMERICAL RESULTS FOR THE VAIHINGEN VALIDATION SET.

	Model	OA	K	AA	F1
full	SP-MSF	79.79	73.28	65.80	66.20
	CNN-PC	82.62	76.97	62.10	64.19
	CNN-SPL	83.04	77.50	64.05	67.89
	CNN-FPL	**83.79**	**78.52**	**69.12**	**73.03**
no bk	SP-MSF	79.86	73.36	74.24	74.40
	CNN-PC	82.68	77.04	74.47	76.95
	CNN-SPL	83.08	77.55	70.76	72.89
	CNN-FPL	**83.83**	**78.57**	**76.45**	**78.76**
er full	SP-MSF	83.64	78.32	70.65	69.92
	CNN-PC	86.67	82.29	65.58	68.01
	CNN-SPL	87.24	83.03	69.17	73.47
	CNN-FPL	**87.80**	**83.79**	**74.94**	**78.60**
er no bk	SP-MSF	83.71	78.39	78.44	77.64
	CNN-PC	86.73	82.36	78.66	81.57
	CNN-SPL	87.28	83.07	75.03	77.78
	CNN-FPL	**87.83**	**83.83**	**81.35**	**83.58**

Emplearon 4 estrategias de evaluación

FULL: todas las clases presentes
NO BK: excluye el background
ER FULL: todas las clases presentes, pero con bordes erosionados
ER NO BK:excluye el background, pero con bordes erosionados

- **Accuracy general (OA)**
- **Kappa (K)**
- **Accuracy promedio de la clase (AA)**
- **puntaje F1 promedio de la clase (F1).**

*Erosión en los bordes (ER): para que sea tolerante a pequeños errores en los bordes de los objetos.

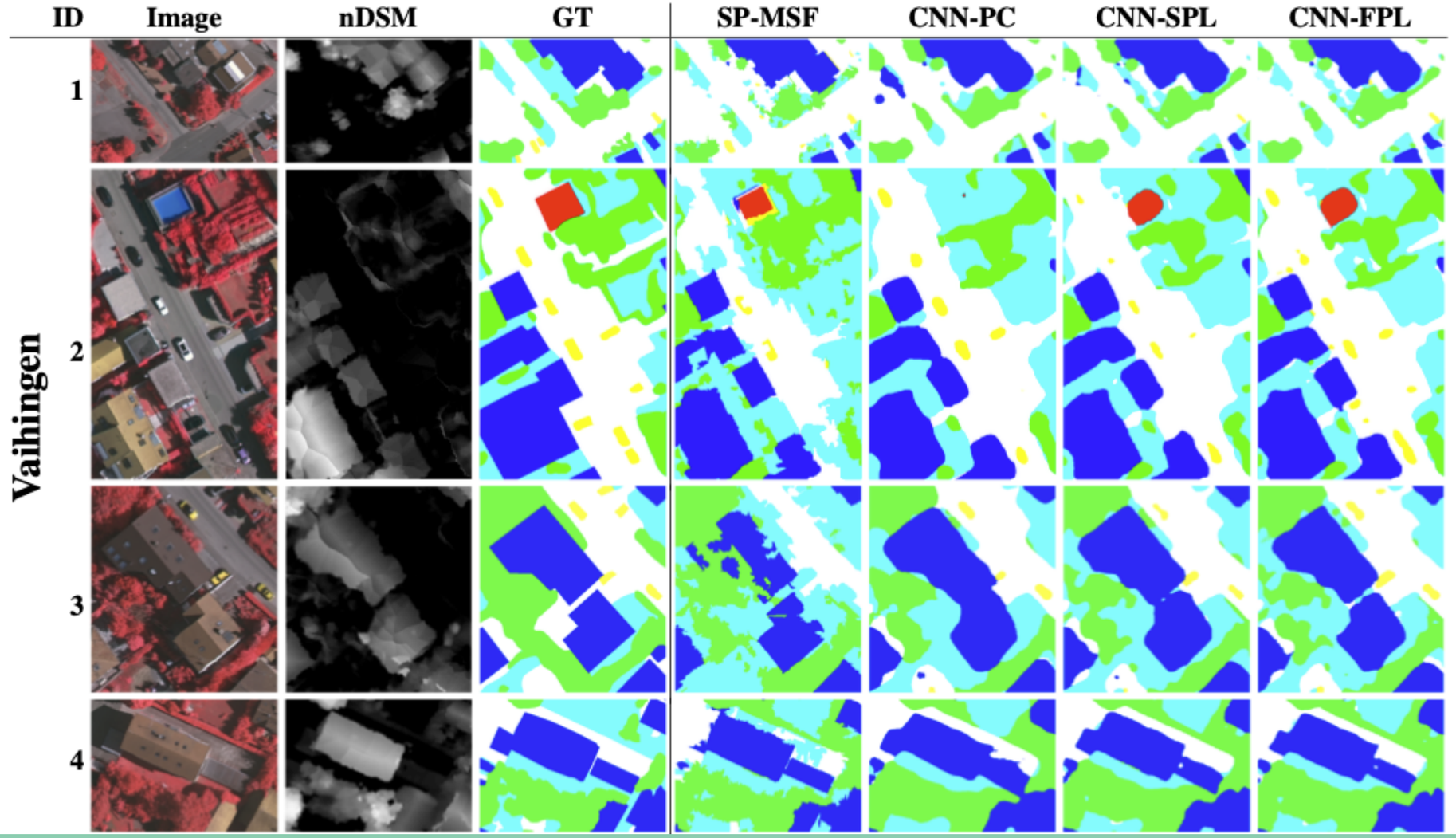


Fig. 4. Example predictions for the Vaihingen (subsets 1-5) and Potsdam (6-9) for the tested architectures. Legend – White: impervious surfaces; Blue: buildings; Cyan: low vegetation; Green: trees; Yellow: cars; Red: clutter, background. Best viewed in color PDF.

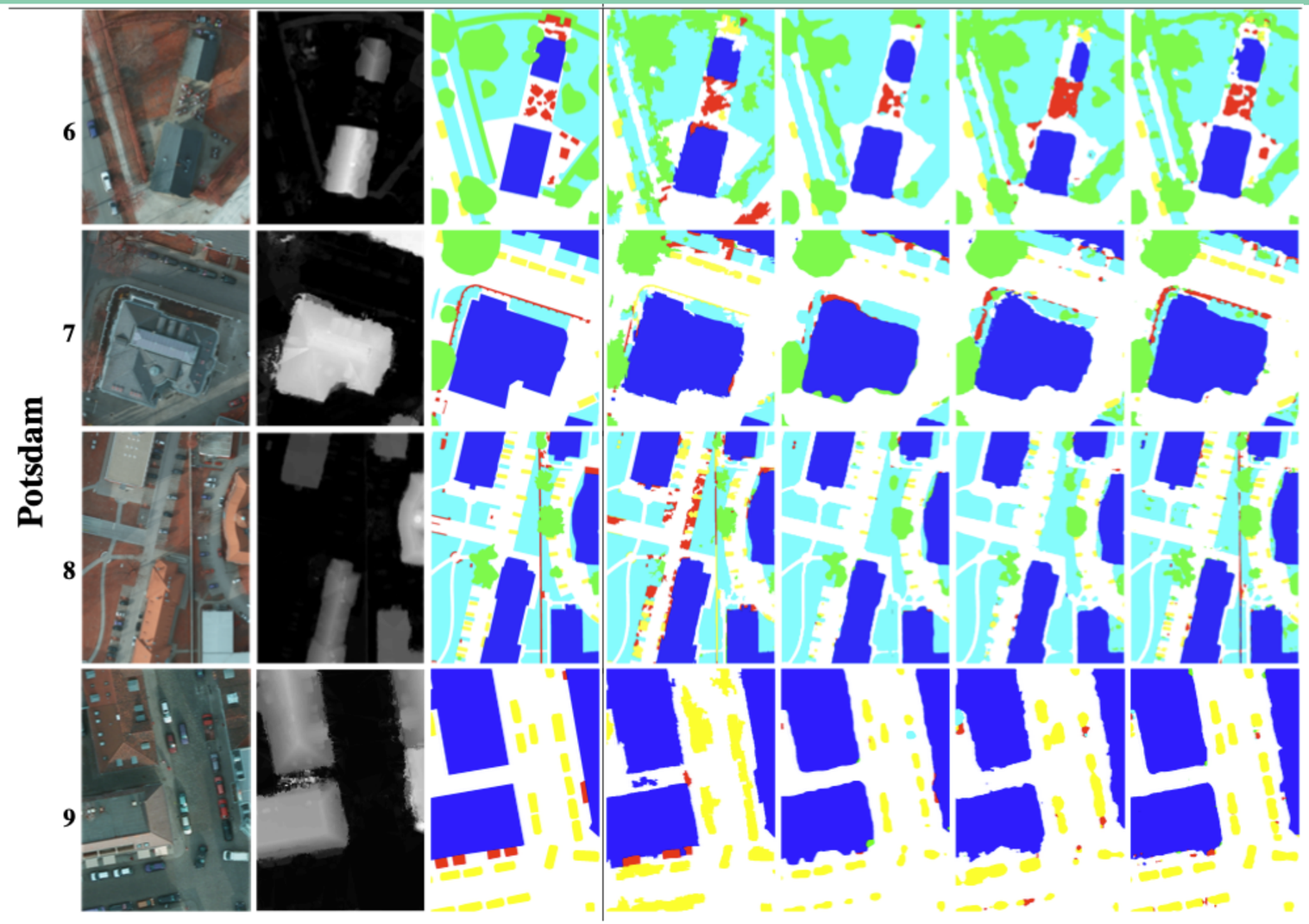


Fig. 4. Example predictions for the Vaihingen (subsets 1-5) and Potsdam (6-9) for the tested architectures. Legend – White: impervious surfaces; Blue: buildings; Cyan: low vegetation; Green: trees; Yellow: cars; Red: clutter, background. Best viewed in color PDF.

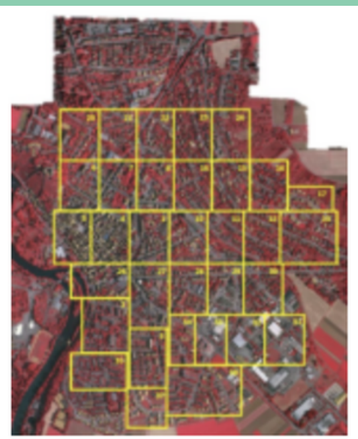
VI. DISCUSSION AND CONCLUSION

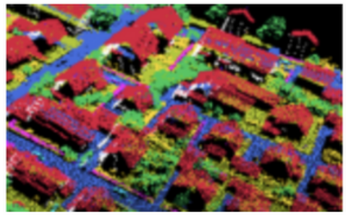
- Los autores propusieron un sistema para realizar un etiquetado semántico denso utilizando redes neuronales convolucionales.

- Predecir la segmentación para parches completos es realmente ventajoso desde las perspectivas de eficiencia y precisión semántica / geométrica.

- Los resultados obtenidos estaban alineados con los modelos state-of-the-art de ese tiempo (2016)

REFERENCIAS





- https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/

- https://www.researchgate.net/publication/305779573_Dense_Semantic_Labeling_of_Subdecimeter_Resolution_Images_With_Convolutional_Neural_Networks

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

DENSE SEMANTIC LABELING OF SUB-DECIMETER RESOLUTION IMAGES WITH CONVOLUTIONAL NEURAL NETWORKS (2016)

Michele Volpi
Devis Tuia - IEEE

CT3115 - Sistemas Urbanos Inteligentes

Pontificia Universidad Católica de Chile, Escuela de Ingeniería
Departamento de Ciencia de la Computación

CAMILA F. VERA VILLA

mayo 6, 2021