Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

INSTYTUT STEROWANIA I ELEKTRONIKI PRZEMYSŁOWEJ
ZAKŁAD STEROWANIA

Praca dyplomowa inżynierska

na kierunku INFORMATYKA STOSOWANA
w specjalności Inżynieria oprogramowania

Poprawa rozdzielczości zdjęć przy użyciu krzyżowo-skalowej korelacji cech.

Aliaksandr Karolik
nr albumu 295138

promotor
dr inż. Grzegorz Sarwas

Warszawa 2020

Wykorzystanie głębokich sieci neuronowych do poprawy rozdzielczości zdjęć

Streszczenie

Niniejsza praca porusza problem wykorzystania ecia pecia do zrobienia czegoś wielkiego. W pracy przeanalizowane zostały algorytmy do wykrywania ecia pecia. Wybrane algorytmy zostały zaimplementowane i przebadane. Najlepsze rozwiązania zostały wykorzystane w zaprojektowanej i zbudowanej aplikacji.

Słowa kluczowe: praca dyplomowa, LaTeX, jakość

THESIS TITLE

Abstract

This thesis presents a novel way of using a novel algorithm to solve complex problems of filter design. In the first chapter the fundamentals of filter design are presented. The second chapter describes an original algorithm invented by the authors. Is is based on evolution strategy, but uses an original method of filter description similar to artificial neural network. In the third chapter the implementation of the algorithm in C programming language is presented. The fifth chapter contains results of tests which prove high efficiency and enormous accuracy of the program. Finally some posibilities of further development of the invented algoriths are proposed.

Keywords: thesis, LaTeX, quality

Warszawa, 1 lutego 2020

POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRYCZNY

OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa inżynierska pt. Poprawa rozdzielczości zdjęć przy użyciu krzyżowo-skalowej korelacji cech.:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Aliaksandr Karolik.....

Spis treści

Rozdział 1

Wstęp

W dobie dużej popularności cyfrowej rejestracji obrazów przy wykorzystaniu urządzeń mobilnych takich jak kamery, czy tablety za pomocą wbudowanych w nich aparatów fotograficznych jakość/rozdzielczość zarejestrowanych obrazów nie zawsze jest zadowalająca. Zarejestrowane materiały są w różnoraki sposób zakłócone, zniekształcone, co nie pozwala nam na wydruk, w odpowiedniej jakości, tego typu materiału. Ponieważ optyka zainstalowana w średniej półki telefonach komórkowych nie pozwala na uzyskanie wystarczającej jakości fotografii, widać wyraźne zapotrzebowanie na algorytmy poprawiające rozdzielczość i jakość zarejestrowanych obrazów.

Czym jest super-rozdzielczość? Super-rozdzielczość (pisana również jako super resolution, superresolution) jest określeniem zestawu metod zwiększania skali wideo lub obrazów. Terminy takie jak „skalowanie w górę", „powiększanie", „konwersja w górę" i „uprez" również opisują wzrost rozdzielczości w przetwarzaniu obrazu lub edycji wideo.

Większość technik super-rozdzielczości opiera się na tym samym pomyśle: wykorzystanie informacji z kilku różnych obrazów do stworzenia jednego powiększonego obrazu. Algorytmy próbują wyodrębnić szczegóły z każdego obrazu w sekwencji, aby zrekonstruować inne ramki. Obraz w wysokiej rozdzielczości oferuje dużą gęstość pikseli, a tym samym więcej szczegółów na temat oryginalnej sceny.

Potrzeba wysokiej rozdzielczości jest powszechna w wizji komputerowej dla lepszej wydajności w rozpoznawaniu wzorów i analizie obrazów. Wysoka rozdzielczość ma znaczenie w obrazowaniu medycznym dla diagnozy. Wiele aplikacji wymaga powiększenia określonego obszaru w którym niezbędna jest wysoka rozdzielczość, np. aplikacje do nadzoru, kryminalistyki i obrazowania satelitarnego.

Praca ta skupiać się będzie na badaniu rozwiązań algorytmicznych w dziedzinie widzenia komputerowego służących do poprawy rozdzielczości, zwa-

nych również algorytmami super-rozdzielczości (super-resolution).

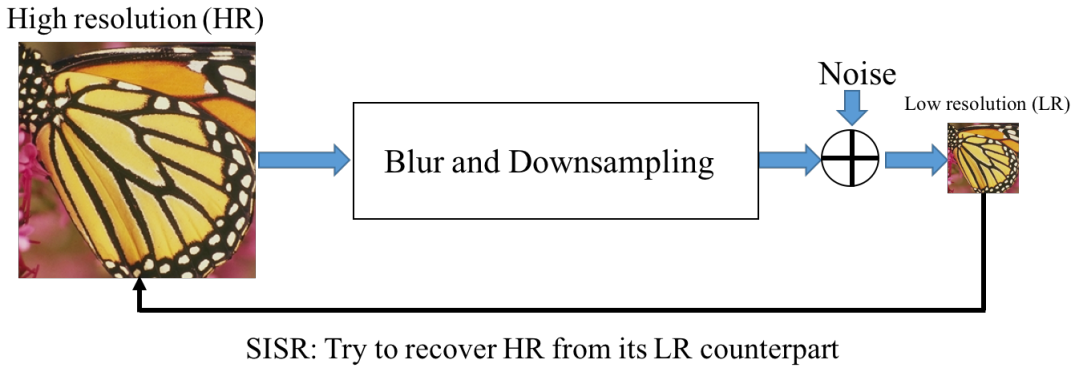
Rozdział 2

Algorytmy do poprawy rozdzielczości zdjęć

W tym rozdziale opisane zostaną podstawowe, jak i obecnie używane architektury sieci neuronowych do poprawy rozdzielczości zdjęć. Przedstawione zostaną teoretyczne podstawy działania oraz główne założenia budowy. Architektury sieci są tak dobrane, aby móc zaprezentować rozwój pomysłów ich twórców.

2.1 Wstęp teoretyczny

Super-rozdzielczość (SR) odnosi się do zadania przywracania obrazów o wysokiej rozdzielczości z jednej lub więcej obserwacji tej samej sceny w niskiej rozdzielczości (LR). Zgodnie z liczbą wejściowych obrazów LR, SR można podzielić na super-rozdzielczość pojedynczego obrazu (SISR) i super-rozdzielczość wielu obrazów (MISR). W porównaniu z MISR, SISR jest znacznie bardziej popularną metodą ze względu na wysoką wydajność. Typowa struktura SISR, jest zaprezentowana na rysunku 2.1.



Rysunek 2.1: Szkic SISR

Głównie algorytmy SISR dzielą się na trzy kategorie: metody oparte na interpolacji, metody oparte na rekonstrukcji oraz metody oparte na uczeniu. Metody SISR oparte na interpolacji, takie jak interpolacja dwusześcienna (bicubic interpolation) i próbkowanie Lanczosa (Lanczos resampling), są bardzo szybkie i proste, ale dość nie dokładne.

Metody SR oparte na rekonstrukcji, często przyjmują zaawansowaną wcześniejszą wiedzę w celu ograniczenia możliwej przestrzeni rozwiązań z korzyścią polegającą na generowaniu elastycznych i ostrych szczegółów. Jednak wydajność wielu metod opartych na rekonstrukcji szybko spada, gdy zwiększa się skala, oraz metody te są zwykle czasochłonne.

Metody SISR oparte na uczeniu, znane również jako metody oparte na przykładach, najczęściej używane ze względu na ich szybkie obliczenia i wyjątkową wydajność. Metody te zwykle wykorzystują algorytmy uczenia maszynowego do analizy związków statystycznych między LR i odpowiadającym mu odpowiednikiem HR z istotnych przykładów szkoleniowych.

Technika MISR wykorzystuję jako wejście zestaw obrazów niskiej rozdzielczości do budowy obrazu HR, ale jak już wcześniej było wspomniane, SISR jest popularniejsza ze względu na wysoką wydajność.

2.2 Konwolucyjna sieć neuronowa

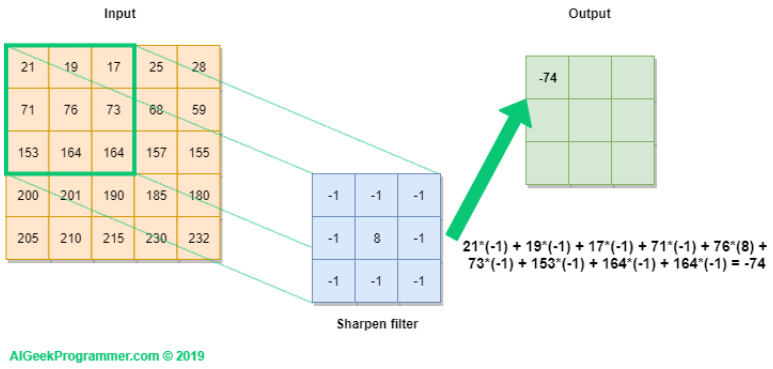
Konwolucyjne sieci neuronowe (CNN) są prawie wszędzie. Jest to prawdopodobnie najbardziej popularna architektura głębokiego uczenia. Niedawny wzrost zainteresowania głębokim uczeniem wynika z ogromnej popularności i skuteczności konwulsyjnych sieci neuronowych. Zainteresowanie CNN rozpoczęło się od AlexNet w 2012 roku i od tego czasu rosło wykładniczo. W

ciągu zaledwie trzech lat, naukowcy przeszli z 8-warstwowej sieci AlexNet do 152-warstwowej sieci ResNet.

CNN jest obecnie modelem go-to dla każdego problemu związanego z obrazem. CNN również stosowane w systemach rekomendacji, przetwarzania języka naturalnego i nie tylko. Główną zaletą sieci CNN w porównaniu z jej poprzednikami jest to, że automatycznie wykrywa ona ważne cechy bez żadnego nadzoru człowieka. Na przykład, biorąc pod uwagę wiele zdjęć kotów i psów, sieć sama uczy się cech charakterystycznych dla każdej klasy.

Podstatowym narzędziem sieci jest warstwa konwolucyjna. Warsta konwolucyjna składa się z zestawu filtrów, zadaniem której jest wykrycie poszczególnych cech ze zdjęcia.

Mnożenie splotowe lub konwolucja to operacja matematyczna polegająca na połączeniu dwóch zestawów informacji. W naszym przypadku konwolucja jest stosowana na danych wejściowych oraz filtru. W wyniku powstaje nowa macierz, która jest nazywana mapą cech, wartości mapy są wynikiem kombinacji liniowej poszczególnych pikseli obrazu wejściowego i przesuwającego się filtra. Poniżej na rysunku 2.2 zaprezentowany jest shemat mnożenia splotowego lub konwolucji:

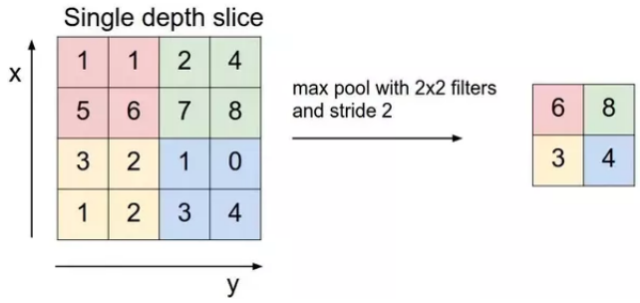


Rysunek 2.2: Schemat mnożenia splotowego (konwolucji)

Tak samo jak w zwykłej sieci, po warstwie konwolucyjnej występuje warstwa aktywacji (najczęściej używana jest funkcja ReLU), zadaniem której jest wprowadzenie nieliniowości do sieci.

Drugim podstawowym elementem sieci konwolucyjnej jest warstwa łączącą (pooling layer). Zadaniem jej jest zmniejszenie wymiarów mapy cech, wyznaczonej w warstwie konwolucyjnej, przy zachowaniu jej kluczowych elementów. Warstwa ta również odpowiada za redukcję szumu. Najczęściej używaną metodą jest „Max pooling".

Algorytm działania metody Max pooling jest następujący definiowany jest filtr oraz krok przesunięcia. Kolejne wartości macierzy wyjściowej są maksymalną wartością objętą filtrem. Na rysunku 2.3 zaprezentowany jest shemat działania metody Max pooling:



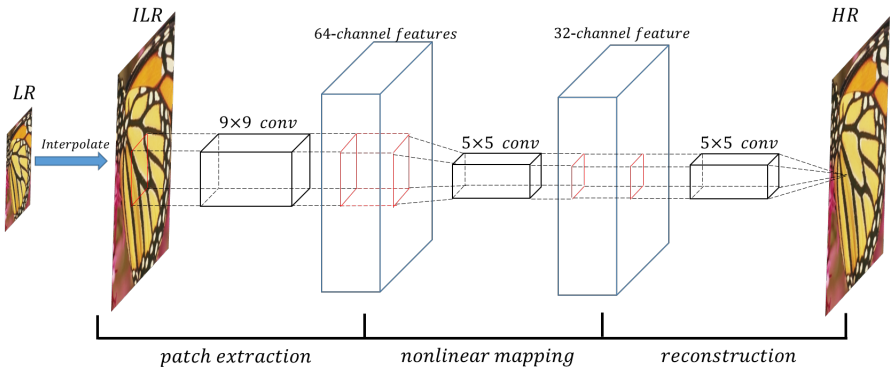
Rysunek 2.3: Schemat metody Max pooling

2.3 Konwolucyjne sieci neuronowe dla superrodzielczości

Pierwszą zapropanowaną architektórę używającą CNN do mapowania obrazów niskiej rozdzielczości do wysokiej jest SRCNN(Image Super-Resolution Using Deep Convolutional Networks). Architekturę SRCNN zaprezentowana jest na rysunku 2.4. SRCNN jest trójwarstwowym CNN (Konwolucyjne sieci neuronowe), w którym znajdują się rozmiary filtrów każdej warstwy $64 \times 1 \times 9 \times 9$, $1 \times 32 \times 5 \times 5$ i $1 \times 32 \times 5 \times 5$. Każda warstwa odpowiada za następujące czynności:

1. Wyodrębnienie i reprezentacja. Obraz jest przepuszczany przez zestaw filtrów. Zadaniem których jest wyodrębnienie cech specyficznych.
2. Mapowanie nieliniowe. Dla każdej mapy cech wyprodókowanych w poprzedniej warstwie, przyporządkowywany jest wektor cech o wysokiej rozdzielczości.
3. Rekonstrukcja. Ostatnia warstwa konwulacyjna układa wektory cech uzyskane w poprzedniej warstwie w jeden obraz wysokiej rozdzielczości.

Autorzy proponujące architekturę SRCNN przewidywali, że zwiększenie ilości warst konwulocyjnych korzystnie wpłynie na wyniki. Niedługo po ich publikacji zaczęły pojawiać się rozwiązania o głębszych architekturach. Kim



Rysunek 2.4: Architektura SRCNN

i in. zaproponowali bardzo głęboki model VDSR z ponad 16 warstwami konwolucyjnymi korzystającymi ze skutecznego uczenia rezydualnego (resztkowego,residual learning). Aby w pełni wykorzystać moc głębokich CNN, Lim i in. zintegrowali bloki rezydualne w framework SR, w wyniku powstały modeli EDSR i MDSR.

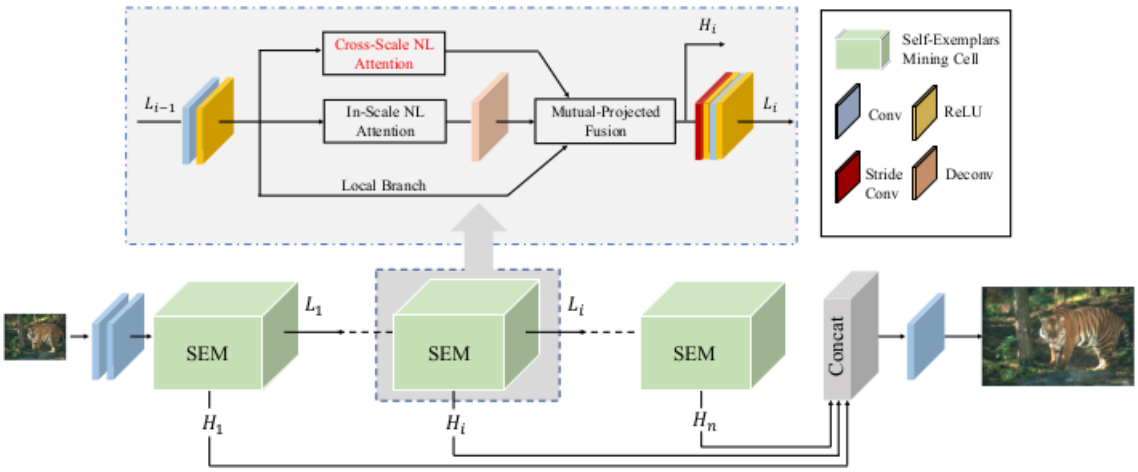
2.4 Model CSNLN

Modeli do poprawy rozdzielczości pojedynczego obrazu oparte na dużej ilości warst konwolucyjcnych wykorzystują korzyści płynące z dużych zewnętrznych zasobów obrazu do lokalnej odbudowy, jednak w większości istniejących prac pominięto dalekosiężne podobieństwa cech charakterystycznych. Niektóre z ostatnich prac z powodzeniem wykorzystały te wewnętrzne korelacje cech, używając nielokalne moduły uwagi.

Jednakże istniejące podejścia do odtwarzania obrazów używały jedynie podobieństwa cech w tej samej skali, ignorując liczne wewnętrzne wzorce LR-HR w różnych skalach, co prowadziło do stosunkowo niskiej wydajności. Wiadomo, że wewnętrzne korelacje HR zawierają bardziej istotne informacje o wysokich częstotliwościach. W tym celu, Yiqun Mei i in. w swoim artykule proponują pierwszy moduł uwagi Cross-Scale Non-Local (CS-NL) z integracją do rekurencyjnej sieci neuronowej.

Proponowana architektura sieci przedstawiona jest na rysunku 2.5. Jest to w zasadzie rekurencyjna sieć neuronowa, w której każda komórka zwana Self-Exemplars Mining (SEM) w pełni integruje uczenie oparte o lokalne czynniki, nielokalne czynniki w tej samej skali obrazów oraz w skali zmiennej używając nowozapropanowany moduł CS-NL.

Powtarzające się komórki SEM są osadzone w sekwencyjną strukturę, jak pokazano na rysunku 2.5. Każda komurka SEM produkuje H_i przekształconą mapę cech powstającą z połączenia wyników dwóch modułów uwagi oraz mapy cech otrzymanej ze zwykłej warstwy konwolucyjnej. Wudobyte mapy cech H_i łączone jedą warstwą konwulacyjną w wyniku łaczenia powstaje obraz wysokej rozdzielczości.



Rysunek 2.5: Architektura CSNLN

2.4.1 Moduły uwagi

Jak było już wspomniano wcześniej wybrany algorytm wyszukuje nielokalne korelacje w obrazach w tej samej skali oraz w skali zmiennej. Korelacje poszukiwane są za pomocą dwóch modułów uwagi nielokalnej o nazwach In-Scale Non-Local Attention (IS-NL) oraz Cross-Scale Non-Local Attention (CS-NL).

Nielokalna uwaga może eksplorować autopróbki poprzez podsumowanie cechy charakterystyczne ze zbioru obrazów wejściowych. Autorzy publikacji używając artykułu zdefiniowali nielokalną uwagę dla wybranej X mapy cech obrazu za pomocą wzora 2.1.

$$Z_{i,j} = \sum_{g,h} \frac{\exp(\phi(X_{i,j}, X_{g,h}))}{\sum_{u,v} \exp(\phi(X_{i,j}, X_{u,v}))} \psi(X_{g,h}), \tag{2.1}$$

gdzie:

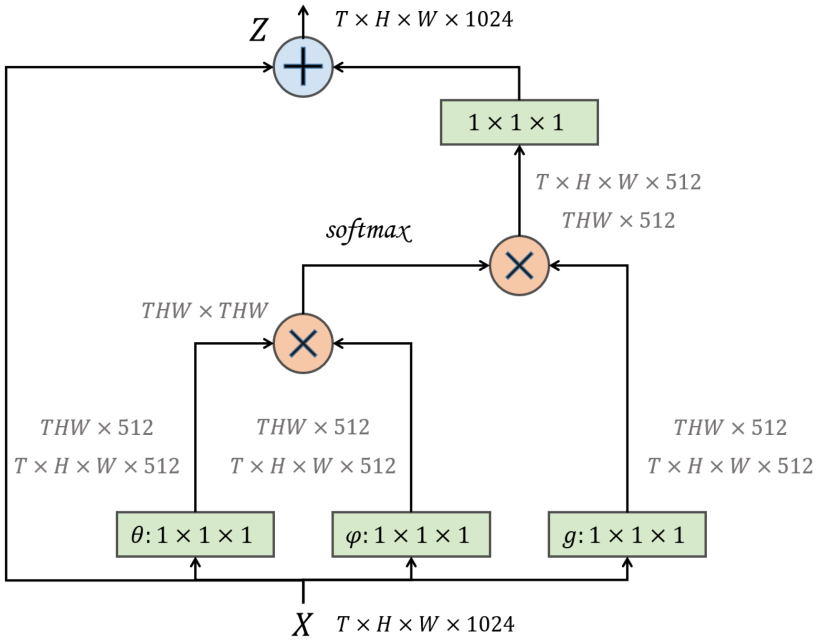
- $(i, j), (g, h)$ oraz (u, v) pary kordynat mapy X
- $\phi(.)$ funkcja transformacji cech
- $\psi(., .)$ funkcją korelacji do pomiaru podobieństwa. Definiowana w sposób 2.2

$$\psi(X_{i,j}, X_{g,h}) = \theta(X_{i,j})^T \delta(X_{g,h}), \tag{2.2}$$

gdzie:

- $\theta(.), \delta(.)$ funkcje transformacji cech

Warto zauważyć że zaproponowany wzór może być wykorzystywany dla obrazów w tej samej skali. W oparciu o wzór 2.1 autorzy zaimplementowali moduł IS-NL poniżej na rysunku 2.6 jest graficzna reprezentacja.

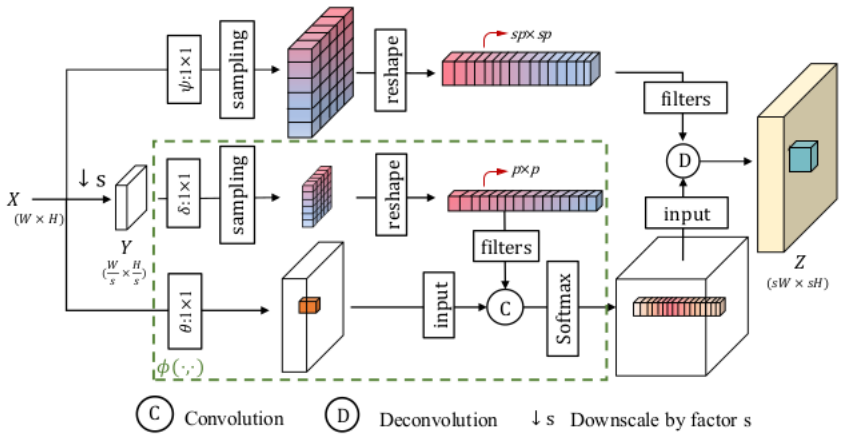


Rysunek 2.6: Moduł IS-NL

Powyższe sformułowanie 2.1 zostało rozszerzone do wersji używającej krzyżowo-skalową korelację cech. Zamiast pomiaru wzajemnej korelacji pikselowej jak jest robione IS-NL module, proponowany nowy moduł uwagi ma na celu pomiar korelacji pomiędzy pikselami o niskiej rozdzielczości a plastrami większej skali obrazu LR.

Na rysunku 2.7 zaprezentowana nowa architektura modułu uwagi wykorzystująca krzyżowo-skalową korelację cech zaprojektowana w oparciu o 2.1

przez twórców artykula. Moduł CS-NL działa w następujący sposób wejściowa mapa cech X o wymiarach W, H przekształcana jest przy pomocy interpolacji dwuliniowej do Y w skali s. Następnie plastry $p \times p$ z mapy X są porównywany do $p \times p$ koordynat w mapie Y aby uzyskać wynik dopasowania softmax. Na samamym końcu jest używana warstwa dekonwolucyjna na wynikach softmaxa oraz wagach plastra o wymiarze sp, sp wydobywana z mapy X. W wyniku powstaje mapa Z o wymiarach (sW, sH) która będzie s razy bardziej określona niż X.



Rysunek 2.7: Moduł CS-NL

2.4.2 Mutual-Projected Fusion

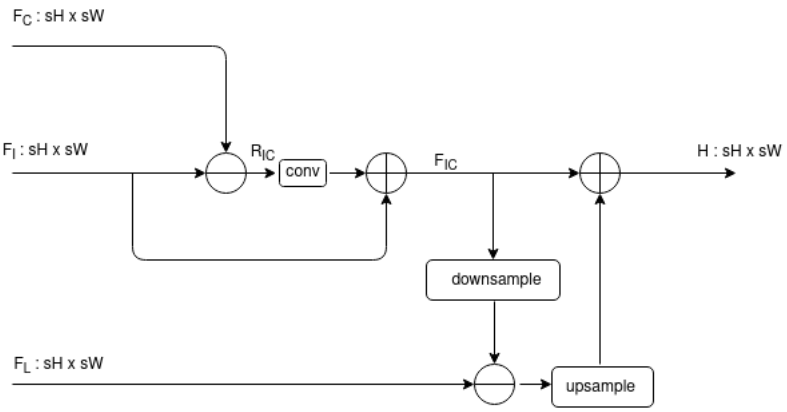
Jak wcześniej było wspomniane każda komurka SEM zawiera trójbranżową strukturę przy pomocy której generuje trzy mapy cech poprzez niezależne wykorzystanie każdego ze źródeł informacji z obrazów LR, niejasne pozostaje, jak połączyć te oddzielne tensory w kompleksową mapę funkcji. Autorzy zaproponowali własny algorytm do stopniowego łączenia cech. Procedura algorytmu została przedstawiona na rysunku 2.8.

Aby pozwolić sieci skupić się na bardziej informacyjnych cechach najpierw jest obliczana różnica R_{IC} pomiędzy dwoma mapami cech F_I i F_C z modułów IS-NL oraz CS-NL odpowiednio. Następnie wynikowa mapa R_{IC} jest przepuszczana przez jedną warstę conwolucyjną później z powrotem jest dodawana mapa F_I jest to robione do przewrócenia straconych informacji przy operacji odejmowania.

$$R_{IC} = F_I - F_C \tag{2.3}$$

$$F_{IC} = conv(R_{IC}) + F_I \tag{2.4}$$

Pozostająca cecha R_{IC} reprezentuje szczegóły istniejące w jednym źródle, a brakujące w drugim. Taka projekcja pozwala sieci skupić się tylko na odrębnych informacjach pomiędzy źródłami, omijając przy tym powszechną wiedzę, co zwiększa jej zdolność dyskryminacyjną.



Rysunek 2.8: Algorytm łączenia map cech

Wzorując się artykułem DBPN, autorzy zadoptowali podejście back-projection w celu włączenia lokalnych informacji, aby dodać regularyzację cech i skorygować błędy rekonstrukcji. Wynikowa mapa H jest obliczana w następujący sposób.

$$e = F_L - downsample(F_{IC}), \tag{2.5}$$

$$H = upsample(e) + F_{IC} \tag{2.6}$$

gdzie F_L jest mapą cech z gałędzie lokalnej (Local branch), *downsample* jest dokanywany przy pomocy warstwy konwolucyjnej do zmienjszenia wymiaru oraz *upsample* odpowiednio warstwa konwolucyjna do przewrócenia wymiarów do $sH \times sW$.

Zaproponawany algorytm gwarantuje uczenie resztkowe (residual learning) przy jednoczesnym łączeniu różnych źródeł cech, co umożliwia bardziej dyskryminacyjne uczenie cech w porównaniu z zwykłym dodawaniem lub łączeniem.

2.5 Metryki porównania jakości modeli

Do porównania wyników działania algorytóm zostaną wykorzystane następujące metryki:

- Szczytowy stosunek sygnału do szumu (PSNR, ang. peak signal-to-noise ratio)
- Podobieństwo strukturalne (SSIM, ang. structure similarity)

2.5.1 Szczytowy stosunek sygnału do szumu

Szczytowy stosunek sygnału do szumu, rzadziej nazywany jako stosunek sygnału szczytowego do szumu (PSNR, ang. peak signal-to-noise ratio) – stosunek maksymalnej mocy sygnału do mocy szumu zakłócającego ten sygnał. Ze względu na szeroki zakres wartości PSNR wyrażany jest w decybelach.

Najczęściej PSNR stosowany jest do oceny jakości kodeków wykorzystujących stratną kompresję obrazów. W takim przypadku sygnałem są nieskompresowane dane źródłowe, a szumem – artefakty (zniekształcenia) spowodowane zastosowaniem kompresji stratnej.

W celu wyznaczenia PSNR należy najpierw obliczyć współczynnik MSE (ang. mean squared error) bazujący na obu porównywanych obrazach, wykorzystując wzór 2.7. A później używając współczynik MSE obliczyć szczytowy stosunek sygnału do szumu za pomocą wzoru 2.8.

$$MSE = \frac{1}{M * N} \sum_{i=1}^{N} \sum_{j=1}^{M} ([f(i,j) - f'(i,j)]^2) \tag{2.7}$$

Gdzie:

- N, M - wymiary obrazu w pikselach,
- $f(i,j)$- wartość piksela o współrzędnych (i,j) obrzu oryginalnego,
- $f'(i,j)$- wartość piksela o wpółrzędnych (i,j) obrazu skompresowanego.

$$PSNR = 10log_{10}\frac{[max(f(i,j))]^2}{MSE} \tag{2.8}$$

Gdzie:

- $max(f(i,j))$ – wartość maksymalna danego sygnału; w przypadku obrazów zwykle jest to wartość stała, np. dla obrazów monochromatycznych o reprezentacji 8-bitowej wynosi 255.

2.5.2 Podobieństwo strukturalne

Podobieństwo strukturalne miarą indeksu (SSIM) jest metodą przewidywania postrzeganej jakości telewizji cyfrowej i obrazów filmowych, a także

innych rodzajów zdjęć i filmów cyfrowych. Indeks SSIM jest metodą pełnego porównania, innymi słowy, mierzy jakość w oparciu o oryginalny obraz (nie skompresowany lub bez zniekształceń). Wskaźnik SSIM jest rozwinięciem tradycyjnych metod, takich jak PSNR (peak signal-to-noise ratio) i standardowej metody średniego błędu kwadratowego (MSE), które okazały się niezgodne z fizjologią ludzkiego postrzegania.

Indeks SSIM zawiera się w przedziale od -1 do $+1$. Wartość $+1$ jest osiągana tylko wtedy, gdy próbki są w pełni autentyczne. Indeks SSIM jest obliczany w różnych oknach obrazu. Miara pomiędzy dwoma oknami x i y o wspólnym rozmiarze $N \times N$ obliczana jest za pomocą wzoru 2.9.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2.9}$$

Gdzie:

- μ_x wartość średnia z x
- μ_y wartość średnia z y
- σ_x^2 wariancja x
- σ_y^2 wariancja y
- σ_{xy} kowariancja x i y
- $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ dwie zmienne stabilizujące podział ze słabym mianownikiem
- L zakres dynamiki wartości pikseli
- $k_1 = 0.01$ i $k_2 = 0.03$

Bibliografia