IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Diterima 2 Maret 2019, diterima 19 Maret 2019, tanggal publikasi 1 April 2019, tanggal versi saat ini 13 April 2019.

Pengenal Objek Digital 10.1109/ACCESS.2019.2908685

Jaringan Konvolusional Padat untuk Semantik Segmentasi

CHAOYI HAN, YIPING DUAN, TAO XIAOMING, (Anggota, IEEE), DAN JIANHUA LU, (Fellow, IEEE)

Departemen Teknik Elektronika, Universitas Tsinghua, Beijing 100084, China
Pusat Penelitian Nasional Beijing untuk Sains dan Teknologi Informasi (BNRist), Universitas Tsinghua, Beijing 100084, Tiongkok
Pusat Inovasi Beijing untuk Chip Masa Depan, Beijing 100084, Tiongkok

Penulis koresponden: Xiaoming Tao (taoxm@mail.tsinghua.edu.cn)

Karya ini didukung sebagian oleh National Science Science Foundation of China (NSFC) di bawah Hibah 61622110 dan Hibah 61801260, dan sebagian oleh Yayasan Ilmu Postdoctoral China di bawah Hibah 2017M620044 dan Hibah 2018T110098.

ABSTRAK Studi terbaru telah sangat mempromosikan pengembangan segmentasi semantik. Sebagian besar metode canggih mengadopsi jaringan konvolusional penuh (FCN) untuk menyelesaikan tugas ini, di mana lapisan yang terhubung sepenuhnya diganti dengan lapisan konvolusi untuk prediksi padat. Namun, konvolusi standar memiliki kemampuan terbatas dalam menjaga kontinuitas antara label yang diprediksi serta memaksa kelancaran lokal. Dalam makalah ini, kami mengusulkan unit konvolusi padat (DCU), yang lebih cocok untuk klasifikasi berdasarkan piksel.
DCU mengadopsi prediksi padat alih-alih cara prediksi tengah yang digunakan dalam lapisan konvolusi saat ini.
Label semantik untuk setiap piksel disimpulkan dari prediksi tengah/luar tengah yang tumpang tindih dari perspektif probabilitas. Ini membantu untuk menggabungkan konteks dan menyematkan koneksi antara prediksi, sehingga berhasil menghasilkan peta segmentasi yang akurat. DCU berfungsi sebagai lapisan klasifikasi dan merupakan opsi yang lebih baik daripada konvolusi standar di FCN. Teknik ini dapat diterapkan dan bermanfaat untuk metode canggih berbasis FCN dan bekerja dengan baik dalam menghasilkan hasil segmentasi. Eksperimen ablasi pada kumpulan data pembanding memvalidasi keefektifan dan kemampuan generalisasi dari pendekatan yang diusulkan dalam tugas segmentasi semantik.

ISTILAH INDEKS Unit konvolusi padat, jaringan konvolusional penuh, prediksi tumpang tindih, segmentasi semantik.

I. PENDAHULUAN

Segmentasi semantik berfungsi sebagai bagian tak terpisahkan dalam analisis konten gambar dan video. Ini bertujuan untuk menetapkan setiap piksel tag semantik yang telah ditentukan dan telah lama menjadi masalah yang sangat menantang karena variabilitas data intra-kelas yang tinggi. Baru-baru ini, banyak hasil yang menjanjikan telah dilaporkan mengenai tugas prediksi yang padat ini, terutama diuntungkan dari kemajuan dalam jaringan saraf konvolusional yang dalam (DCNNs) [1]–[8]. Sebagian besar metode tercanggih [9]–[13] menganggap segmentasi semantik sebagai masalah klasifikasi berdasarkan piksel dan mengadopsi jaringan konvolusional penuh (FCN) [14] untuk menyelesaikannya. Performa luar biasa telah dicapai pada berbagai dataset benchmark [15]–[23].

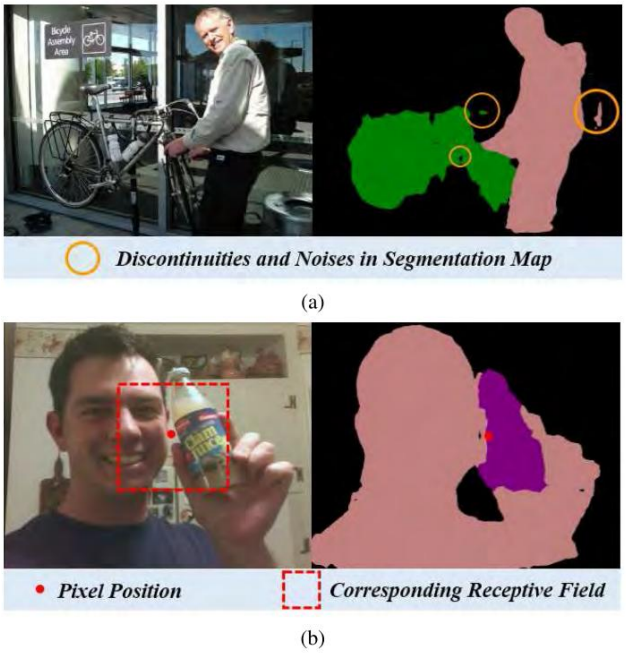
Diinisialisasi dengan parameter DCNN yang dilatih sebelumnya pada kumpulan data berskala besar seperti ImageNet [24], FCN melakukan relatif baik pada tugas segmentasi semantik dengan hanya lapisan konvolusi tambahan yang berfungsi sebagai pengklasifikasi [14]. Kemampuan superior DCNN dalam mengekstraksi semantik tingkat tinggi menentukan bahwa bahkan jaringan end-to-end yang dirakit secara langsung dapat lebih efektif daripada metode yang dirancang khusus dengan fitur buatan tangan.
Oleh karena itu, sebagian besar karya berikut mewarisi semangat belitan penuh dan beralih ke penemuan struktur yang lebih cocok dan efektif untuk jaringan prediksi padat. Di antara studi tersebut, dilatasi konvolusi dan dekonvolusi telah menjadi teknik yang sangat populer karena kinerjanya yang luar biasa dalam menghasilkan peta fitur beresolusi tinggi. Jaringan dasar yang lebih kuat seperti ResNet [2] lebih lanjut mempromosikan pengembangan komunitas segmentasi semantik. Saat ini, kerangka kerja berbasis FCN khas dapat dilihat di keluarga Deeplab [10], [25].

Namun demikian, ada dua masalah yang tidak harmonis antara FCN dan tugas segmentasi semantik. Pertama, tata ruang

Associate editor yang mengoordinasikan review naskah ini dan menyetujuinya untuk diterbitkan adalah Michele Magno.

hubungan konteks antara piksel tidak secara eksplisit digambarkan dalam FCN. Mulai dari jaringan klasifikasi yang dalam, FCN mengganti beberapa lapisan terakhir yang terhubung sepenuhnya dengan lapisan konvolusi dan mengoptimalkan seluruh jaringan dengan fungsi kehilangan piksel. Akibatnya, proses prediksi label untuk satu piksel sebagian besar tidak bergantung pada yang lain. Hal ini tidak dapat dihindari menyebabkan banyak diskontinuitas dan noise pada peta segmentasi, seperti yang ditunjukkan pada Gambar 1(a). Untuk menghilangkan diskontinuitas dan kebisingan tersebut, banyak pendekatan telah dikembangkan. Bidang acak bersyarat (CRF) adalah model yang diadopsi secara luas dalam metode tradisional. CRF dapat, sampai batas tertentu, menyempurnakan peta segmentasi, tetapi tidak nyaman bagi mereka untuk dilatih secara end-to-end dengan FCN [26], [27]. Metode lain yang terbukti penting dan cukup efektif adalah memasukkan lebih banyak konteks global ke dalam lapisan klasifikasi akhir [11].

Masalah lainnya terkait dengan mode komputasi yang kaku dari konvolusi 2D standar. Dalam FCN saat ini, konvolusi dilakukan melalui jaringan reguler yang telah ditentukan sebelumnya. Di dalam lapisan yang sama, bidang reseptif untuk semua konvolusi sebenarnya sama. Mekanisme ini menyimpulkan bahwa aktivasi dihasilkan melalui cara yang identik, default, dan tidak dapat diubah. Meskipun tumpukan lapisan konvolusi menunjukkan kemampuan representasi yang kuat, peningkatan kedalaman tidak dapat sepenuhnya mengimbangi kelemahan ini. Seperti yang ditunjukkan dalam [28], konvolusi standar tidak dapat beradaptasi dengan batas semantik. Ini membatasi kinerja jaringan saraf convolutional (CNN). Dalam jaringan segmentasi semantik, cukup banyak prediksi dilakukan pada area yang samar-samar semantik, seperti yang diilustrasikan pada Gambar 1(b).



GAMBAR 1. Di FCN, prediksi untuk setiap piksel relatif independen. Oleh karena itu, banyak diskontinuitas dan 'noise' yang ada di peta segmentasi. Selain itu, karena bidang reseptif konvolusi 2D standar yang kaku, banyak prediksi yang tidak dapat diandalkan. (a) Diskontinuitas dan kebisingan. (B) daerah semantik kabur.

Singkatnya, masalah utama terletak pada pengklasifikasian FCN, yang saat ini dilakukan oleh lapisan konvolusi tunggal. Dalam jaringan klasifikasi, lapisan yang terhubung sepenuhnya atau lapisan penyatuan global [2], [29] membantu memungkinkan penyematan konteks yang memadai di seluruh peta fitur, yang sangat penting untuk prediksi yang akurat. Namun, rekan mereka di FCN tidak memiliki pandangan global seperti itu. Mode konvolusi penuh menentukan bahwa hanya sebagian dari peta input yang terlibat dalam proses prediksi label untuk piksel. Kami mengamati bahwa peta probabilitas yang dihasilkan oleh FCN tipikal hanya menunjukkan tingkat kepercayaan yang tinggi pada piksel di dekat pusat objek (area 'kanan'). Rupanya, banyak prediksi tidak dibuat di area yang sesuai secara semantik. Keputusan tentang piksel yang ada di antara objek sama acaknya dengan lemparan dadu.

Untuk mengatasi masalah di atas, kami mengusulkan skema "konvolusi padat" dan mengembangkan unit konvolusi padat (DCU) yang sesuai sebagai pengklasifikasi yang lebih baik untuk jaringan saat ini. Seperti namanya, konvolusi padat menghasilkan keluaran padat untuk satu bidang reseptif. Ini memungkinkan prediksi di luar pusat yang tidak tersedia dalam konvolusi standar. Label semantik terakhir untuk setiap piksel disimpulkan dari prediksi pusat/luar pusat tersusun. Gagasan 'padat' serupa dapat dilihat di [30]–[32]. Dengan mekanisme ini, bidang reseptif yang "benar" tidak hanya menentukan label semantik dari piksel pusatnya, tetapi juga tetangganya. Jadi, piksel diimbangi dari posisi yang tepat dapat memiliki kesempatan untuk menyempurnakan label mereka lebih lanjut. Kontribusi dari makalah ini antara lain sebagai berikut: • Kami memberikan penjelasan yang komprehensif tentang motivasi dan mentalitas jaringan konvolusional padat dan menyajikan struktur jaringan secara rinci. • Formulasi lengkap dan algoritma menyeluruh tentang pelatihan jaringan konvolusional padat dikembangkan.

Analisis teoritis mengenai desain peta berat di DCU dan ilustrasi visual yang sesuai juga disertakan.

• Kami melakukan eksperimen ekstensif pada kumpulan data tolok ukur dan melaporkan hasil yang sesuai. Studi dan eksperimen ablasi menunjukkan kinerja superiornya dibandingkan konvolusi standar dan pesaing lainnya.

Versi konferensi awal dari pekerjaan kami dapat dilihat di [33]. Dalam makalah ini, hubungan dan perbedaan dengan karya yang ada ditambahkan untuk menunjukkan sebab dan akibat. Selain studi ablasi asli pada model dasar, kami menggeneralisasikan DCU ke jaringan terkenal lainnya serta kumpulan data tambahan dan mencapai peningkatan yang signifikan. Berikut ini, pertama-tama kami akan memberikan ulasan untuk pekerjaan terkait di Bagian II. Kemudian, kami menyajikan detail teknis di Bagian III dan menunjukkan bahwa konvolusi padat dapat diimplementasikan secara efisien dan bekerja dengan baik pada tugas segmentasi semantik. Hasil eksperimen dan diskusi terkait diberikan di Bagian IV. Kesimpulan dan pekerjaan masa depan ditarik dalam Bagian V.

II. PEKERJAAN

TERKAIT Perbaikan terbaru pada segmentasi semantik sebagian besar dapat dikaitkan dengan penerapan jaringan konvolusional penuh (FCNs) [14]. Jaringan saraf convolutional tipikal (CNN) yang digunakan untuk tugas klasifikasi diubah menjadi jaringan prediksi padat dengan cara langsung.

Jaringan seperti itu menunjukkan kemampuan luar biasa pada tugas segmentasi semantik, sehingga telah menjadi paradigma dalam metode canggih. Sementara itu, banyak karya berfokus pada kelemahan intrinsik dari struktur konvolusional penuh dan mengembangkan banyak varian, yang diringkas sebagai berikut.

A. PEMULIHAN RESOLUSI

Jaringan dalam biasanya mengadopsi lapisan penyatuan berurutan untuk menyematkan lebih banyak konteks, serta untuk mengurangi persyaratan komputasi yang tinggi, yang telah terbukti penting dalam pemrosesan gambar [1]–[3]. Namun, metode berbasis FCN banyak mengalami downsampling karena hilangnya resolusi gambar. Untuk mengurangi kerugian tersebut, dilatasi konvolusi diusulkan dan telah diadopsi secara luas [34]. Itu dapat memperbesar bidang reseptif dengan memasukkan lubang (nol) ke filter. Dengan menghilangkan pooling layer dan meningkatkan dilation rate dari subsequent convolutions, fitur resolusi dapat dipertahankan tanpa mengubah struktur dan ukuran parameter model. Dekonvolusi [35]–[37] adalah metode populer lainnya untuk memulihkan resolusi. Ini dapat dianggap sebagai kebalikan dari konvolusi. Dekonvolusi mewujudkan upampling menggunakan bobot yang dapat dipelajari alih-alih metode nonparametrik (mis., interpolasi bilinear) dan menunjukkan kinerja yang unggul. Ide serupa dapat ditemukan di jaringan encoder-decoder [38].

B. PENYEMATAN KONTEKS MULTISKALA

Tidak seperti jaringan klasifikasi tipikal yang inputnya disetel ke ukuran tetap, jaringan prediksi padat menerima gambar dengan ukuran acak. Sayangnya, input mutatif membatasi keakuratan segmentasi karena jaringan saraf tidak dapat beradaptasi dengan skala yang berbeda secara efektif. Sebagian besar karya yang ada memilih untuk menanamkan konteks multiskala ketika berhadapan dengan masalah ini [39], [40]. Di antara metode tersebut, pengujian multiskala bersifat langsung dan mudah dilakukan. Metode ini mengubah ukuran gambar input ke skala yang berbeda kemudian menerapkan jaringan terlatih (sama atau berbeda) secara terpisah. Hasil akhir diperoleh dari kombinasi peta segmentasi dalam beberapa skala. Struktur piramid fitur [10] menggabungkan operasi penyisipan multiskala ke dalam jaringan dengan cara end-to-end. Pooling piramida spasial (SPP) [11] berbagi ide yang sama dan memperoleh margin yang signifikan dibandingkan dengan metode yang ada. Usulan Scale Normalization for Image Pyramids (SNIP) baru-baru ini [41] terbukti lebih baik menangani objek dengan skala kecil atau besar, tetapi belum diterapkan dalam tugas segmentasi semantik.

C. PENYEMPURNAAN BATAS

Performa segmentasi yang buruk di sepanjang batas objek mungkin menjadi masalah umum untuk semua pembelajaran mendalam metode. Kebalikan dari konteks global, informasi lokal tidak ditangkap dengan baik oleh DNN. Meskipun peningkatan kapasitas jaringan (kedalaman) mengarah pada hasil yang terus-menerus lebih baik [10], biaya respons kor dalam runtime dan memori akan segera menjadi tak tertahankan. Oleh karena itu, dalam banyak karya canggih, CRF diadopsi sebagai pasca-pemrosesan karena kinerjanya yang luar biasa di sepanjang batas objek. Pelatihan CRF end-to-end dengan DNN juga telah dipelajari dalam banyak karya [26], [27], tetapi pemodelan khusus dan metode inferensi hampir tidak dapat diperluas ke jaringan lain. Cara yang lebih populer cenderung menambahkan unit penyempurnaan batas ke dalam FCN [12], [42]. Biasanya, ukuran kernel kecil diadopsi untuk karakterisasi konteks lokal yang lebih baik. Namun, karena batas menyumbang sebagian kecil dari keseluruhan gambar, metode tersebut membawa perbaikan terbatas dibandingkan dengan teknik seperti dilatasi konvolusi, SPP, dan sebagainya.

D. EVOLUSI STRUKTUR

Studi terbaru memikirkan kembali struktur DCNN untuk menyematkan semantik tingkat tinggi. Convnet yang dapat dideformasi [28] menambah lokasi pengambilan sampel tetap tersebut di jaringan tipikal dengan parameter tambahan. Sebagai metode berbasis data, ini dapat beradaptasi dengan skala dan bentuk objek, yang tidak dapat dicapai dalam konvolusi standar. Konvolusi upsampling padat [32] membagi seluruh peta keluaran menjadi beberapa subbagian yang sama dan menghasilkan setiap bagian secara bersamaan menggunakan saluran yang berbeda. Ide serupa dapat ditemukan di FCIS [30] yang digunakan dalam segmentasi instan, di mana saluran yang berbeda menyandikan fitur sensitif posisi dan kemudian digabungkan untuk mendapatkan peta segmentasi akhir.

Metode semacam ini mengeksploitasi beberapa struktur spesifik untuk jaringan prediksi padat yang dimotivasi oleh berbagai observasi atau asumsi. Keberhasilan mereka menunjukkan bahwa, meskipun tugas klasifikasi ditransfer secara efisien ke segmentasi semantik, desain khusus tugas masih sangat diperlukan untuk terobosan lebih lanjut.

Kami akan menyebutkan bahwa sebagian besar metode di atas telah diuji pada beberapa dataset benchmark untuk perbandingan yang adil. Namun, segmentasi yang sepenuhnya dianotasi dalam kumpulan data yang ada relatif jarang, sehingga melatih DNN dari awal menjadi sulit.

Pertunjukan canggih dicapai melalui pembelajaran transfer berdasarkan jaringan klasifikasi. Oleh karena itu, metode yang mengadopsi pembelajaran dengan pengawasan lemah dan semi pengawasan [43]–[47] memiliki masa depan yang sangat menjanjikan karena mereka dapat memanfaatkan data dengan anotasi yang lemah atau bahkan tanpa anotasi. Dalam hal itu, fungsi kerugian yang dirancang khusus dan pelatihan rekursif diadopsi secara luas. Rincian lebih lanjut dapat ditemukan di referensi.

Makalah ini dekat dengan karya-karya tentang evolusi struktur dalam hal motivasi. Kami berpendapat bahwa klasifikasi harus dilakukan pada area semantik yang sesuai sehingga tingkat kepercayaan yang tinggi dapat dipertahankan. Namun, berbeda dengan convnet yang dapat dideformasi di mana setiap aktivasi mencoba menemukan area semantiknya, konvolusi padat memungkinkan area semantik menghasilkan lebih banyak keputusan dan menyebarkan hasilnya melalui peta bobot berdasarkan piksel. Saluran ekstra di jaringan kami tidak dimaksudkan untuk menyandikan lebih banyak fitur untuk konten yang sama, yang lebih seperti over-fitting, tetapi dibuat untuk menyandikan konten yang berbeda.

Makalah ini mencoba untuk menunjukkan bahwa tidak semua bidang reseptif memiliki kepentingan yang sama ketika memprediksi label semantik dan prediksi offset dari "area tengah" lebih kredibel daripada prediksi pusat dari "area offset".

AKU AKU AKU. PENDEKATAN YANG DIUSULKAN

Jaringan saraf convolutional membuka jalan bagi keberhasilan penerapan jaringan saraf tiruan untuk visi komputer. Konvolusi (terkadang disingkat sebagai konv) memanfaatkan properti lokal di dalam gambar dan banyak digunakan dalam pemrosesan gambar digital. Lapisan konvolusi standar melakukan proses berikut (peta fitur dan vektor ditampilkan dalam huruf tebal):

$$\mathbf{y}^{(o)}_{h,w} = \sum_{im} \ddot{\mathbf{y}}^{(i,o)}_{MN} \cdot \mathbf{x}^{(i)}_{h+m,w+n} + b^{(o)} \qquad (1)$$

di mana ***x*** dan ***y*** menunjukkan peta fitur input dan output, **ÿ** menunjukkan bobot kernel conv, i, *o* mewakili saluran yang berbeda dan *h, w,* m, *n* mewakili indeks dalam peta fitur dan kernel conv yang sesuai. Jaringan end-to-end sederhana yang dapat dilatih untuk segmentasi semantik kemudian diselesaikan setelah lapisan kerugian diterapkan setelahnya. Fungsi kerugian softmax untuk jaringan di atas didefinisikan sebagai:

$$L_{softmax} = -\frac{1}{N}\sum_{hw} \log(\exp(y^{label}_{h,w}) / \sum_{Hai} \exp(y^{(o)}_{h,w})) \qquad (2)$$

di mana *label* menunjukkan label kebenaran dasar pada posisi *(h,w), N* adalah jumlah elemen dalam peta segmentasi dan sama dengan *h* · w.

A. PREDIKSI PADAT BUKAN PREDIKSI PUSAT

Dari perspektif pemrosesan sinyal, lapisan konvolusi terakhir di FCN memberikan probabilitas kelas *o* pada posisi *(o) (h,w)*: *P* h,w . Serupa dengan prediksi maju dan prediksi mundur *h,w* , yang umum dalam pemrosesan sinyal, kita dapat menyebut cara prediksi konvolusi Secara intuitif, korespondensi seperti itu sebagai prediksi pusat. prediksi padat seharusnya tidak dibatasi sedemikian rupa. Faktanya, korespondensi antara satu piksel dan bidang reseptifnya yang "sempurna" sangat bervariasi ketika terletak pada posisi objek yang berbeda. Untuk lebih memanfaatkan kemampuannya, kami memperluas konvolusi standar dan menyimpulkan bentuk umum sebagai konvolusi padat.

Untuk presentasi yang jelas, pada bagian ini kami menggunakan notasi yang mirip dengan [28], yang menyatakan lokasi 2D (h,w) dalam bentuk vektor dan mengabaikan dilatasi dan langkah untuk kesederhanaan. Maka konvolusi standar pada peta masukan tunggal ***x*** dapat dirumuskan kembali sebagai berikut:

$$\mathbf{y(p_0)} = \sum_{saya=1}^{k^2} \ddot{\mathbf{y}}(\mathbf{p_i}) \cdot \mathbf{x}(\mathbf{p_0} + \mathbf{p_i}) \qquad (3)$$

di mana p0 menunjukkan lokasi di ***x*** dan p_i mewakili item ke-i dari yang berikut:

$$\{(\ddot{y}\,\tfrac{k}{2}, \ddot{y}\,\tfrac{kk}{2}), (\ddot{y}\,\tfrac{2}{}, \tfrac{kk}{2} + 1), \ldots, (\tfrac{2}{2}\,\tfrac{kk}{}\,\ddot{y}\,1), (\tfrac{2}{2}, \tfrac{k}{2})\}$$

yang menyebutkan semua kemungkinan lokasi dalam kernel konvolusi *k* × *k* . Dalam praktik umum, k diset ke bilangan ganjil dan k/2 dibulatkan ke bawah menjadi bilangan bulat. Biasanya bantalan akan diadopsi jika terjadi ketidaksesuaian dimensi. input diambil sampelnya dan

Dalam setiap proses konvolusi, k^2 digabungkan dengan parameter yang dapat dipelajari bersama. Setiap kisi *k* × *k* yang disampel sama dalam ***x*** bertanggung jawab atas satu aktivasi dalam ***y*** yang terletak di pusatnya. Rupanya, jumlah masukan dan keluaran dalam konvolusi standar adalah *k* Konvolusi padat yang diusulkan ^{2} lawan 1. memperkenalkan keluaran tambahan ke dalam 2 proses di atas. Untuk kisi *k* × *k* ***R*** dalam x, ini menghitung aktivasi *k* (juga kisi):

$$y_j = \sum_{saya=1}^{k^2} \ddot{y}_j(p_i) \cdot \mathbf{R}(\mathbf{p_i}), \; j = 1, 2, \ldots, k^2 \qquad (4)$$

di mana $\ddot{y}_j$ menunjukkan bobot kernel jth conv. Jika melakukan konvolusi seperti itu pada seluruh peta input ***x*** dan merakit aktivasi ke-j, kita mendapatkan k^2 peta keluaran:

$$y_j\,\mathbf{(p_0)} = \sum_{saya=1}^{k^2} \ddot{y}_j(p_i) \cdot \mathbf{x}(\mathbf{p_0} + p_i + p_j) \qquad (5)$$

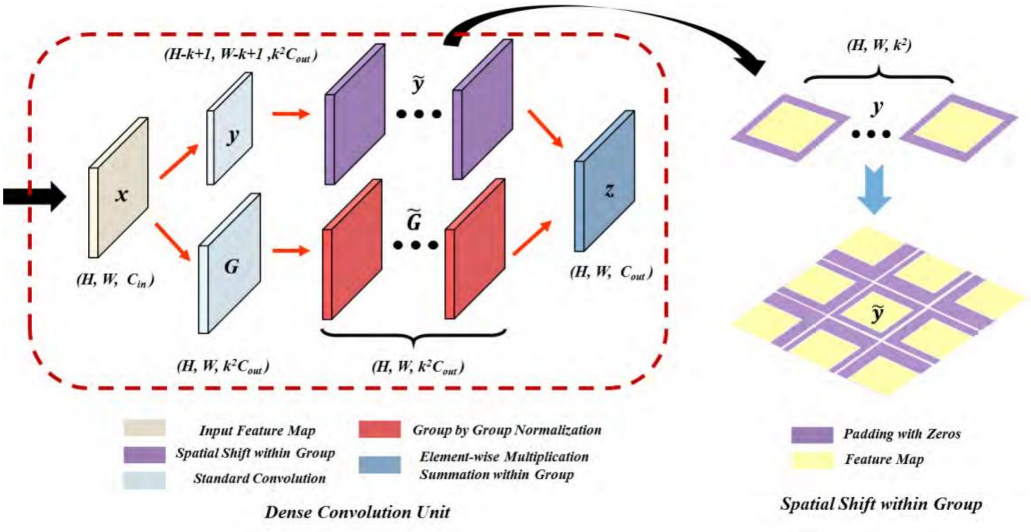
di mana p_j memiliki arti yang sama dengan p_i dan mengkodekan pergeseran spasial dari y_j ke y. Peta di y_j memiliki ukuran yang sama tetapi tidak selaras secara spasial karena mengandung elemen yang berbeda dari kisi keluaran.

Tidak perlu meningkatkan saluran output dari lapisan konvolusi yang ada *k* kali untuk mencapai prediksi yang begitu padat. Jumlahnya, kita semua ^{2} dalam setiap lapisan, oleh karena itu, kita yang perlu dilakukan adalah membagi saluran menjadi beberapa kelompok dan memelihara *k* saluran dalam setiap kelompok. Di setiap ^{2} saluran grup, *k* dapat dilihat sebagai prediksi padat dari satu peta fitur. Untuk mengaktifkan ini, kami mengatur ulang posisi relatif saluran tersebut, seperti yang diilustrasikan dalam bagian "Pergeseran Spasial dalam Grup" pada Gambar. 2. Dengan cara ini, kami mengubah cara prediksi tengah dari konvolusi standar sambil mempertahankan parameter serupa model yang ada. Pembagian seperti itu mencegah pertumbuhan cepat ukuran parameter saat *k* meningkat.

Setelah melakukan ini, konvolusi padat membawa masalah nyata yang disebut tumpang tindih spasial. Masalah ini muncul ketika peta keluaran tumpang tindih satu sama lain. Sebenarnya, tumpang tindih seperti itu dapat dihindari dengan meningkatkan langkah konvolusi ke k. Namun, dalam keadaan seperti itu, proses di atas menjadi konvolusi blok-ke-blok yang menunjukkan partisi artifisial yang jelas. Oleh karena itu, kami mempertahankan langkahnya tidak berubah, yang disetel ke 1 di lapisan klasifikasi tipikal, dan menemukan strategi untuk menangani peta yang tumpang tindih dan menghasilkan hasil akhir.

B. STRATEGI KOMBINASI: DARI PERSPEKTIF PROBABILITAS

Seperti yang ditunjukkan di atas, konvolusi padat dikembangkan sebagai perpanjangan dari konvolusi standar. Perhatikan bahwa, pada lapisan konvolusi asli, aktivasi menunjukkan probabilitas untuk *(o) (o)* setiap kelas *o* pada posisi *(h,* w): *P* h,w . Demikian pula, multioutput dari konvolusi padat juga diharapkan



GAMBAR 2. Struktur unit konvolusi padat (DCU) ketika k = 3. Semua operasi dilakukan kelompok demi kelompok. Setiap grup memiliki peta k2 bobot dan sesuai ke peta fitur satu. Tiga kali lipat merupakan operasi pergeseran spasial untuk peta fitur yang sesuai dan Untuk ilustrasi yang lebih baik, peta fitur mereka diatur secara spasial untuk menunjukkan perbedaan padding. Terbaik dilihat dalam warna.

mewakili semacam probabilitas. Karena aktivasi yang tumpang tindih tersebut memprediksi probabilitas yang sama tetapi dilakukan pada kisi pengambilan sampel yang berbeda (bidang reseptif), masuk akal untuk melihat aktivasi yang tumpang tindih sebagai probabilitas yang dikondisikan *(o)* pada input yang berbeda. Dalam hal ini, hasil akhir $P_{h,w}$ harus dihitung mengikuti rumus probabilitas total:

$$P^{(Hai)}_{h,w} = \sum_j G^{(o)}_{h,w,j} P^{(o)}_{h,w|j}. \tag{6}$$

Di sini, probabilitas bersyarat $P^{(Hai)}_{h,w|j}$ diwakili oleh *(o)* peta yang tumpang tindih y_j (h,w) dan probabilitas $G_{h,w,j}$ bertindak sebagai koefisien yang dinormalisasi. Mempertimbangkan bahwa lokasi yang berbeda dapat menyandikan semantik yang berbeda dan dengan demikian memiliki ketergantungan yang berbeda pada prediksi yang tumpang tindih *(h,w),* peta koefisien dapat digunakan alih-alih berbagi koefisien di semua posisi. Peta koefisien G_j dihasilkan melalui lapisan konvolusi pada peta masukan x. *(Hai)*

Persamaan (6) menentukan bahwa $G_{h,w,j}$ harus selalu menjumlahkan hingga satu untuk satu posisi *(h,w).* Oleh karena itu, kami menerapkan fungsi softmax pada peta tersebut (disebut peta bobot dalam makalah ini) untuk memenuhi batasan ini. Normalisasi softmax digunakan karena memiliki stabilitas numerik yang baik dan mudah untuk dibedakan, yaitu G_j dinormalisasi menggunakan berikut ini:

$$G_j(p_0) = \exp(G_j(p_0)) / \sum_{m=1}^{k^2} \exp(G_m(p_0)), \; j = 1, \ldots, k^2. \tag{7}$$

Oleh karena itu, dalam skema konvolusi padat yang diusulkan, penjumlahan bobot diadopsi sebagai strategi kombinasi untuk peta fitur yang tumpang tindih. Ini adalah metode yang sederhana namun efisien. Akhirnya, peta keluaran ***z*** dapat diperoleh melalui a penjumlahan tertimbang sebagai berikut:

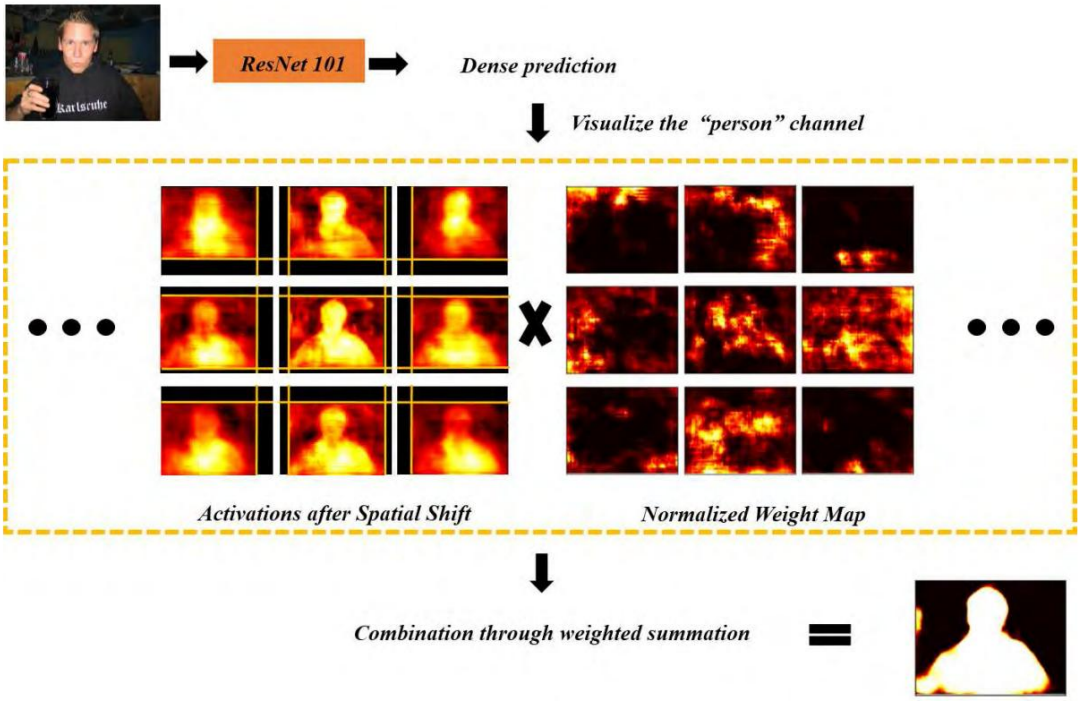
$$z(p_0) = \sum_{j=1}^{k^2} G_j(p_0) \cdot \sum_{saya=1}^{k^2} \tilde{y}_j(p_i) \cdot x(p_0 + p_i + p_j). \tag{8}$$

Jika laju dilatasi bukan 1, cukup ganti kisi sampel ***R*** pada (4) dengan kisi yang sesuai dan prosesnya masih dapat diselesaikan. Ketika stride *s* >1, konvolusi dapat diubah menjadi dilatasi konvolusi dengan stride 1, yang mempertahankan resolusi fitur; Dengan demikian, makalah ini meninggalkan situasi seperti itu.

Di bawah pengaturan yang paling lengkap, koefisien yang dinormalisasi (peta bobot) tidak dibagi oleh posisi yang berbeda atau oleh saluran keluaran yang berbeda. Dalam praktiknya, kompromi-kompromi tertentu dapat dilakukan dengan mempertimbangkan kompleksitas. Gambar 3 mengilustrasikan bagaimana konvolusi padat bekerja pada satu gambar input dan memvisualisasikan peta fitur yang sesuai. Setelah proses ekstraksi fitur sebelumnya, konvolusi padat terlebih dahulu menghasilkan beberapa prediksi tengah. Kemudian, unit pergeseran spasial membantu mengoreksi koordinat relatifnya dan memindahkannya dengan benar. Sementara itu, peta bobot sensitif posisi yang dinormalisasi dihasilkan melalui lapisan konv standar lainnya dan membantu menggabungkan beberapa prediksi menengah tersebut. Saluran "benar" mencapai aktivasi terkuat dan dengan demikian menekan saluran lain setelah operasi softmax. Terakhir, peta segmentasi disintesis menggunakan fungsi argmax.

C. PELATIHAN DENSE CONVOLUTIONAL NETWORK

Untuk mengintegrasikan konvolusi padat ke dalam kerangka saat ini, unit konvolusi padat (DCU) dikembangkan. Struktur keseluruhan ditunjukkan pada Gambar. 2. Ini mencakup dua konvolusi 2D standar, satu normalisasi grup demi grup, satu unit pergeseran spasial dan satu perkalian berdasarkan elemen diikuti dengan penjumlahan.



GAMBAR 3. Visualisasi prediksi tengah dan peta bobot yang sesuai di DCU. Bagian yang tumpang tindih digabungkan melalui perkalian dan penjumlahan berdasarkan elemen (ukuran kernel adalah 3 × 3). Terbaik dilihat dalam warna.

Berdasarkan kerangka pembelajaran mendalam yang ada, hanya diperlukan sedikit modifikasi. Titik kunci terletak pada operasi pergeseran spasial. Bagian ini implementsyj berikut (4) dan (5). Misalkan peta fitur input memiliki saluran *Cin* , peta fitur keluaran memiliki saluran *Cout* dan ukuran kernel konvolusi adalah *k* × k. Untuk mencapai konvolusi padat, pertama-tama DCU menghasilkan

k 2*Cout* peta tengah *yj* menggunakan konvolusi standar lalu tambahkan offset ke dalamnya melalui lapisan pergeseran spasial dan dapatkan *yj* . Perhatikan bahwa di Bagian III-A, konvolusi padat dijelaskan pada satu lapisan masukan. Formulasi multichannel dapat dengan mudah disimpulkan sebagai berikut:

$$z^{(Co)}(\boldsymbol{p0}) = \sum_{j=1}^{2k} G_j^{(Bersama)}(\boldsymbol{p0}) \cdot \left(\sum_{saya=1}^{Cip} \ddot{y}j^{(Ci,Co)}(\boldsymbol{pi}) \cdot x^{(Ci)}(\boldsymbol{p0} + pi + pj) + b(Co) \right) \quad (9)$$

di mana *Co* dan *Ci* menunjukkan saluran output dan input yang berbeda. Sekarang output memiliki saluran *Cout* , ***y*** andy dibagi menjadi grup *Cout* dengan masing-masing grup memiliki *k* peta, sehingga berbeda dari konvolusi dapat dilakukan operasi pergeseran spasial kelompok demi kelompok. Konvolusi tidak menggunakan bantalan karena pergeseran spasial akan memulihkan dimensi yang dikurangi. Artinya ***y*** memiliki dimensi yang sama dengan input x, namun ***y*** lebih kecil. Demikian pula, peta bobot ternormalisasi *k* 2*Cout* ***G*** dapat diperoleh dengan konvolusi standar lain yang diikuti oleh satu normalisasi grup-demi-grup menggunakan (7). Sebaliknya, padding digunakan kali ini.

Dalam proses perambatan maju, peta keluaran dapat dihitung sebagai berikut (9). Untuk perambatan mundur, rumus diferensiasi harus ditetapkan. Untuk tetap konsisten dengan Bagian III-A, pembahasan berikut ini berada dalam satu kelompok dan Pengadilan superskrip *dihilangkan* . Makalah ini hanya menyajikan turunan untuk operasi pergeseran spasial.

Sisanya dapat diselesaikan dengan propagasi balik dari konvolusi standar dan lapisan softmax.

Diberikan *ÿLoss/ÿz* untuk peta keluaran z, gradien wrt *yj* dan *Gj* dapat dihitung sebagai

$$\frac{\ddot{y}Kerugian}{\ddot{y}yj(\boldsymbol{p0})} = \frac{\ddot{y}Kerugian}{\ddot{y}z(p0 \ddot{y} pj)} \cdot Gj(p0 \ddot{y} pj) \quad (10)$$

dan

$$\frac{\ddot{y}Kerugian}{\ddot{y}Gj(p0)} = \frac{\ddot{y}Kerugian}{\ddot{y}z(p0)} \cdot yj(\boldsymbol{p0} + pj). \quad (11)$$

Di sini, korespondensi *yj* ***(p0) =yj (p0*** *ÿ pj)* digunakan. Dalam skema di atas, setiap saluran keluaran memiliki peta bobotnya sendiri karena diharapkan mengkodekan semantik yang berbeda.

Dengan demikian, propagasi balik untuk *yj* dan *Gj* dalam grup yang berbeda adalah independen satu sama lain. Namun, selama eksperimen, kami menemukan bahwa jika peta bobot ***G*** dibagikan di seluruh saluran keluaran, jaringan masih berfungsi dengan baik dan hanya sedikit kehilangan akurasi yang teramati. Dalam hal ini, gradien ÿLoss/ÿGj(p0) harus merangkum semua saluran *Cout* dan rumusnya menjadi:

$$\frac{\ddot{y}Kerugian}{\ddot{y}Gj(p0)} = \sum_{(Co)} \frac{\ddot{y}Kerugian}{\ddot{y}z^{(Co)}(\boldsymbol{p0})} \cdot y_i^{(Co)}(\boldsymbol{p0} + pj). \quad (12)$$

Algoritma 1 Pelatihan End-to-End Dense Convolutional Jaringan

Inisialisasi: Siapkan lapisan konvolusi dan atur ukuran kernel DCU ke *k* × k. Parameter jaringan dilambangkan sebagai Jumlah iterasi maksimum: · *max_iter.*

1: **while** *iter* < *max_iter* **do** 2: **Forward:**

Co=1...Cout *Co Co* 3: Generate ***{y G }*** layer sebelumnya *j , j*

4: **untuk** *Co* = 1 ke *Cout* **do**

5: Tambahkan pergeseran spasial ke *yj* menggunakan korespondensi $y_j^{Co}(p0) = y_j^{bersama}(p0 + pj)$

6: Normalisasikan G_j^{Co} *Co* menggunakan (7) $\ddot{y}_j$.

7: Keluaran ^{bersama} menggunakan (9).

z 8: **akhir untuk**

9: Hitung kerugian softmax melalui (2) lalu dapatkan *Co ÿLoss/ÿz* ·

10 : **Mundur :** 11 :

untuk *Co* = 1 ke *Cout* **do**

12: Terapkan (10) ÿ $\frac{\ddot{y}Kerugian}{\ddot{y}G_j^{Co}}$.

13: Terapkan (11) ÿ $\frac{\ddot{y}Loss}{\ddot{y}y_i^{Co}}$.

14: **end for** 15:

Back propagation ke layer sebelumnya.

16: Perbarui .

17: *iter* ÿ *iter* + 1 18:

end while Keluaran:

Jaringan terlatih untuk inferensi.

Prosedur keseluruhan untuk melatih jaringan konvolusional padat dirangkum dalam Algoritma 1.

IV. EKSPERIMEN

Kami melakukan eksperimen ekstensif pada dataset benchmark PASCAL VOC 2012 [48] dan Cityscapes [49] untuk mendemonstrasikan keefektifan skema konvolusi padat yang diusulkan.

Untuk memulainya, model dasar yang sering digunakan (dengan lapisan konvolusional tunggal sebagai pengklasifikasi) dibuat mengikuti praktik umum. Kemudian, beberapa pengaturan penting untuk akurasi akhir diberikan dan didiskusikan. Setelah itu, konvolusi padat diterapkan pada model dasar untuk menunjukkan kinerjanya yang unggul dibandingkan dengan konvolusi standar. Namun, karena metode yang diusulkan memperkenalkan lebih banyak parameter (dua lapisan konvolusional menengah) ke dalam jaringan saat ini, sulit untuk mengetahui apakah peningkatan tersebut berasal dari struktur yang dirancang khusus atau parameter tambahan. Untuk mengamati lebih lanjut kapasitasnya di bawah parameter yang sama atau bahkan lebih sedikit daripada konvolusi standar, konvolusi padat diintegrasikan ke dalam metode PSPNet yang terkenal [11]. Parameter dalam konvolusi standar dan konvolusi padat dikontrol dengan hati-hati, dan hasil yang sesuai ditampilkan. Selain itu, waktu berjalan dan kompleksitas juga dilaporkan untuk perbandingan yang komprehensif. Kami menyediakan kode di https://github.com/hancy16/DCU.

A. PENGATURAN EKSPERIMENTAL

Metrik Evaluasi: Untuk mengevaluasi pendekatan kami, standard mean junction over union (mIoU) digunakan. Metrik ini lebih memilih kelas dengan area yang luas. Namun, dalam dataset seperti PASCAL VOC 2012, jumlah piksel bervariasi secara signifikan untuk kelas yang berbeda. Luas latar belakang bisa beberapa lusin kali lebih besar dari kelas lain (sebagai hasilnya, latar belakang selalu memiliki akurasi tertinggi). Oleh karena itu, kami juga memberikan skor IoU berdasarkan kelas untuk perbandingan yang lebih baik. Menurut [14], metrik evaluasi dihitung sebagai:

- Keakuratan piksel: rasio piksel yang diklasifikasikan dengan benar terhadap total jumlah piksel, didefinisikan sebagai $\frac{\sum_i t_{ii}}{\sum_i T_i}$.
- Rata-rata IoU: persimpangan rata-rata atas persentase gabungan atas semua kelas, didefinisikan sebagai $\frac{1}{N}\sum_i \frac{t_{ii}}{T_i + \sum_j t_{ji} - t_{ii}}$.

di mana *tji* menunjukkan jumlah piksel milik kelas *j* sedangkan diprediksi menjadi kelas i, *Ti* menunjukkan jumlah total piksel di kelas i, *N* adalah jumlah kelas.

Model Baseline: Kerangka kerja Deeplab [10] diadopsi sebagai model baseline. Menurut [14], pertama-tama kami memodifikasi ResNet-101 [2] menjadi mode konvolusi penuh untuk mengaktifkan prediksi padat. Ini dilakukan dengan mengganti beberapa lapisan terakhir di belakang *conv5_x* dengan lapisan konvolusi (klasifikasi).

Dengan cara ini, bagian ResNet berfungsi sebagai modul ekstraksi fitur dan model prapelatihan menyediakan inisialisasi yang tepat. Tingkat pembelajaran relatif untuk lapisan sebelumnya adalah 1 dan untuk lapisan klasifikasi adalah 10. Kemudian, dua lapisan penyatuan terakhir di ResNet asli dibuang untuk mempertahankan resolusi fitur dan pelebaran konvolusi digunakan untuk menjaga agar bidang reseptif tidak berubah. Tingkat pelebaran di *conv4_x* diatur ke 2 dan di *conv5_x* diatur ke 4, sehingga seluruh jaringan memiliki tingkat downsample 8. Untuk jaringan kami, lapisan conv di belakang *conv5_x* diganti dengan unit konvolusi padat. Dalam pengaturan kami, lapisan konvolusi di DCU tidak menggunakan normalisasi batch. Selama pelatihan dan pengujian, lapisan normalisasi batch di ResNet menggunakan parameter yang disimpan, dan tingkat pembelajarannya nol. Agar sesuai dengan ukuran peta keluaran, peta groundtruth diturunkan sampelnya sebanyak 8 sebelum dimasukkan ke dalam lapisan kerugian. Pada tahap inferensi, peta keluaran di-upampling sebanyak 8 menggunakan interpolasi bilinear. Lapisan kerugian menggunakan fungsi softmax dan fungsi kerugian logistik multinomial (lihat (2)) untuk menghitung kerugian serta gradien piksel untuk propagasi balik dan diterapkan setelah lapisan klasifikasi.

B. HASIL PADA PASCAL VOC 2012

PASCAL VOC 2012 [48] adalah dataset segmentasi semantik yang banyak digunakan. Mengikuti praktik umum [10], [11], anotasi tambahan yang disediakan oleh [50] disertakan dalam pelatihan.

Terakhir, kami memiliki 13.487 gambar yang dianotasi dengan baik, dimana 10.582 gambar digunakan untuk pelatihan, 1.449 untuk validasi, dan 1.456 untuk pengujian.

Dataset PASCAL VOC2012 memiliki 20 kelas objek dan satu kelas latar belakang. Selain itu, dataset berisi anotasi lain yang berarti area tersebut sulit untuk dipisahkan

ment, biasanya batas objek. Saat ini area seperti itu diabaikan dalam pelatihan dan pengujian.

Protokol Pelatihan: Saat melatih jaringan prediksi padat, kami secara acak mencerminkan dan memotong gambar pelatihan untuk augmentasi data. Ukuran pemotongan diatur ke 321 × 321. Seperti yang disarankan dalam [10], [11], kebijakan tingkat pembelajaran "poli" diadopsi, daya diatur ke 0,9, tingkat pembelajaran dasar diatur ke 0,00025, momentum dan peluruhan bobot masing-masing ditetapkan ke 0,9 dan 0,0005 . Implementasi kami didasarkan pada platform Caffe [51] dengan modifikasi dari [10]. Ukuran batch diatur ke 30. Model baseline memiliki konsumsi memori yang sangat besar, sehingga ukuran batch yang besar dapat dicapai dengan mengatur ukuran batch ke 3 dan iter_size ke 10 di Caffe saat melatih GPU tunggal. Pelatihan berjalan selama 30K iterasi pada dataset pelatihan PASCAL VOC2012 yang ditambah. Lapisan baru yang ditambahkan diinisialisasi secara acak menggunakan cara "msra" [52].

1) STUDI ABLASI PADA ARSITEKTUR JARINGAN

Pertama, kami ingin melaporkan rincian lebih lanjut tentang hasil yang disajikan dalam makalah konferensi kami dan memberikan beberapa diskusi.

Ada faktor penting dalam model baseline yang mempengaruhi akurasi secara signifikan, yaitu tingkat dilatasi *r* pada lapisan klasifikasi. Meskipun lapisan yang ditumpuk di ResNet 101 membawa bidang reseptif yang cukup besar untuk gambar PASCAL VOC2012, tingkat pelebaran yang besar masih diperlukan di lapisan klasifikasi. Pengamatan serupa juga dapat ditemukan di [10]. Fenomena ini sebagian dijelaskan dalam [53], yaitu bidang reseptif efektif tidak sebesar yang diharapkan. Oleh karena itu, kami melatih beberapa model dasar dengan *r* berbeda mengikuti protokol yang diperkenalkan di atas. Ketika *r* = 1, IoU rata-rata adalah 69,2% pada set validasi; saat *r* menjadi lebih besar, rata-rata IoU meningkat. Kami akhirnya memilih *r* = 12 sebagai model referensi, yang mencapai 70,1% mIoU. Dalam percobaan berikut, kami menjaga bidang reseptif dari lapisan klasifikasi tidak berubah untuk menghilangkan efek yang dibawa oleh bidang reseptif yang berbeda.

a: HASIL STRATEGI PETA BERAT YANG BERBEDA

Pada model dasar, klasifikasi dilengkapi dengan lapisan konvolusi 2D standar dengan ukuran kernel *k* = 3 dan laju pelebaran *r* = 12. Jaringan kami mengganti lapisan konvolusi standar dengan unit konvolusi padat dengan parameter hiper yang sama. Untuk menyelidiki strategi mana yang paling baik menggabungkan aktivasi tengah, tiga metode dicoba untuk menghasilkan peta bobot G:

- M1: Semua aktivasi (baik posisi maupun saluran) di peta keluaran memiliki bobot yang sama, yang berarti hanya diperlukan 9 bobot. Dalam hal ini, bobot tidak dihasilkan oleh lapisan sebelumnya dan karena itu tidak terlibat dalam perambatan balik selanjutnya. Gradien dihitung menggunakan rumus yang mirip dengan (12) dan menjumlahkan keseluruhan aktivasi keluaran. Akhirnya jaringan mencapai 70,5% mIoU pada set validasi.
- M2: Aktivasi yang berbeda dalam kelompok yang sama diberikan bobot yang berbeda; dengan demikian, peta bobot dibuat. Saluran keluaran yang berbeda berbagi peta bobot. Di bawah pengaturan ini, 9 peta bobot diproduksi dan diharapkan mengkodekan semantik di posisi yang berbeda. mIoU yang sesuai meningkat menjadi 71,4%.
- M3: Berdasarkan pengaturan sebelumnya, peta bobot tidak lagi dibagikan di seluruh saluran keluaran. Dibandingkan dengan M2, peningkatan mIoU kurang dari 0,05% diamati.

Kami akhirnya memilih metode terakhir; sebenarnya, yang kedua sama baiknya. Pada percobaan tersebut, laju dilatasi lapisan konvolusi yang menghasilkan ***G*** dan ***y*** adalah sama, dilambangkan dengan ***rG*** dan ***ry***. Jika ***rG*** diubah menjadi 1, diamati kehilangan 0,01% pada mIoU. Ini menunjukkan bahwa bidang reseptif yang besar tidak diperlukan saat membuat peta bobot. Oleh karena itu, di bagian selanjutnya, ***rG*** selalu sama dengan ***ry*** untuk penyederhanaan dan tidak akan disebutkan lagi. Perbandingan strategi-strategi tersebut diberikan dalam Tabel. 1.

TABEL 1. mIoU strategi peta bobot yang berbeda.

	Dilation r_G	Dilation r_y	Param_Size	mIoU
baseline(r=12)	—	12	—	70.1
DCU(k=3)_M1	—	12	9	70.5
DCU(k=3)_M2	12	12	166K	71.4
DCU(k=3)_M3	12	12	3.5M	71.4(+0.05)
DCU(k=3)_M3'	1	12	3.5M	71.4(-0.01)

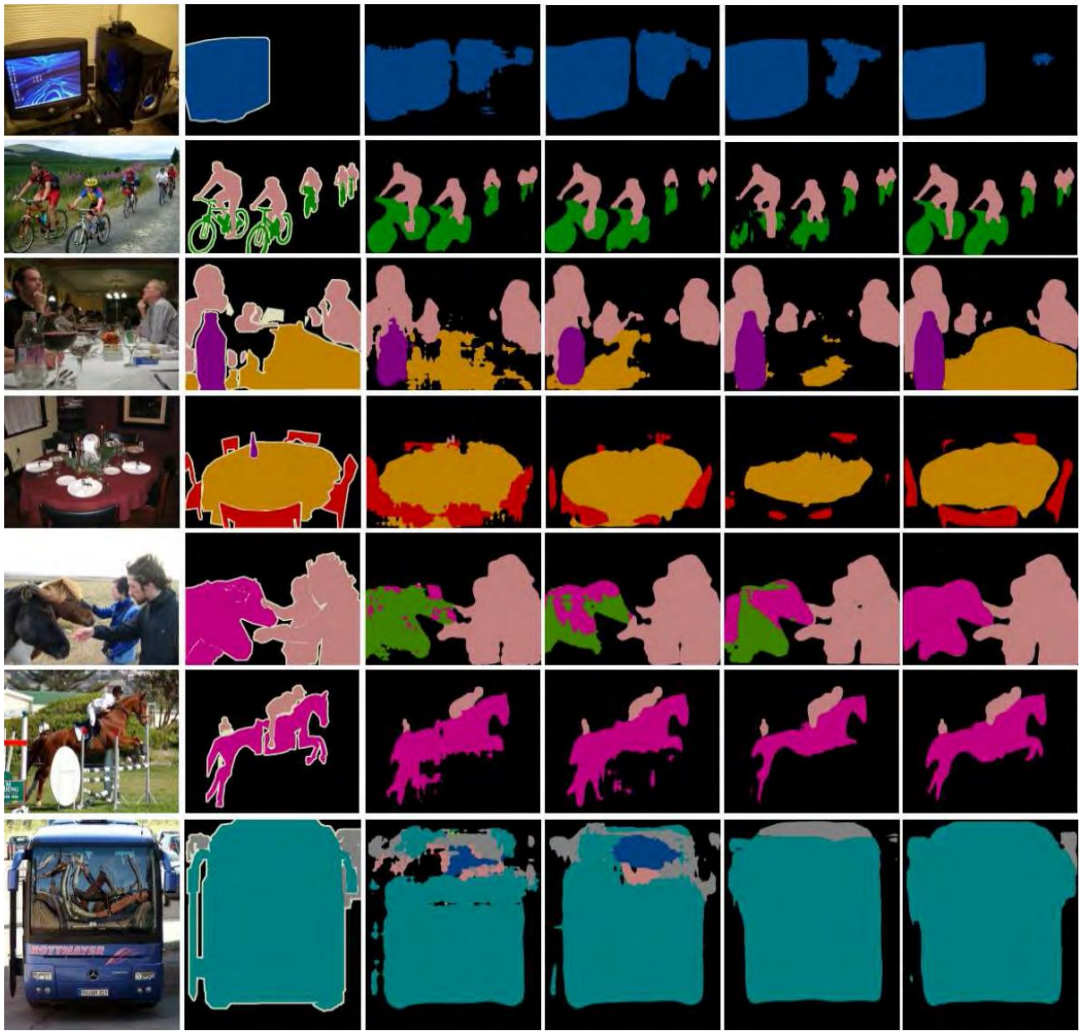
b: PERBANDINGAN METODE YANG BERBEDA

Dalam DCU, ukuran parameter meningkat pesat dengan k. Oleh karena itu, hanya satu percobaan tambahan mengenai ukuran kernel yang dilakukan dengan *k* diatur ke 5. Karena bidang reseptif sangat penting untuk akurasi akhir, tingkat pelebaran yang sesuai disesuaikan dengan 6. Dengan cara ini, bidang reseptif tetap konstan. Hasil IoU berdasarkan kelas dari pengaturan ini bersama dengan eksperimen sebelumnya tercantum dalam Tabel. 2. Untuk *k* = 5, mIoU mencapai 71,7%. Di bawah bidang reseptif yang sama, ukuran kernel yang lebih besar dapat menghasilkan hasil segmentasi yang lebih baik. Untuk ukuran kernel yang besar, metode dekomposisi kernel yang diusulkan pada [53] dapat diadopsi untuk menyeimbangkan akurasi dan kom Perbandingan dengan beberapa metode terkait lainnya ditunjukkan di bagian atas Tabel. 3. Secara keseluruhan, DCU memiliki keunggulan dibanding kompetitornya yang masih dalam skema konvolusi standar.

Untuk menunjukkan efek konvolusi padat dengan jelas dan menghindari pengaruh faktor lain, banyak trik peningkatan akurasi yang umum tidak digunakan di sini, seperti menjalankan penskalaan ulang gambar pelatihan secara domly, merata-ratakan hasil di seluruh skala input serval, melakukan prapelatihan jaringan pada skala yang lebih besar dataset kemudian finetuning pada set pelatihan PASCAL VOC2012 standar, menggunakan CRF sebagai postprocess dan seterusnya. Tidak ada normalisasi batch yang diadopsi di DCU untuk perbandingan yang adil dengan model baseline.

TABEL 2. IoU berdasarkan kelas dari pengaturan yang berbeda.

	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
baseline(r=12)	92.3	85.5	38.6	82.7	64.1	77.5	89.3	82.0	87.2	33.9	74.5	42.7	79.4	71.9	76.6	80.8	53.0	78.0	35.1	81.0	66.6	70.1
DCU(k=3)_M1	92.5	86.4	38.7	82.6	62.5	77.7	89.3	82.6	87.2	33.8	71.4	44.3	80.0	73.0	77.0	82.3	54.5	78.5	38.7	78.6	68.4	70.5
DCU(k=3)_M2	92.8	88.0	40.0	85.0	65.1	78.2	90.3	83.5	88.2	35.7	70.9	42.2	80.8	72.0	78.9	82.9	56.4	79.0	40.1	79.9	70.0	71.4
DCU(k=3)_M3	92.8	86.8	39.7	85.1	66.1	79.1	90.3	82.9	87.9	35.5	72.1	42.8	81.0	72.1	78.8	82.6	56.3	78.8	39.2	80.3	69.7	71.4
DCU(k=5)	92.8	87.0	39.8	84.6	65.0	78.8	90.7	83.6	88.0	36.7	71.8	44.4	80.5	71.5	80.0	83.3	55.4	79.8	39.1	80.1	71.3	71.7



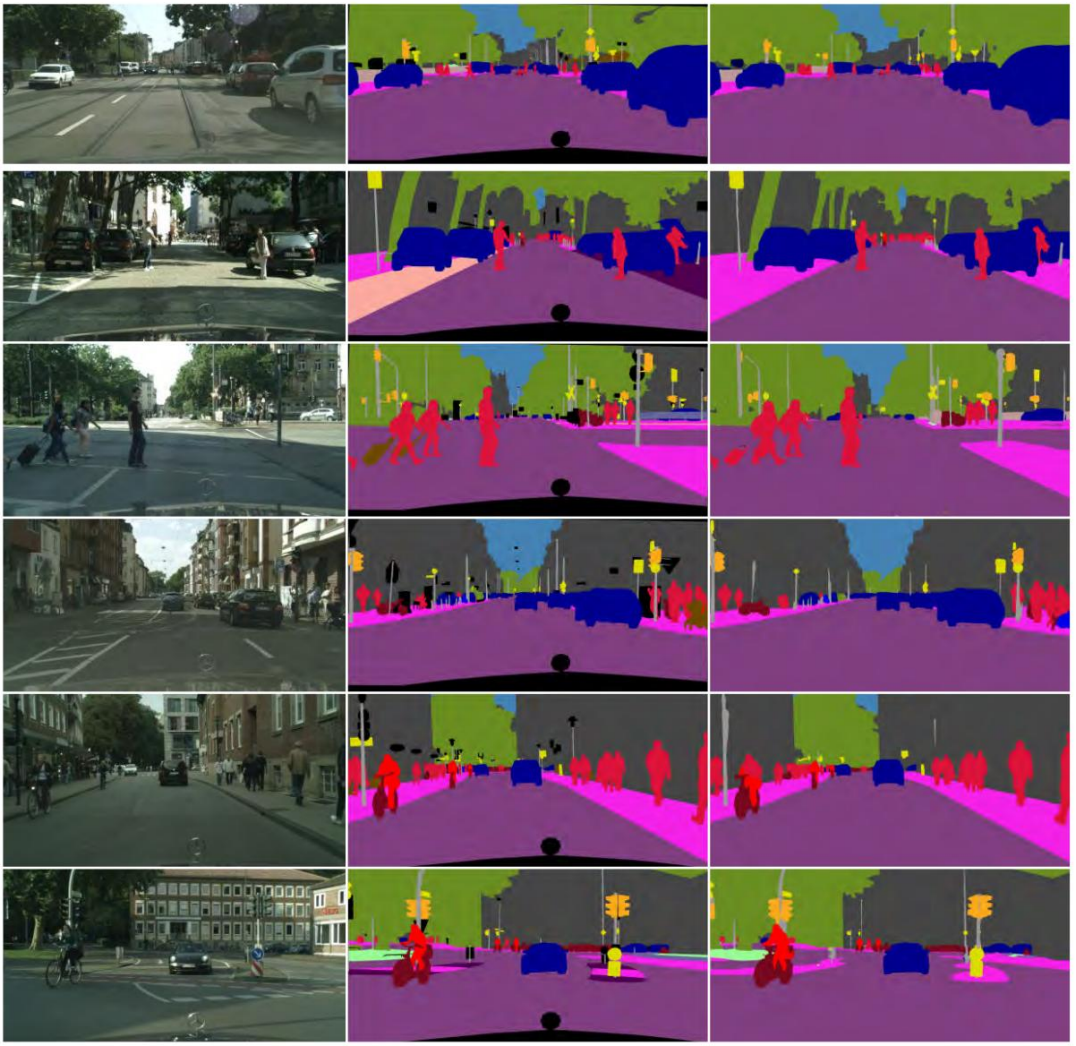
GAMBAR 4. Perbandingan visual dari berbagai metode. Dari kiri ke kanan: Gambar Asli, Kebenaran Dasar, Res101-Conv, Res101-DCU, Res101-SPP-Conv, Res101-SPP-DCU. Conv menunjukkan konvolusi standar, DCU menunjukkan unit konvolusi padat. Terbaik dilihat dalam warna.

2) EKSPLORASI LEBIH LANJUT PADA DCU

Dalam eksperimen di atas, unit konvolusi rapat dibandingkan dengan model baseline dan beberapa metode terkait. Namun, jaringan tersebut tidak memiliki parameter yang sebanding. Hal ini membuat hasilnya kurang meyakinkan karena diketahui bahwa untuk jaringan yang dalam, peningkatan dapat dicapai melalui penumpukan lapisan konvolusi atau dekonvolusi. Sementara itu, kompatibilitas dan generalisasi DCU perlu diuji karena diusulkan sebagai pengganti konvolusi standar. Berdasarkan pertimbangan tersebut, kami menambahkan modul penggabungan piramida spasial (SPP) yang digunakan dalam [11] ke model baseline dan menerapkan DCU setelahnya. Bagian ini mencoba untuk menunjukkan bahwa 1) DCU menunjukkan kinerja yang unggul dibandingkan dengan lapisan susun langsung dan 2) dapat dengan mudah diintegrasikan ke dalam kerangka kerja lain dan tetap efektif

a: RINCIAN STRUKTUR JARINGAN

Perhatikan bahwa dalam PSPNet asli [11], klasifikasi dilakukan oleh dua lapisan konvolusi standar. Lapisan pertama memiliki 512 saluran keluaran dengan ukuran kernel $k = 3$.



GAMBAR 5. Contoh visual dari hasil kami pada dataset validasi Cityscapes. Dari kiri ke kanan: Gambar Asli, Kebenaran Dasar, Res101-SPP-DCU. Model dilatih di set pelatihan. Terbaik dilihat dalam warna.

TABEL 3. Rata-rata IoU dari metode yang berbeda pada set validasi PASCAL VOC2012.

	mIoU
baseline(r=1)	69.2
baseline(r=12): LargeFOV	70.1
DeepLab-v2 MSC [10]	71.2
GCN(k=3) [53]	70.1
GCN(k=5) [53]	71.1
DCU(k=3)	71.4
DCU(k=5)	71.7
Deformable Conv [28]	75.3
Res101-SPP-DCU	76.3

Lapisan kedua memiliki 21 saluran keluaran dengan ukuran kernel $k = 1$. Pengaturan seperti itu memberi kita kesempatan untuk melakukan eksperimen terkontrol karena mereka memiliki parameter yang sebanding dengan jaringan yang diusulkan. Untuk DCU, lapisan konvolusi yang digunakan untuk menghasilkan $\boldsymbol{G}$ dan $\boldsymbol{y}$ memiliki total 378 saluran, dan bagian sisanya tidak memiliki parameter yang dapat dilatih. Agar tetap konsisten dengan [11], satu lapisan konvolusi 1 × 1 ditempatkan setelah DCU, ukuran krop diatur ulang ke 473 dan ukuran peta fitur keluaran adalah 60 × 60 setelah downsampling 8 kali. Modul pooling piramida spasial terdiri dari empat layer pooling dengan ukuran kernel dan stride setara dengan 60, 30, 20, dan 10. Peta output SPP kemudian digabungkan dengan inputnya. Semua lapisan konvolusi setelah modul SPP memiliki tingkat dilatasi $r = 1$. Dalam percobaan, lapisan di belakang modul SPP dilatih dari awal dengan inisialisasi acak msra [52], dan bagian sebelumnya diinisialisasi dengan pretrained bobot di [11] karena modelnya sedikit berbeda dari Res101 asli. Di bawah pengaturan seperti itu, hanya ukuran batch 1 yang tersedia karena keterbatasan memori fisik dari satu GPU. Model dilatih pada 4 GPU TITAN XP. Dibatasi oleh jumlah GPU, parameter normalisasi batch dibekukan untuk model dasar dan tidak disertakan dalam lapisan yang baru ditambahkan setelah SPP.

b: PENINGKATAN YANG KONSISTEN

Kedua jaringan secara singkat dilambangkan sebagai Res101-SPP-Conv, Res101-SPP-DCU dan dilatih tentang set pelatihan tambahan untuk

TABEL 4. Perbandingan ukuran parameter, waktu maju dan mIoU untuk metode yang berbeda.

	Param_Size	Forward Time(sec)	mIoU(%)
Res101-Conv	0.4M	0.071	69.2
Res101-DCU	6.6M	0.082	71.4
Res101-SPP-Conv	13.0M	0.101	71.4
Res101-SPP-DCU	10.6M	0.110	76.3

Iterasi 30K menggunakan protokol yang dijelaskan di atas. Demikian pula, model dasar sebelumnya dengan pelebaran $r = 1$ dan jaringan berbasis konvolusi padat pasangannya dengan ukuran kernel $k = 3$ dan pelebaran $r = 12$ disebut Res101-Conv, Res101-DCU, secara bersamaan. Perbandingan keempat metode tersebut disajikan pada Tabel. 4 dalam hal ukuran parameter (blok Res101 tidak termasuk), waktu maju seluruh jaringan dan rata-rata IoU. Kesimpulan utamanya adalah sebagai berikut: 1) Mulai dari model baseline, mengganti lapisan konvolusi tunggal dengan modul SPP diikuti oleh dua lapisan konvolusi menghasilkan peningkatan mIoU sebesar 2,2%.

Hasilnya sangat dekat dengan yang dicapai oleh DCU dengan FOV (dilation rate) yang besar. Namun, DCU hanya memiliki setengah parameter, yang menunjukkan kinerja superiornya dibandingkan modul penyatuan piramida spasial.

2) Menerapkan DCU pada SPP sangat meningkatkan akurasi, dengan peningkatan 4,9% mIoU. Ini tidak mengherankan karena konvolusi padat dirancang untuk meningkatkan koneksi dan interaksi antara teks konteks. Dengan demikian, DCU memiliki angin di punggungnya ketika dilengkapi dengan modul penyematan konteks yang kuat, yaitu SPP.

3) Dalam hal tugas prediksi yang padat, DCU lebih cocok daripada konvolusi standar. Itu mencapai banyak hal akurasi yang lebih tinggi bahkan dengan parameter yang lebih sedikit.

4) Komputasi dan kedalaman yang dibawa oleh DCU menghasilkan sekitar 14% waktu berjalan ekstra berdasarkan implementasi CPU kami yang tidak dioptimalkan. Ini adalah setengah dari yang dibutuhkan oleh SPP. Faktanya, DCU tidak memuat banyak perhitungan dan dapat dioptimalkan dengan cara perkalian matriks dan dilakukan di GPU, yang selanjutnya akan mengurangi kebutuhan waktu.

Untuk pelatihan dan pengujian multiskala, gambar diubah ukurannya menggunakan faktor 0.5, 0.75, 1.25, 1.5, dan 1.75. Selama inferensi, peta probabilitas dirata-ratakan pada berbagai skala. Gambar terbalik juga disertakan. Dalam kondisi ini, model mencapai 76,8% mIoU. Peningkatan tersebut tetap konsisten dengan karya lainnya. Secara tidak terduga, menyetel laju dilatasi DCU pada Res101-SPP-DCU ke $r = 12$ mengakibatkan penurunan akurasi. Hasilnya tercantum dalam Tabel. 5.

Penurunan mIoU menunjukkan bahwa modul pooling piramida spasial menyediakan bidang reseptif yang cukup besar untuk mengikuti klasifikasi; dengan demikian, FoV besar di DCU membawa konteks global terbatas yang tidak dapat mengimbangi kerugian secara detail.

TABEL 5. Hasil percobaan dengan pengujian MS_Mirror dan dilatasi yang lebih besar kecepatan.

MS_Mirror	Dilation Rate	mIoU(%)	Pixel Accuracy(%)
	1	76.3	94.26
√	1	76.8	94.48
√	12	75.9	93.44

TABEL 6. Hasil eksperimen pada perangkat uji pemandangan kota. hanya data halus yang digunakan selama pelatihan. † set pelatihan dan validasi.

Method	IoU Cla(%)	iIoU Cla(%)	IoU (%)	iIoU Cat(%)
CRF-RNN [26]	62.5	34.4	82.7	66.0
FCN [14]	65.3	41.7	85.7	70.1
SiCNN+CRF [54]	66.3	44.9	85.0	71.2
DPN [55]	66.8	39.1	86.0	69.1
Dilation10 [34]	67.1	42.0	86.5	71.1
LRR [56]	69.7	48.0	88.2	74.7
DeepLab [10]	70.4	42.6	86.4	67.7
Piecewise [57]	71.6	51.7	87.3	74.1
RefineNet [12]	73.6	47.2	87.9	70.6
FoveaNet [58]	74.1	52.4	89.3	77.6
PEARL [59]	75.4	51.6	89.2	75.1
TuSimple [32]	77.6	53.6	90.1	75.2
SAC-multiple [60]	78.1	55.2	90.6	78.3
PSPNet [11]	78.4	56.7	90.6	78.6
ResNet-38 [61]	78.4	59.1	90.9	81.1
Res101-SPP-Conv	77.0	54.4	89.9	77.1
Res101-SPP-DCU	77.8	56.1	90.3	78.3
Res101-SPP-DCU‡	78.9	57.8	90.5	78.0

c: PERBANDINGAN VISUAL

Pada Gambar 4 disajikan beberapa hasil segmentasi visual untuk keempat metode tersebut. Berdasarkan visualisasi, unit konvolusi yang rapat menghasilkan diskontinuitas yang jauh lebih sedikit dan batas yang lebih halus pada peta segmentasi. Dalam konvolusi standar, beberapa posisi tidak dapat dihindari untuk menghadapi konteks yang membingungkan dalam bidang reseptif yang sesuai. Maka noise pasti akan muncul, terutama pada adegan yang rumit. Namun, diskontinuitas tersebut dapat dikoreksi melalui prediksi yang tumpang tindih di DCU, karena prediksi yang tepat kemungkinan besar dapat ditemukan dari tetangganya.

C. HASIL PADA CITYSCAPES

Kami juga melakukan percobaan pada dataset Cityscapes [49] dengan model Res101-SPP-DCU. Cityscapes adalah kumpulan data pemandangan jalanan yang berisi 2975, 500, dan 1525 gambar yang dianotasikan dengan halus untuk pelatihan, validasi, dan pengujian. Semua gambar berukuran 2048×1024 *px.* Kami menggunakan pengaturan yang sama seperti di atas kecuali bahwa proses pelatihan memerlukan iterasi 90K pada kumpulan data ini. Perbedaan utama dalam struktur model adalah bahwa DCU berisi 504 saluran ***(total y*** dan ***G),*** yang mencapai parameter yang serupa dengan pasangan dasarnya.

Hasil pada set pengujian Cityscapes dilaporkan dalam Tabel. 6 dan Tabel. 7. Fusing dan mirroring multiskala digunakan untuk pengujian. Beberapa contoh visual set validasi Cityscapes disajikan pada Gambar. 5. Karena kurangnya lapisan normalisasi batch dan kerugian tambahan serta

TABEL 7. Set pengujian IoU berdasarkan kelas pada pemandangan kota. ‡ berarti bahwa model dilatih pada set pelatihan dan validasi.

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
CRF-RNN [26]	96.3	73.9	88.2	47.6	41.3	35.2	49.5	59.7	90.6	66.1	93.5	70.4	34.7	90.1	39.2	57.5	55.4	43.9	54.6	62.5
FCN [14]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
SiCNN+CRF [54]	96.3	76.8	88.8	40.0	45.4	50.1	63.3	69.6	90.6	67.1	92.2	77.6	55.9	90.1	39.2	51.3	44.4	54.4	66.1	66.3
DPN [55]	97.5	78.5	89.5	40.4	45.9	51.1	56.8	65.3	91.5	69.4	94.5	77.5	54.2	92.5	44.5	53.4	49.9	52.1	64.8	66.8
Dilation10 [34]	97.6	79.2	89.9	37.3	47.6	53.2	58.6	65.2	91.8	69.4	93.7	78.9	55.0	93.3	45.5	53.4	47.7	52.2	66.0	67.1
LRR [56]	97.7	79.9	90.7	44.4	48.6	58.6	68.2	72.0	92.5	69.3	94.7	81.6	60.0	94.0	43.6	56.8	47.2	54.8	69.7	69.7
DeepLab [10]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
Piecewise [57]	98.0	82.6	90.6	44.0	50.7	51.1	65.0	71.7	92.0	72.0	94.1	81.5	61.1	94.3	61.1	65.1	53.8	61.6	70.6	71.6
RefineNet [12]	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70.0	73.6
FoveaNet [58]	98.2	83.2	91.5	44.4	51.2	63.2	70.8	75.5	92.7	70.1	94.5	83.3	64.2	94.6	60.8	70.7	63.3	63.0	73.2	74.1
PEARL [59]	98.4	84.5	92.1	54.1	56.6	60.4	69.0	74.0	92.9	70.9	95.2	83.5	65.7	95.0	61.8	72.2	69.6	64.8	72.8	75.4
TuSimple [32]	98.5	85.5	92.8	58.6	55.5	65.0	73.5	77.9	93.3	72.0	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8	77.6
SAC-multiple [60]	98.7	86.5	93.1	56.3	59.5	65.1	73.0	78.2	93.5	72.6	95.6	85.9	70.8	95.9	71.2	78.6	66.2	67.7	76.0	78.1
PSPNet [11]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
ResNet-38 [61]	98.5	85.7	93.1	55.5	59.1	67.1	74.8	78.7	93.7	72.6	95.5	86.6	69.2	95.7	64.5	78.8	74.1	69.0	76.7	78.4
Res101-SPP-Conv	98.5	85.5	92.5	53.0	60.1	62.7	73.4	77.2	93.1	71.4	94.7	84.4	66.0	95.5	63.3	79.5	68.8	68.0	75.2	77.0
Res101-SPP-DCU	98.6	86.2	92.7	53.7	61.2	64.1	74.4	78.4	93.3	72.0	94.7	85.4	67.8	95.7	64.7	80.4	70.3	69.3	76.2	77.8
Res101-SPP-DCU‡	98.6	86.6	92.8	51.8	62.6	64.4	74.7	78.8	93.4	72.4	94.9	85.9	69.6	95.8	71.7	83.7	74.9	69.9	76.9	78.9

ukuran tanaman kecil (473 × 473 di sini, dan 713 × 713 di [11]), hasil baseline kami sedikit lebih rendah dari PSPNet asli.
Faktanya, peningkatan terbaru pada kumpulan data Cityscapes berjalan lambat. Sebagian besar kemajuan memperkenalkan perhitungan berat dan lebih banyak parameter. Namun, teknik kami masih memperoleh manfaat dengan parameter serupa dan cukup ringkas untuk direproduksi.

V. KESIMPULAN DAN PEKERJAAN KE DEPAN

Dalam makalah ini, kami mengusulkan unit konvolusi padat (DCU) untuk segmentasi semantik. DCU menghasilkan multioutput dan memperkenalkan tumpang tindih spasial ke dalam konvolusi saat ini. Untuk mendapatkan peta segmentasi akhir, peta bobot diadopsi untuk memungkinkan kombinasi yang efektif dari aktivasi yang tumpang tindih dengan parameter yang dapat dipelajari. Studi ablasi pada dataset benchmark menunjukkan keefektifan pendekatan yang diusulkan dan kemampuannya yang unggul dibandingkan konvolusi standar. Secara keseluruhan, unit konvolusi padat merupakan komponen yang menjanjikan untuk segmentasi semantik dan dapat diadopsi secara luas dalam jaringan prediksi padat.

Di masa mendatang, kami akan mengeksplorasi strategi yang lebih terpadu dan elegan untuk mengaktifkan konvolusi yang padat di seluruh jaringan. Ini menunjukkan kaskade unit konvosi padat dan membutuhkan struktur yang lebih efektif untuk perhitungan dan kombinasi. Struktur yang dirancang khusus untuk jaringan segmentasi semantik seperti itu membantu mengurangi perhitungan berat dalam metode canggih dan sangat penting untuk penerapannya dalam aplikasi praktis.

REFERENSI

[1] K. Simonyan dan A. Zisserman, "Jaringan konvolusional yang sangat dalam untuk pengenalan gambar berskala besar," dalam *Proc. Int. Konf. Mempelajari. Mewakili.,* 2015, hlm. 1–14.

[2] K. He, X. Zhang, S. Ren, dan J. Sun, "Pembelajaran sisa yang mendalam untuk pengenalan gambar," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola,* Juni 2016, hlm. 770–778.

[3] G. Huang, Z. Liu, L. van der Maaten, dan KQ Weinberger, "Jaringan konvolusional yang terhubung rapat," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juli 2017, hlm. 2261–2269.

[4] L. Zhang *et al.,* "Meningkatkan segmentasi gambar semantik dengan bidang acak kondisional padat berbasis proba bilistic superpixel," *IEEE Access,* vol. 6, hlm. 15297–15310, 2018.

[5] L. Fan, W. Wang, F. Zha, dan J. Yan, "Menjelajahi tulang punggung baru dan modul perhatian untuk segmentasi semantik dalam adegan jalanan," *IEEE Access,* vol. 6, hlm. 71566–71580, 2018.

[6] TD Nguyen, A. Shinya, T. Harada, dan R. Thawonmas, "Penyempurnaan topeng segmentasi menggunakan transformasi gambar," *IEEE Access,* vol. 5, hlm. 26409–26418, 2017.

[7] Z. Jiang, Q. Wang, dan Y. Yuan, "Pemodelan dengan prasangka: Pembelajaran sampel kecil melalui musuh untuk segmentasi semantik," *Akses IEEE,* vol. 6, hlm. 77965–77974, 2018.

[8] M. Naseer, S. Khan, dan F. Porikli, "Indoor scene understanding in 2.5/3D for autonomous agents: A survey," *IEEE Access,* vol. 7, hlm. 1859–1887, 2019.

[9] J. Fu, J. Liu, Y. Wang, dan H. Lu, "Jaringan dekonvolusional yang terhubung rapat untuk segmentasi semantik," dalam *Proc. Int. Konf. Image Process.,* September 2017, hlm. 3085–3089.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, dan AL Yuille, "DeepLab: Segmentasi gambar semantik dengan jaring konvolusional yang dalam, konvolusi atrous, dan CRF yang terhubung sepenuhnya," *IEEE Trans. Pola Anal. Mesin Intell.,* vol. 40, tidak. 4, hlm. 834–848, April 2017.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, dan J. Jia, "Jaringan parsing adegan piramida," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juni 2017, hlm. 6230–6239.

[12] G. Lin, A. Milan, C. Shen, dan I. Reid, "RefineNet: Jaringan penyempurnaan multi-jalur untuk segmentasi semantik beresolusi tinggi," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juli 2017, hlm. 5168–5177.

[13] AW Harley, KG Derpanis, dan I. Kokkinos, "Jaringan konvolusional sadar segmentasi menggunakan topeng perhatian lokal," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Oktober 2017, hlm. 5048–5057.

[14] J. Long, E. Shelhamer, dan T. Darrell, "Jaringan yang sepenuhnya konvolusional untuk segmentasi semantik," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola,* Juni 2015, hlm. 3431–3440.

[15] X. Chen, G. Wang, C. Zhang, T.-K. Kim, dan X. Ji, "SHPR-NET: Regresi pose tangan semantik yang dalam dari awan titik," *IEEE Access,* vol. 6, hlm. 43425–43439, 2018.

[16] R. Wang, Y. Meng, W. Zhang, Z. Li, F. Hu, dan L. Meng, "Segregasi semantik penginderaan jauh untuk ekstraksi informasi air: Optimasi sampel melalui kinerja kesalahan pelatihan," *Akses IEEE,* vol. 7, hlm. 13383–13395, 2019.

[17] Z. Yang, H. Yu, W. Sun, Z. Mao, dan M. Sun, "Fitur yang dibagikan secara lokal: Alternatif yang efisien untuk bidang acak bersyarat untuk tasi segmen semantik," *IEEE Access,* vol. 7, hlm. 2263–2272, 2019.

[18] M. Martin-Abadal, E. Guerrero-Font, F. Bonin-Font, dan Y. Gonzalez-Cid, "Segmentasi semantik mendalam dalam AUV untuk identifikasi padang rumput Posidonia Oceanica Online," *IEEE Access,* vol. 6, hlm. 60956–60967, 2018.

[19] X. Jiang, Y. Gao, Z. Fang, P. Wang, dan B. Huang, "Segmentasi manusia end-to-end berdasarkan wilayah mengusulkan jaringan konvolusional penuh," *IEEE Access,* vol. 7, hlm. 16395–16405, 2019.

[20] L. Fan, H. Kong, W.-C. Wang, dan J. Yan, "Semantic segmentation with global encoding and dilated decoder in street scenes," *IEEE Access,* vol. 6, hlm. 50333–50343, 2018.

[21] Y. Duan *et al.,* "Model laten multinomial hierarkis dengan distribusi G0 untuk segmentasi semantik gambar radar aperture sintetis," *IEEE Access,* vol. 6, hlm. 31783–31797, 2018.

[22] S. Saito, R. Arai, dan Y. Aoki, "Seamline penentuan berdasarkan segmentasi semantik untuk mosaik gambar udara," *IEEE Access,* vol. 3, hlm. 2847–2856, 2015.

[23] D. Sakkos, ESL Ho, dan HPH Shum, "Illumination-aware multi-task GANs for foreground segmentation," *IEEE Access,* vol. 7, hlm. 10976–10986, 2019.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, dan FF Li, "ImageNet: Database gambar hierarkis berskala besar," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juni 2009, hlm. 248–255.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Memikirkan kembali konvolusi atrous untuk segmentasi gambar semantik." [Online]. Tersedia: https://arxiv.org/abs/1706.05587

[26] S. Zheng *et al.,* "Bidang acak bersyarat sebagai jaringan saraf berulang," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Desember 2015, hlm. 1529–1537.

[27] S. Chandra dan I. Kokkinos, "Inferensi cepat, tepat, dan multiskala untuk segmentasi gambar semantik dengan CRF Gaussian yang dalam," dalam *Proc. eur. Konf. Komputer.* Lihat , 2016, hlm. 402–418.

[28] J. Dai *et al.,* "Jaringan konvolusional yang dapat dideformasi," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Oktober 2017, hlm. 764–773.

[29] M. Lin, Q. Chen, dan S. Yan, "Jaringan dalam jaringan", dalam *Proc. Int. Konf. Mempelajari. Mewakili.,* 2014, hlm. 1–10.

[30] Y. Li, H. Qi, J. Dai, X. Ji, dan Y. Wei, "segmentasi semantik yang sadar sepenuhnya convolutional instance", dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juli 2017, hlm. 4438–4446.

[31] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, dan KQ Wein berger. (2017). "Jaringan konvolusional padat multi-skala untuk prediksi yang efisien." [Online]. Tersedia: https://pdfs.semanticscholar.org/6b86/ 7004cda7d2682159bc745d5ea1ef1bff48fc.pdf

[32] P. Wang *dkk.* (2017). "Memahami konvolusi untuk segmentasi semantik." [Online]. Tersedia: https://arxiv.org/abs/1702.08502

[33] C. Han, X. Tao, Y. Duan, dan J. Lu, "Konvolusi padat untuk segmentasi semantik," dalam *Proc. IEEE Int. Konf. Image Process.,* Oktober 2018, hlm. 2222–2226.

[34] F. Yu dan V. Koltun, "Agregasi konteks multi-skala oleh konvolusi melebar," dalam *Proc. Int. Konf. Mempelajari. Mewakili.,* 2016, hlm. 1–13.

[35] H. Noh, S. Hong, dan B. Han, "Belajar jaringan dekonvolusi untuk segmentasi semantik," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Desember 2015, hlm. 1520–1528.

[36] J. Fu, J. Liu, Y. Wang, dan H. Lu. (2017). "Jaringan dekon volusional bertumpuk untuk segmentasi semantik." [Online]. Tersedia: https://arxiv.org/abs/1708.04943

[37] O. Ronneberger, P. Fischer, dan T. Brox, "U-Net: Jaringan konvolusional untuk segmentasi gambar biomedis," dalam *Proc. Int. Konf. Kedokteran Komputasi Gambar. Comput.-Assist. Intervensi.,* 2015, hlm. 234–241.

[38] V. Badrinarayanan, A. Kendall, dan R. Cipolla, "SegNet: A deep convolu tional encoder-decoder architecture for scene segmentation," *IEEE Trans. Pola Anal. Mesin Intell.,* vol. 39, tidak. 12, hlm. 2481–2495, Desember 2017.

[39] L.-C. Chen, Y. Yang, J. Wang, W. Xu, dan AL Yuille, "Perhatian terhadap skala: segmentasi gambar semantik sadar-skala," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juni 2016, hlm. 3640–3649.

[40] M. Mostajabi, P. Yadollahpour, dan G. Shakhnarovich, "Segmentasi semantik umpan maju dengan fitur zoom-out", dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola,* Juni 2015, hlm. 3376–3385.

[41] B. Singh dan LS Davis. (2017). "Analisis invarian skala dalam deteksi objek–SNIP." [Online]. Tersedia: https://arxiv.org/abs/1711.08189 [42] T. Pohlen, A. Hermans, M. Mathias, dan B. Leibe, "Jaringan residu beresolusi penuh untuk segmentasi semantik dalam adegan jalanan," dalam *Prok. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juli 2017, hlm. 3309–3318.

[43] S. Hong, H. Noh, dan B. Han, "Decoupled jaringan saraf dalam untuk segmentasi semantik semi diawasi," dalam *Proc. Inf saraf. Proses. Sistem,* 2015, hlm. 1495–1503.

[44] G. Papandreou, L.-C. Chen, KP Murphy, dan AL Yuille, "Pembelajaran yang lemah dan semi-diawasi dari jaringan konvolusional yang dalam untuk segmentasi gambar semantik," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Desember 2015, hlm. 1742–1750.

[45] A. Bearman, O. Russakovsky, V. Ferrari, dan FF Li, "Apa gunanya: segmentasi semantik dengan pengawasan titik," dalam *Proc. eur. Konf. Com put.* Lihat , 2016, hlm. 549–565.

[46] P. Luo, G. Wang, L. Lin, dan X. Wang, "Pembelajaran ganda mendalam untuk segmentasi gambar semantik," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola,* Juni 2017, hlm. 2718–2726.

[47] I. Ahn dan C. Kim, "Pelabelan wilayah wajah dan rambut menggunakan segmentasi multipel berbasis pengelompokan spektral semi-diawasi," *IEEE Trans. Multime dia,* vol. 18, tidak. 7, hlm. 1414–1421, Juli 2016.

[48] M. Everingham, L. Van Gool, CKI Williams, J. Winn, dan A. Zisserman, "Tantangan kelas objek visual Pascal (VOC)," *Int. J.Komput. Lihat,* vol. 88, tidak. 2, hlm. 303–338, September 2009.

[49] M. Cordts *et al.,* "The cityscapes dataset for semantik urban scene under standing," in *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola,* Juni 2016, hlm. 3213–3223.

[50] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, dan J. Malik, "Kontur semantik dari detektor terbalik," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* November 2011, hlm. 991–998.

[51] J. Jia *et al.,* "Caffe: Convolutional architecture for fast feature embedding," dalam *Prok. 22th ACM Int. Konf. Multimedia,* 2014, hlm. 675–678.

[52] K. He, X. Zhang, S. Ren, dan J. Sun, "Mendalami penyearah: Melampaui kinerja tingkat manusia pada klasifikasi imagenet," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Desember 2015, hlm. 1026–1034.

[53] C. Peng, X. Zhang, G. Yu, G. Luo, dan J. Sun, "Masalah kernel besar— Tingkatkan segmentasi semantik dengan jaringan konvolusional global," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juli 2017, hlm. 1743–1751.

[54] I. Krešo, D. ŷauševiŷ, J. Krapac, dan S. Šegviŷ, "Invarian skala konvolusional untuk segmentasi semantik," dalam *Proc. Konferensi Jerman Pengenalan Pola.* Cham, Swiss: Springer, 2016, hlm. 64–75.

[55] Z. Liu, X. Li, P. Luo, C.-C. Loy, dan X. Tang, "Semantic image segmentation via deep parsing network," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Desember 2015, hlm. 1377–1385.

[56] G. Ghiasi dan CC Fowlkes, "rekonstruksi dan penyempurnaan piramida Laplacian untuk segmentasi semantik," dalam *Proc. eur. Konf. Komputer. Vis.* Cham, Swiss: Springer, 2016, hlm. 519–534.

[57] G. Lin, C. Shen, A. Van Den Hengel, dan I. Reid, "Pelatihan sepotong-sepotong yang efisien dari model terstruktur dalam untuk segmentasi semantik," dalam *Proc. Konferensi IEEE Komputer. Vis. Pengenalan Pola.,* Juni 2016, hlm. 3194–3203.

[58] X. Li *dkk.* (2017). "FoveaNet: Penguraian adegan perkotaan yang sadar perspektif." [Online]. Tersedia: https://arxiv.org/abs/1708.02421 [59] X. Jin *et al.,* "Penguraian adegan video dengan pembelajaran fitur prediktif," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Oktober 2017, hlm. 5581–5589.

[60] R. Zhang, S. Tang, Y. Zhang, J. Li, dan S. Yan, "konvolusi adaptif-skala untuk penguraian adegan," dalam *Proc. IEEE Int. Konf. Komputer. Vis.,* Oktober 2017, hlm. 2050–2058.

[61] Z. Wu, C. Shen, dan A. van den Hengel, "Lebih luas atau lebih dalam: Meninjau kembali model ResNet untuk pengenalan visual," *Pengenalan Pola,* vol. 90, hlm. 119–133, Juni 2019.

CHAOYI HAN menerima gelar BS dalam bidang teknik elektro dari Universitas Tsinghua, pada tahun 2016, di mana dia saat ini sedang mengejar gelar Ph.D. derajat. Minat penelitiannya meliputi pemrosesan gambar dan pembelajaran mesin.

YIPING DUAN menerima gelar BS dari School of Computer Science and Technol ogy, Henan Normal University, Xinxiang, China, pada tahun 2010, dan Ph.D. gelar dari School of Computer Science and Technology, Universitas Xidian, Xi'an, China, pada tahun 2016. Saat ini ia menjadi Asisten Riset di Departemen Teknik Elektro, Universitas Tsinghua. Minat penelitian sewanya saat ini meliputi penambangan semantik, pembelajaran mesin, dan pemrosesan gambar.

XIAOMING TAO menerima gelar BE dari Sekolah Teknik Telekomunikasi, Universitas Xidian, pada tahun 2003, dan gelar Ph.D. gelar dari Departemen Teknik Elektronika, Universitas Tsinghua, Beijing, China, pada tahun 2008, di mana dia saat ini menjadi Profesor di Departemen Teknik Elektronika. Minat penelitiannya meliputi komunikasi dan jaringan nirkabel, dan pemrosesan sinyal multimedia.

JIANHUA LU menerima gelar BE dan MS dari Universitas Tsinghua, Beijing, China, masing-masing pada tahun 1986 dan 1989, dan gelar Ph.D. gelar di bidang teknik listrik dan elektronik dari The Hong Kong University of Science and Technology, Hong Kong. Sejak 1989, dia bekerja di Departemen Teknik Elektro, Universitas Tsinghua, di mana dia saat ini menjabat sebagai Profesor. Dia telah menulis lebih dari 180 makalah teknis di jurnal internasional dan prosiding konferensi. Minat penelitiannya meliputi komunikasi nirkabel broadband, pemrosesan sinyal multimedia, dan jaringan nirkabel. Dia juga anggota IEEE Communication Society dan IEEE Signal Processing Society. Dia telah menjabat sebagai anggota Komite Program Teknis di berbagai konferensi IEEE dan menjabat sebagai Ketua Ketua Simposium Umum IEEE ICC 2008, serta Ketua Bersama Komite Program Konferensi Internasional IEEE Kesembilan tentang Informasi Kognitif matics, pada tahun 2010. Dia telah menjadi Anggota Aktif dari masyarakat profesional.

Dia adalah penerima penghargaan makalah terbaik di IEEE International Conference on Communications, Circuits and Systems 2002, ChinaCom 2006, dan IEEE Embedded-Com 2012, dan National Distinguished Young Scholar Fund oleh NSF Committee of China, pada tahun 2005. Dia adalah sekarang menjadi Kepala Ilmuwan Program Riset Dasar Nasional (973), Tiongkok.