Appendix

A Toy Experiment

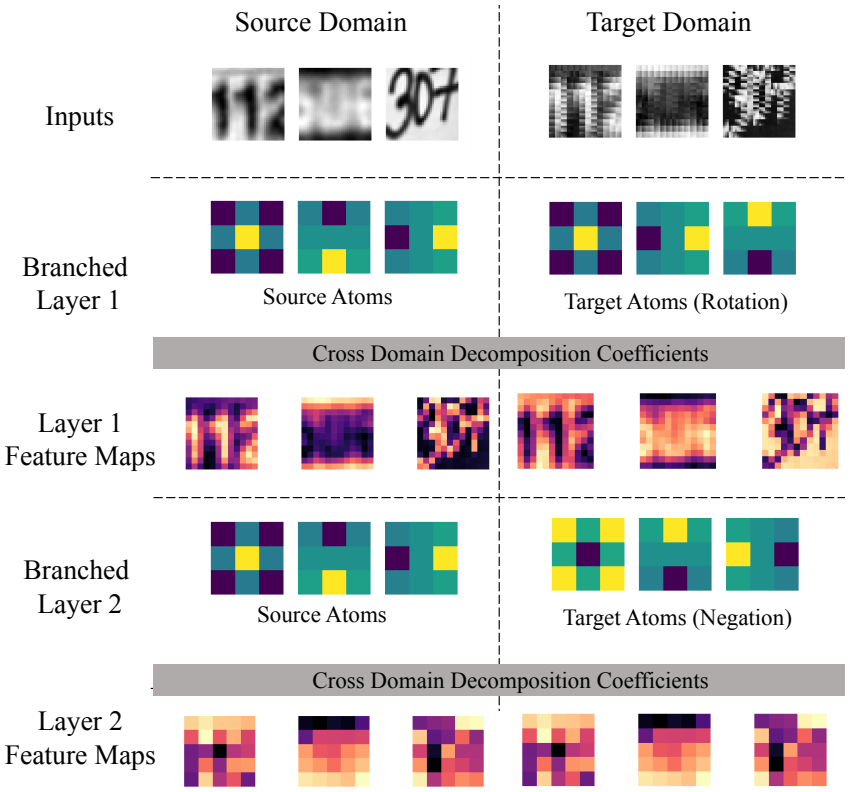


Figure A.1: Visualization of the toy example. The two columns visualize the inputs, features, and atoms of the source domain and the target domain, respectively. Only the output feature in the first channel of each convolutional layer is visualized for comparison. Domain invariant features, the last row, are obtained by manually adapting source domain atoms to generated target domain atoms.

B Dataset Samples and Qualitative Results

B.1 Unsupervised DA for Image Segmentation

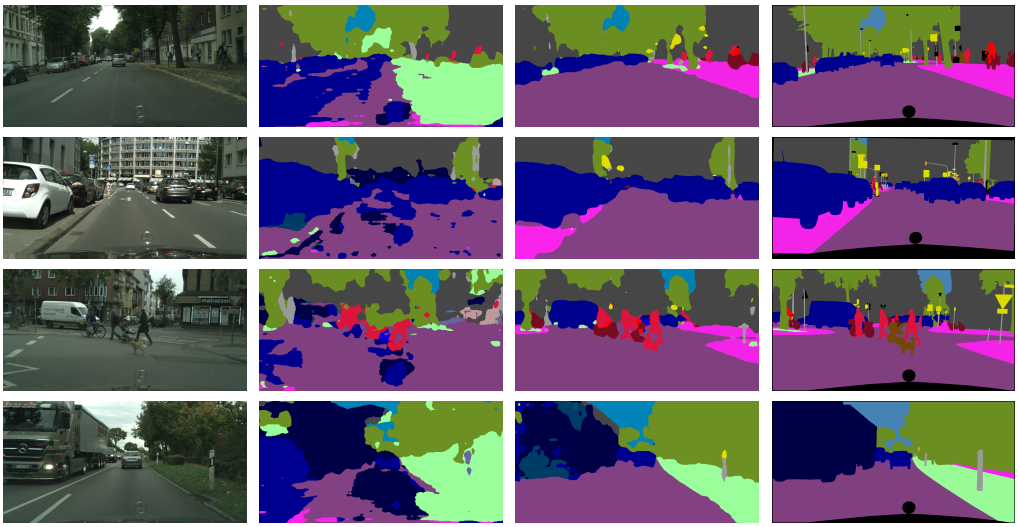
For the image segmentation experiments in Section 5.2, we provide more qualitative results in Figure A.2.

C Computation and Parameters

In Table A.1, we provide comparisons on additional parameters and computation introduced by one extra domain with and without the proposed domain-adaptive filter decomposition. The comparison reveals that domain-adaptive filter decomposition not only delivers superior performances but also saves both parameters and computation significantly.

D Comparisons to Domain Separation Networks (DSN).

Domain separation networks (DSN) [1] share the motivation with us by using 'domain specific' and 'domain shared' network components to improve domain invariant feature learning. However,



(a) Target domain image. (b) Before adaptation. (c) After adaptation. (d) Ground truth.

Figure A.2: Qualitative results for domain adaptation segmentation. The samples are randomly selected from the validation subsets of Cityscapes.

Table A.1: Comparisons on additional parameters and computation introduced by one extra domain. Comparisons are performed on VGG-16, with 6 dictionary atoms and the input size of 224×224.

Model	Regular VGG	VGG with DAFD
Parameters	14.71M	0.0007M
Flops	15.38G	10.75G

- Our method works as a plug-and-play module as demonstrated with the numerous architectures in the experiments, with no additional loss functions as in [1], which consequently introduces additional hyperparameters to tune.

- Our method introduces only hundreds of parameters to model one extra domain; while in DSN, three encoder networks and one decoder networks are required to model two domains, which introduces many times more parameters.

- The aforementioned additional costs also prevent DSN from being extended to large-scale experiments like the unsupervised image segmentation which can be however easily performed by using the proposed DAFD with no additional training objectives and neglectable parameter overheads.

Despite the remarkable simplicity, DAFD is comparable to DSN according to the performance on SVHN->MNIST. Due to limited overlap of the experiments reported in DSN with ours, we show additional comparisons in Table A.2 by reimplementing DSN (since the code link provided with the original paper is taken down now), and we will add the discussion to the final revision and more experiments in the supplementary.

Table A.2: Comparisons to domain separatioon networks (DSN) with DANN as underlying method. Datasets include USPS (U), SVHN (S), MNIST (M), MNIST-M (MM), Synth Digits (SD), Synth Signs (SS), and GTSRB(G). $*$ denotes numbers obtained by our reimplementation.

Methods	M→U	U→M	S→M	M→MM	SD→S	SS→G
DSN	90.6$*$	92.1$*$	82.7	83.2	91.2	93.1
Ours	92.3	95.4	83.2	86.2	91.7	94.0

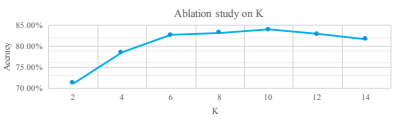


Figure A.3: Ablation study on K performed on SVHN→MNIST with DANN as underlying method.

E Ablation Study on the Number of Atoms.

We present the ablation study on K here in Figure A.3, which shows that DAFD is only sensitive to very small K, which degrades the expressiveness.

F Atom Visualizations

We visualize trained atoms for the digit experiments in Figure A.4. Strong correspondence is observed from atoms across domains.

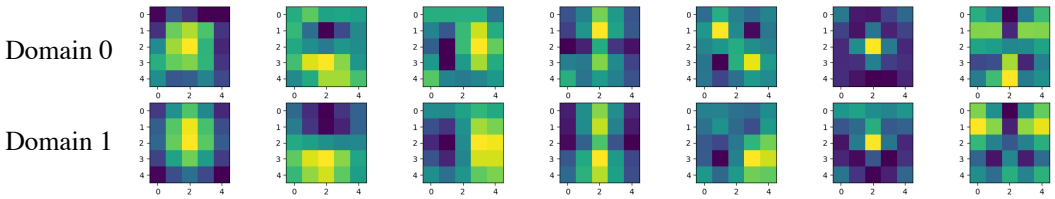


Figure A.4: Visualizations of atoms for both domains trained on the digit experiment. Atoms across domains have strong correspondence.

G Correction of a Single Filter Transform

We first analyze the "symmetric" correction of one filter spatial transform D_τ in one layer. The inclusion of linear correspondence transform is more direct. For technical reasons, we assume that the displacement field τ is a small distortion, namely $\|\nabla\tau\|_\infty \ll 1$, and then D_τ is invertible. Example includes rotation by a small angle and a small factor rescaling (dilation).

For simplicity we only consider one input and output channel in each of the multiple convolutional layers. The argument extends to multiple channels by modifying the boundedness condition of the filters. Then the forward mapping in one convolutional layer can be written as $y = \sigma(x * w + b)$, where x is the input activation, y is the output, w is the filter, b is the constant bias, and σ is the nonlinear activation function, e.g., ReLU. As we take a continuous formulation in the analysis, the activations x and y are assumed to be smooth functions supported on domain $\Omega \subset \mathbb{R}^2$, typically $\Omega = [-1, 1]^2$. The filter w is a function supported on $2^j B$, B being the unit disk, and 2^j is layer scale (diameter of filter patches) . The 1-norm of a function is defined to be $\|x\|_1 = \int_{\mathbb{R}^2} |x(u)|du$.

Lemma 1. *Suppose that the two filters w, f are supported on $2^{j_w} B$ and $2^{j_f} B$ respectively. $\sigma : \mathbb{R} \to \mathbb{R}$ is non-expansive, D_τ is a spatial transform where τ is odd, i.e., $\tau(-u) = -\tau(u)$, and $|\nabla\tau|_\infty < \frac{1}{5}$. Then*

$$\begin{aligned}&\|\sigma_b(x * D_\tau w) * f - \sigma_b(x * w) * D_\tau^{-1} f\|_1 \\ &\quad \le 2|\nabla\tau|_\infty \|w\|_1 \|f\|_1 \left\{(2^{j_w} + 2^{j_f})\|\nabla x\|_1 + 4\|x\|_1\right\},\end{aligned}$$

where σ_b denotes the nonlinear function with the bias. The second term vanishes if $(I_d - \tau)$ is a rigid motion, e.g., rotation.

Proof of Lemma 1. We establish a few facts:

Fact 1. $|\nabla\tau|_\infty < \frac{1}{5}$ guarantees that, $\rho := I_d - \tau$,

$$||J\rho| - 1|, ||J\rho^{-1}| - 1| \le 4|\nabla\tau|_\infty, \tag{A.1}$$

where $Jf = \det(\nabla f)$ denotes the determinate of the Jacobian matrix of the mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$. The inequality can be verified by elementary calculation. When ρ is a rigid motion then the r.h.s of (A.1) is zero.

Fact 2. ρ is invertible, and odd symmetry of τ implies that ρ and thus ρ^{-1} are odd, namely $-\rho^{-1}(-u) = \rho^{-1}(u)$.

Define

$$\begin{aligned} y_1(u) &:= \sigma_b(x * D_\tau w) * f(u) \\ &= \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+v-z) w(\rho(z)) dz \right) f(-v) dv \\ &= \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+v-\rho^{-1}(\tilde{z})) w(\tilde{z}) |J\rho^{-1}(\tilde{z})| d\tilde{z} \right) f(-v) dv \end{aligned}$$

and

$$\hat{y}_1(u) := \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+v-\rho^{-1}(\tilde{z})) w(\tilde{z}) d\tilde{z} \right) f(-v) dv.$$

We have that

$$\begin{aligned} |y_1(u) - \hat{y}_1(u)| &\le \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} |x(u+v-\rho^{-1}(\tilde{z}))| |w(\tilde{z})| \left| |J\rho^{-1}| - 1 \right| |f(-v)| d\tilde{z} dv \text{ (by } \sigma_b \text{ non-expansive)} \\ &\le 4 |\nabla \tau|_\infty \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} |x(u+v-\rho^{-1}(\tilde{z}))| |w(\tilde{z})| |f(-v)| d\tilde{z} dv \text{ (by Fact 1)} \end{aligned}$$

and thus

$$\|y_1 - \hat{y}_1\|_1 \le 4 |\nabla \tau|_\infty \|x\|_1 \|w\|_1 \|f\|_1. \tag{A.2}$$

When ρ is a rigid motion, $y_1 = \hat{y}_1$.

Also, let

$$\begin{aligned} y_2(u) &:= \sigma_b(x * w) * D_\tau^{-1} f(u) \\ &= \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+v-z) w(z) dz \right) f(-\rho^{-1}(v)) dv \quad \text{(by Fact 2)} \\ &= \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+\rho(\tilde{v})-z) w(z) dz \right) f(-\tilde{v}) |J\rho(\tilde{v})| d\tilde{v} \end{aligned}$$

and

$$\hat{y}_2(u) := \int_{\mathbb{R}^2} \sigma_b \left(\int_{\mathbb{R}^2} x(u+\rho(\tilde{v})-z) w(z) dz \right) f(-\tilde{v}) d\tilde{v}.$$

Similar to the proof of (A.2), one can verify that

$$\|y_2 - \hat{y}_2\|_1 \le 4 |\nabla \tau|_\infty \|x\|_1 \|w\|_1 \|f\|_1, \tag{A.3}$$

and the bound is zero when ρ is a rigid motion.

It remains to bound $\|\hat{y}_1 - \hat{y}_2\|_1$. Note that by σ_b being non-expansive again

$$|\hat{y}_1(u) - \hat{y}_2(u)| \le \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} |x(u+v-\rho^{-1}(z)) - x(u+\rho(v)-z)| |w(z)| dz |f(-v)| dv. \tag{A.4}$$

We claim that

$$\int_{\mathbb{R}^2} |x(u+v-\rho^{-1}(z)) - x(u+\rho(v)-z)| du \le |\nabla \tau|_\infty 2(2^{j_w} + 2^{j_f}) \|\nabla x\|_1 \tag{A.5}$$

uniformly for v and z. If true, with (A.4) it gives that

$$\int_{\mathbb{R}^2} |\hat{y}_1(u) - \hat{y}_2(u)| du \le |\nabla \tau|_\infty 2(2^{j_w} + 2^{j_f}) \|\nabla x\|_1 \|w\|_1 \|f\|_1$$

which proves the lemma together with (A.2) and (A.3).

Proof of (A.5): We verify that for any fixed v, z,

$$\int_{\mathbb{R}^2} |x(u+v-\rho^{-1}(z)) - x(u+\rho(v)-z)| du \le \|\nabla x\|_1 |\nabla \tau|_\infty |v - \rho^{-1}(z)|, \tag{A.6}$$

by a direct calculation:

$$\begin{aligned} \text{(l.h.s)} &\le \|\nabla x\|_1 |(v - \rho^{-1}(z)) - (\rho(v) - z)| \\ &= \|\nabla x\|_1 |\tau(v) - \tau(\rho^{-1}(z))| \\ &\le \|\nabla x\|_1 |\nabla \tau|_\infty |v - \rho^{-1}(z)|. \end{aligned}$$

Then, combined with that $v \in 2^{j_f}B$ thus $|v| \leq 2^{j_f}$, and $z \in 2^{j_w}B$ and thus $|\rho^{-1}(z)| \leq \frac{1}{1-|\nabla\tau|_\infty}2^{j_w} \leq 22^{j_w}$ ($\tau(0) = 0$ by that τ is odd, and then $|\tau(\rho^{-1}(z))| \leq |\nabla\tau|_\infty|\rho^{-1}(z)|$), the r.h.s of (A.6) $\leq 2(2^{j_w} + 2^{j_f})|\nabla\tau|_\infty\|\nabla x\|_1$, which proves (A.5). □

Proof of Theorem 1. We need a slightly generalized form of Lemma 1, which inserts multiple plain convolutional layers between $*w$ and $*f$, presented in Lemma 2.

Under the setting of the theorem, in the generative CNNs,

$$X_s = \sigma(\cdots\sigma(h * w_s^{(-L)} + b_s^{(-L)})\cdots * w_s^{(-1)} + b_s^{(-1)}) \tag{A.7}$$

$$X_t = \sigma(\cdots\sigma(h * w_t^{(-L)} + b_t^{(-L)})\cdots * w_t^{(-1)} + b_t^{(-1)}) \tag{A.8}$$

where $w_t^{(l)}$ and $b_t^{(l)}$ are defined by, $l = -L, \cdots, -1$,

$$w_t^{(l)} = D_l w_s^{(l)}, \quad \tilde{x}_0^{(l)} * w_t^{(l)} + b_t^{(l)} = \tilde{x}_0^{(l)} * w_s^{(l)} + b_s^{(l)}. \tag{A.9}$$

The notation $\tilde{x}^{(l)}$ stands for the l-th layer output in the target net from the input in the bottom ($(-L)$-th) layer as $\tilde{x}^{(-L)} = h$, $\tilde{x}^{(0)} = X_t$, and $\tilde{x}_0^{(l)}$ for that from zero input in the bottom. In the feature CNNs, the L-th layer outputs are

$$F_s = \sigma(\cdots\sigma(X_s * w_s^{(1)} + b_s^{(1)})\cdots * w_s^{(L)} + b_s^{(L)}) \tag{A.10}$$

$$F_t = \sigma(\cdots\sigma(X_t * w_t^{(1)} + b_t^{(1)})\cdots * w_t^{(L)} + b_t^{(L)}) \tag{A.11}$$

where for $l = 1, \cdots, L$,

$$w_t^{(l)} = D_l w_s^{(l)}, \quad b_t^{(l)} = b_s^{(l)}.$$

The proof is by applying Lemma 2 recursively to the pair of layers indexed by l and $-l$, from $l = 1$ to L. Denote $w_s^{(l)}$ by $w^{(l)}$, then $w_t^{(l)} = D_l w^{(l)}$, where $D_{-l} = D_l = D_{\tau_l}$, $l = 1, \cdots, L$. We also denote $b_s^{(l)}$ by $b^{(l)}$ and keep notation $b_t^{(l)}$ for negative l.

First, $l = 1$, in the target net,

$$\tilde{x}^{(1)} := \sigma(\sigma(\tilde{x}^{(-1)} * D_1 w^{(-1)} + b_t^{(-1)}) * D_1 w^{(1)} + b^{(1)})$$

Use the centering $\tilde{x}_c^{(-1)} := \tilde{x}^{(-1)} - \tilde{x}_0^{(-1)}$, it can be written as

$$\tilde{x}^{(1)} = \sigma(\sigma(\tilde{x}_c^{(-1)} * D_1 w^{(-1)} + \tilde{x}_0^{(-1)} * D_1 w^{(-1)} + b_t^{(-1)}) * D_1 w^{(1)} + b^{(1)}) \tag{A.12}$$

$$= \sigma(\sigma(\tilde{x}_c^{(-1)} * D_1 w^{(-1)} + (\tilde{x}_0^{(-1)} * w^{(-1)} + b^{(-1)})) * D_1 w^{(1)} + b^{(1)}) \text{ (by (A.9))} \tag{A.13}$$

Applying Lemma 2 (or Lemma 1 for this case), taking $\tilde{x}_0^{(-1)} * w^{(-1)} + b^{(-1)}$ as the effective "b", we have that (using the non-expansiveness of σ to take r outside the last σ)

$$\tilde{x}^{(1)} = \sigma(\sigma(\tilde{x}_c^{(-1)} * w^{(-1)} + \tilde{x}_0^{(-1)} * w^{(-1)} + b^{(-1)}) * w^{(1)} + b^{(1)}) + r^{(1)} \tag{A.14}$$

$$= \sigma(\sigma(\tilde{x}^{(-1)} * w^{(-1)} + b^{(-1)}) * w^{(1)} + b^{(1)}) + r^{(1)} \tag{A.15}$$

$$:= \hat{x}^{(1)} + r^{(1)} \tag{A.16}$$

where, since $w^{(-1)}$, $w^{(1)}$ are supported on $2^{j_1}B$,

$$\|r^{(1)}\|_1 \leq 4\varepsilon\left\{2^{j_1}\|\nabla\tilde{x}_c^{(-1)}\|_1 + 2\|\tilde{x}_c^{(-1)}\|_1\right\}. \tag{A.17}$$

Next,

$$\tilde{x}^{(2)} := \sigma(\tilde{x}^{(1)} * D_2 w^{(2)} + b^{(2)}) \tag{A.18}$$

$$= \sigma((\hat{x}^{(1)} + r^{(1)}) * D_2 w^{(2)} + b^{(2)}) \text{ (by (A.16))} \tag{A.19}$$

$$= \sigma(\hat{x}^{(1)} * D_2 w^{(2)} + b^{(2)}) + r^{(1)'} \tag{A.20}$$

where $\|r^{(1)'}\|_1 \le \|r^{(1)}\|_1$ and observe the same bound as (A.17), since neither $*w_t^{(2)}$ (Lemma 3(i)) nor applying σ with bias expands the 1-norm. Using the brief notation σ_l to denote the non-linear mapping with biases $b^{(l)}$, consider

$$\begin{aligned}\sigma_2(\hat{x}^{(1)} * D_2 w^{(2)}) &= \sigma_2(\sigma_1(\sigma_{-1}(\tilde{x}^{(-1)} * w^{(-1)}) * w^{(1)}) * D_2 w^{(2)}) \\ &= \sigma_2(\sigma_1(\sigma_{-1}(\sigma(\tilde{x}^{(-2)} * D_2 w^{(-2)} + b_t^{(-2)}) * w^{(-1)}) * w^{(1)}) * D_2 w^{(2)}) \\ &= \sigma_2(\sigma_1(\sigma_{-1}(\sigma(\tilde{x}_c^{(-2)} * D_2 w^{(-2)} + \tilde{x}_0^{(-2)} * w^{(-2)} + b^{(-2)}) \\ &\qquad * w^{(-1)}) * w^{(1)}) * D_2 w^{(2)}), \quad \text{(by (A.9))}\end{aligned}$$

by Lemma 2, it equals (using the non-expansiveness of σ_2 to take $r^{(2)}$ outside)

$$\begin{aligned}&\sigma_2(\sigma_1(\sigma_{-1}(\sigma(\tilde{x}_c^{(-2)} * w^{(-2)} + \tilde{x}_0^{(-2)} * w^{(-2)} + b^{(-2)}) * w^{(-1)}) * w^{(1)}) * w^{(2)}) + r^{(2)} \\ &= \sigma_2(\sigma_1(\sigma_{-1}(\sigma(\tilde{x}^{(-2)} * w^{(-2)} + b^{(-2)}) * w^{(-1)}) * w^{(1)}) * w^{(2)}) + r^{(2)} \\ &:= \hat{x}^{(2)} + r^{(2)}\end{aligned}$$

where

$$\|r^{(2)}\|_1 \le 4\varepsilon \left\{ 2^{j_2} \|\nabla \tilde{x}_c^{(-2)}\|_1 + 2\|\tilde{x}_c^{(-2)}\|_1 \right\}. \tag{A.21}$$

Inserting back to (A.20),

$$\tilde{x}^{(2)} = \hat{x}^{(2)} + r^{(1)'} + r^{(2)}$$

thus $\|\tilde{x}^{(2)} - \hat{x}^{(2)}\|_1$ is bounded by the sum of (A.17) and (A.21).

Continue the process, $\hat{x}^{(l)}$ denotes the l-th layer output in the source CNN (after l times correction in the target CNN) by feeding $\tilde{x}^{(-l-1)}$ from the $(-l)$-th layer, where $\tilde{x}^{(-l-1)}$ is the output in the (un-corrected) generative target CNN after the first $(L-l)$ layers. By that $\tilde{x}^{(-L)} = x^{(-L)} = h$, and that $F_t = \tilde{x}^{(L)}$, $F_s = x^{(L)}$, repeating the argument L times gives that

$$\|F_s - F_t\|_1 \le 4\varepsilon \sum_{l=1}^{L} (2^{j_l} \|\nabla \tilde{x}_c^{(-l)}\|_1 + 2\|\tilde{x}_c^{(-l)}\|_1),$$

and when $(I_d - \rho_l)$ are rigid motions, the 2nd term for each l vanishes.

We claim that

Claim 3. For $l = -L, \cdots, -1$, $\|\nabla \tilde{x}_c^{(l)}\|_1 \le \|\nabla h\|_1$, and $\|\tilde{x}_c^{(l)}\|_1 \le \|h\|_1$.

which suffices to prove the theorem.

Proof of Claim 3: No that in the bottom layer $\tilde{x}_c^{(-L)} = \tilde{x}^{(-L)} = h$. For $l = -L+1, \cdots, -1$,

$$\begin{aligned}\|\tilde{x}_c^{(l)}\|_1 &= \|\tilde{x}^{(l)} - \tilde{x}_0^{(l)}\|_1 \\ &= \|\sigma_l(\tilde{x}^{(l-1)} * w_t^{(l-1)}) - \sigma_l(\tilde{x}_0^{(l)} * w_t^{(l-1)})\|_1 \\ &\le \|\tilde{x}^{(l-1)} * w_t^{(l-1)} - \tilde{x}_0^{(l)} * w_t^{(l-1)}\|_1 \quad \text{(by that } \sigma_l \text{ non-expansive)} \\ &\le \|\tilde{x}^{(l-1)} - \tilde{x}_0^{(l)}\|_1 \quad \text{(by that } \|w_t^{(l-1)}\|_1 \le 1 \text{ and Lemma 3(i))} \\ &= \|\tilde{x}_c^{(l-1)}\|_1.\end{aligned}$$

Recursing the inequality gives that $\|\tilde{x}_c^{(l)}\|_1 \le \|h\|_1$. Similarly,

$$\begin{aligned}\|\nabla \tilde{x}_c^{(l)}\|_1 = \|\nabla \tilde{x}^{(l)}\|_1 &= \text{TV}[\sigma_l(\tilde{x}^{(l-1)} * w_t^{(l-1)})] \\ &\le \text{TV}[\tilde{x}^{(l-1)} * w_t^{(l-1)}] \quad \text{(by that } \sigma_l \text{ does not increase total variation)} \\ &= \|\nabla(\tilde{x}^{(l-1)} * w_t^{(l-1)})\|_1 \\ &\le \|\nabla \tilde{x}^{(l-1)}\|_1 = \|\nabla \tilde{x}_c^{(l-1)}\|_1, \quad \text{(by that } \|w_t^{(l-1)}\|_1 \le 1 \text{ and Lemma 3(ii))}\end{aligned}$$

and thus $\|\nabla \tilde{x}_c^{(l)}\|_1 \le \|\nabla h\|_1$. This proves Claim 3. □

Lemma 2. *Suppose filters w, $f_1, \cdots, f_m$, f satisfy that the 1-norm are all bounded by 1, and w and f are supported on $2^j B$. The sequence of σ_l, denoting non-linear function with bias, for $l = 0, \cdots, m$ are non-expansive. D_τ is a spatial transform where τ is odd and $|\nabla\tau|_\infty \leq \varepsilon < \frac{1}{5}$. Then*

$$\sigma_m(\cdots\sigma_1(\sigma_0(x * D_\tau w) * f_1)\cdots * f_m) * f$$

approximates

$$\sigma_m(\cdots\sigma_1(\sigma_0(x * w) * f_1)\cdots * f_m) * D_\tau^{-1} f$$

up to an error whose 1-norm is bounded by

$$4\varepsilon\left\{2^j\|\nabla x\|_1 + 2\|x\|_1\right\},$$

and the second term vanishes if $(I_d - \tau)$ is a rigid motion.

Proof of Lemma 2. The proof uses the same technique as in the proof of Lemma 1. Omitting subscript $\mathbb{R}^2$ in the integral, let

$$\begin{aligned} y_1(u) = \int \sigma_m(\int \cdots \sigma_1(\int \sigma_0(\int x(u + v_1 + \cdots + v_m + v - \rho^{-1}(z))w(z)|J\rho^{-1}|dz) \\ f(-v_1)dv_1)\cdots f_m(-v_m)dv_m)f(-v)dv, \\ \hat{y}_1(u) = \int \sigma_m(\int \cdots \sigma_1(\int \sigma_0(\int x(u + v_1 + \cdots + v_m + v - \rho^{-1}(z))w(z)dz) \\ f(-v_1)dv_1)\cdots f_m(-v_m)dv_m)f(-v)dv. \end{aligned}$$

By Fact 1, that σ_j are all non-expansive and that the 1-norm of all the filters are bounded by 1,

$$\int |y_1(u) - \hat{y}_1(u)|du \leq 4\varepsilon\|x\|_1.$$

Also,

$$\begin{aligned} y_2(u) = \int \sigma_m(\int \cdots \sigma_1(\int \sigma_0(\int x(u + v_1 + \cdots + v_m + \rho(v) - z)w(z)dz) \\ f(-v_1)dv_1)\cdots f_m(-v_m)dv_m)f(-v)|J\rho|dv, \\ \hat{y}_2(u) = \int \sigma_m(\int \cdots \sigma_1(\int \sigma_0(\int x(u + v_1 + \cdots + v_m + \rho(v) - z)w(z)dz) \\ f(-v_1)dv_1)\cdots f_m(-v_m)dv_m)f(-v)dv. \end{aligned}$$

Similarly,

$$\int |y_2(u) - \hat{y}_2(u)|du \leq 4\varepsilon\|x\|_1.$$

Same as before, with ρ being a rigid motion, $\|y_1 - \hat{y}_1\|$ and $\|y_2 - \hat{y}_2\|$ are both zero.

It remains to bound $\|\hat{y}_1 - \hat{y}_2\|_1$. Observe that

$$\begin{aligned} \int |\hat{y}_1(u) - \hat{y}_2(u)|du \leq \int \cdots \int dv|f(-v)|dv_m|f(-v_m)|\cdots dv_1|f(-v_1)|dz|w(z)| \\ \int du|x(u + v_1 + \cdots + v_m + v - \rho^{-1}(z)) - x(u + v_1 + \cdots + v_m + \rho(v) - z)|, \quad \text{(A.22)} \end{aligned}$$

and similarly as in proving Lemma 1, one can verify that for any fixed $v_1, \cdots, v_m, v, z$,

$$\begin{aligned} \int |x(u + v_1 + \cdots + v_m + v - \rho^{-1}(z)) - x(u + v_1 + \cdots + v_m + \rho(v) - z)|du \\ \leq \|\nabla x\|_1|\nabla\tau|_\infty|v - \rho^{-1}(z)| \leq \varepsilon 2(2^j + 2^j)\|\nabla x\|_1. \end{aligned}$$

Inserting back to (A.22), and again by that the 1-norm of all the filters are bounded by 1, we have that $\|\hat{y}_1 - \hat{y}_2\|_1 \leq 4\varepsilon 2^j\|\nabla x\|_1$. □

Lemma 3. *Let x and w be smooth and compactly supported on $\mathbb{R}^2$, then*

(i) $\|x * w\|_1 \leq \|x\|_1\|w\|_1$.

(ii) $\|\nabla(x * w)\|_1 \leq \|\nabla x\|_1\|w\|_1$.

Proof of Lemma 3. For (i),

$$\|x * w\|_1 = \int_{\mathbb{R}^2} |\int_{\mathbb{R}^2} x(u-v)w(v)dv|du \leq \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} |x(u-v)||w(v)|dudv = \|x\|_1\|w\|_1.$$

For (ii),

$$\begin{aligned}
\|\nabla(x * w)\|_1 &= \int_{\mathbb{R}^2} |\nabla_u(\int_{\mathbb{R}^2} x(u-v)w(v)dv)|du \\
&= \int_{\mathbb{R}^2} |\int_{\mathbb{R}^2} \nabla_u x(u-v)w(v)dv|du \\
&\leq \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} |\nabla_u x(u-v)||w(v)|dudv \\
&= \|\nabla x\|_1\|w\|_1.
\end{aligned}$$

□