A Proof

Proof of Theorem 2. To see $\mathcal{H}^\circ$ is still an RKHS under the new norm, note that a Hilbert space is an RKHS if point evaluation functionals are bounded, i.e., there exists a $C > 0$ such that for any $x \in \mathcal{X}$ and f in the space, it holds that $|f(x)| \le C \|f\|$. Since $\mathcal{H}$ is an RKHS, so $|f(x)| \le C \|f\|_{\mathcal{H}}$. Since $\|f\|_{\mathcal{H}^\circ} \ge \|f\|_{\mathcal{H}}$, it follows trivially that $|f(x)| \le C \|f\|_{\mathcal{H}^\circ}$. So $\mathcal{H}^\circ$ is an RKHS.

Clearly $k^\circ(x, \cdot) = k(x, \cdot) - z(x)^\top (I + K_z)^{-1} z(\cdot)$ is in $\mathcal{H}^\circ$ as it linearly combines $k(x, \cdot)$ and $\{z_i\}$ which are in $\mathcal{H}$, and $\mathcal{H}^\circ$ consists of the same set of functions as $\mathcal{H}$. So it suffices to show $k^\circ(x, \cdot)$ is a representer of point evaluation at x in $\mathcal{H}^\circ$.

For any $f \in \mathcal{H}^\circ$ (or equivalently $f \in \mathcal{H}$), denote $z_f := (\langle z_1, f\rangle_{\mathcal{H}}, \ldots, \langle z_m, f\rangle_{\mathcal{H}})^\top$. It follows that for all $x \in \mathcal{X}$,

$$
\begin{aligned}
\langle k^\circ(x,\cdot), f\rangle_{\mathcal{H}^\circ} &= \langle k^\circ(x,\cdot), f\rangle_{\mathcal{H}} + \sum\nolimits_i \langle z_i, f\rangle_{\mathcal{H}} \langle z_i, k^\circ(x,\cdot)\rangle_{\mathcal{H}} \quad \text{(by the definition of } \langle\cdot,\cdot\rangle_{\mathcal{H}^\circ}) \\
&= \left\langle k(x,\cdot) - z(x)^\top (I+K_z)^{-1} z(\cdot), f\right\rangle_{\mathcal{H}} + \sum\nolimits_i \langle z_i, f\rangle_{\mathcal{H}} \left\langle z_i, k(x,\cdot) - z(x)^\top (I+K_z)^{-1} z(\cdot)\right\rangle_{\mathcal{H}} \\
&= f(x) - z(x)^\top (I+K_z)^\top z_f + z(x)^\top z_f - z(x)^\top (I+K_z)^\top K_z z_f = f(x),
\end{aligned}
$$

where the last equality follows from the simple fact that $-(I+K_z)^{-1} + I - (I+K_z)^{-1} K_z = \mathbf{0}$ (to see it, left multiply both sides by the invertible matrix $I + K_z$). So k° is the reproducing kernel of $\mathcal{H}^\circ$. □

Proof of Theorem 3. Let $v := \sum_i \alpha_i z(x_i) - \sum_j \beta_j z(y_j)$. Then

$$
\begin{aligned}
0 &\le \left\| \sum_i \alpha_i \varphi^\circ(x_i) - \sum_j \beta_j \varphi^\circ(y_j) \right\|^2_{\mathcal{H}^\circ} && (19) \\
&= \sum_{i,i'} \alpha_i \alpha_{i'} k^\circ(x_i, x_{i'}) + \sum_{j,j'} \beta_j \beta_{j'} k^\circ(y_j, y_{j'}) - 2 \sum_{i,j} \alpha_i \beta_j k^\circ(x_i, y_j) && (20) \\
\text{by (3)} \quad &= \sum_{i,i'} \alpha_i \alpha_{i'} (k(x_i, x_{i'}) - z(x_i)^\top M z(x_{i'})) \quad \text{where} \quad M := (I+K_z)^{-1} && (21) \\
&\quad + \sum_{j,j'} \beta_j \beta_{j'} (k(y_j, y_{j'}) - z(y_j)^\top M z(y_{j'})) - 2 \sum_{i,j} \alpha_i \beta_j (k(x_i, y_j) - z(x_i)^\top M z(y_j)) && (22) \\
&= \left\| \sum_i \alpha_i \varphi(x_i) - \sum_j \beta_j \varphi(y_j) \right\|^2_{\mathcal{H}} - v^\top M v = -v^\top M v \le 0, && (23)
\end{aligned}
$$

So we conclude $\left\| \sum_i \alpha_i \varphi^\circ(x_i) - \sum_j \beta_j \varphi^\circ(y_j) \right\|_{\mathcal{H}^\circ} = 0$, i.e., $\sum_i \alpha_i \varphi^\circ(x_i) = \sum_j \beta_j \varphi^\circ(y_j)$. □

Property 1. *The warping operator is non-expansive.*

Proof. For any $f = \sum_i \alpha_i \varphi(x_i)$, denote $z(x_i) = (z_1(x_i), \ldots, z_m(x_i))^\top$. Then

$$
\begin{aligned}
\|f^\circ\|^2_{\mathcal{H}^\circ} &= \left\| \sum_i \alpha_i \varphi^\circ(x_i) \right\|^2_{\mathcal{H}^\circ} = \sum_{ij} \alpha_i \alpha_j k^\circ(x_i, x_j) && (24) \\
&= \sum_{ij} \alpha_i \alpha_j \left(k(x_i, x_j) - z(x_i)^\top (I+K_z)^{-1} z(x_j) \right) && (25) \\
&= \|f\|^2_{\mathcal{H}} - \left(\sum_i \alpha_i z(x_i) \right)^\top (I+K_z)^{-1} \left(\sum_j \alpha_j z(x_j) \right) \le \|f\|^2_{\mathcal{H}}. && (26)
\end{aligned}
$$

So $\|f^\circ\|_{\mathcal{H}^\circ} \le \|f\|_{\mathcal{H}}$ as K_z is PSD. □

Proof of Lemma 1. The proof follows that of Proposition 4 in [29], but inserts $\|[W_{k-1}, L_\tau]\|$ as needed. Define $(MPAW)_{k:j} := M_k P_k A_{k-1} W_{k-1} M_{k-1} P_{k-1} A_{k-2} W_{k-2} \ldots M_j P_j A_{j-1} W_{j-1}$. Noting that $\|A_k\| \leq 1$, $\|P_k\| = 1$, W_k and M_k are non-expansive, we obtain

$$
\begin{aligned}
&\|\Psi_n(L_\tau x) - \Psi_n(x)\| && (27)\\
&= \|A_n(MPAW)_{n:2} M_1 P_1 A_0 L_\tau x - A_n(MPAW)_{n:2} M_1 P_1 A_0 x\| && (28)\\
&\leq \|A_n(MPAW)_{n:2} M_1 P_1 A_0 L_\tau x - A_n(MPAW)_{n:2} M_1 L_\tau P_1 A_0 x\| && (29)\\
&\quad + \|A_n(MPAW)_{n:2} M_1 L_\tau P_1 A_0 x - A_n(MPAW)_{n:2} M_1 P_1 A_0 x\| && (30)\\
&\overset{(a)}{\leq} \|[P_1 A_0, L_\tau]\| \, \|x\| + \|A_n(MPAW)_{n:2} L_\tau M_1 P_1 A_0 x - A_n(MPAW)_{n:2} M_1 P_1 A_0 x\| && (31)\\
&\overset{(b)}{\leq} \|[P_1 A_0, L_\tau]\| \, \|x\| + \|A_n(MPAW)_{n:3} M_2 P_2 A_1 W_1 L_\tau y_1 - A_n(MPAW)_{n:3} M_2 P_2 A_1 L_\tau W_1 y_1\| && (32)\\
&\quad + \|A_n(MPAW)_{n:3} M_2 P_2 A_1 L_\tau W_1 y_1 - A_n(MPAW)_{n:3} M_2 P_2 A_1 W_1 y_1\| && (33)\\
&\overset{(c)}{\leq} \|[P_1 A_0, L_\tau]\| \, \|x\| + \|[W_1, L_\tau]\| \, \|x\| + \|A_n(MPAW)_{n:3} M_2 P_2 A_1 L_\tau z_1 - A_n(MPAW)_{n:3} M_2 P_2 A_1 z_1\|. && (34)
\end{aligned}
$$

Here (a) is by $M_1 L_\tau = L_\tau M_1$, (b) is by defining $y_1 = M_1 P_1 A_0 x$, (c) is by defining $z_1 = W_1 y_1$. Noting that the last line is isomorphic to the first line and $\|z_1\| \leq \|x\|$, we can unfold this recursion and prove Lemma 1. $\square$

Proof of Theorem 4. We use Schur's test by reformulating $[W, L_\tau]$ as an integral operator and bounding its kernel [Lemma A.1, 29]. Letting $\xi = (I - \tau)^{-1}$ and noting the Jacobian in change of variable for integral, we have

$$
\begin{aligned}
[W, L_\tau] f(z) &= W L_\tau f(z) - L_\tau W f(z) && (35)\\
&= \int W(z, u) f(u - \tau(u)) \, \mathrm{d}u - \int W(z - \tau(z), u) f(u) \, \mathrm{d}u && (36)\\
&= \int W(z, \xi(s)) f(s) \left|\frac{\mathrm{d}u}{\mathrm{d}s}\right| \mathrm{d}s - \int W(z - \tau(z), s) f(s) \, \mathrm{d}s && (37)
\end{aligned}
$$

Noting that $\alpha := \left|\frac{\mathrm{d}u}{\mathrm{d}s}\right| = \det(I - \nabla\tau(u))^{-1}$, we derive the kernel

$$
\begin{aligned}
k(z, s) &= \alpha W(z, \xi(s)) - W(z - \tau(z), s) && (38)\\
&= \underbrace{(\alpha - 1) W(z, \xi(s))}_{=:A} + \underbrace{W(z, \xi(s)) - W(z, s)}_{=:B} + \underbrace{W(z, s) - W(z - \tau(z), s)}_{=:C}. && (39)
\end{aligned}
$$

Since $\det(I - \nabla\tau(u)) \geq (1 - \|\nabla\tau\|_\infty)^d \geq 1 - d\|\nabla\tau\|_\infty$, it follows that $\alpha \in [1, 1 + 2d\|\nabla\tau\|_\infty]$. We can then bound each term as

$$
\begin{aligned}
|A| &\leq 2d \|\nabla\tau\|_\infty |W(z, \xi(s))| \overset{(a)}{\leq} 2d \|\nabla\tau\|_\infty, && (40)\\
|B| &\leq L_w \|\xi(s) - s\| = L_w \|\tau(\xi(s))\| \leq L_w \|\tau\|_\infty, && (41)\\
|C| &\leq L_w \|\tau\|_\infty, && (42)
\end{aligned}
$$

where (a) is because W is non-expansive. As Ω is bounded, we can bound both $\int |k(z, s)| \, \mathrm{d}z$ and $\int |k(z, s)| \, \mathrm{d}s$ by $C_1 \|\nabla\tau\|_\infty + C_2 L_w \|\tau\|_\infty$, where C_1 and C_2 depend on Ω only. Then Schur's test directly implies (18). $\square$

B Derivative for End-to-end Training of Single Hidden Layer Network

Suppose we have invariance representers with finite approximation $Z = \{z_1, \ldots, z_m\}$ (we dropped the tilde to simplify notation as here we only deal with finite approximations). Let there be n_c classes, and the output layer weight be a matrix O with each column corresponding to a class. Let $\xi(x)$ be the FA of x using the Fourier samples $B := (\omega_1, \ldots, \omega_p)$. Then the end-to-end empirical risk minimization can be written as

$$
\min_{B, O} \mathop{\mathbb{E}}_{(x,l) \sim \tilde{p}} \left[L(O^\top (I + Z Z^\top)^{-1/2} \xi(x), l) \right], \tag{43}
$$

where $\tilde{p}$ is the empirical distribution over feature/label pair (x, l). Both $\xi(x)$ and Z depend on B.

Denote $f(B, O) = L(O^\top (I + ZZ^\top)^{-1/2} \xi(x), l)$. Then trivially

$$\nabla_O f(B, O) = (I + ZZ^\top)^{-1/2} \xi(x) \cdot r^\top, \tag{44}$$

where $r := \nabla L(O^\top (I + ZZ^\top)^{-1/2} \xi(x), l) \in \mathbb{R}^{n_c}$ and ∇L denotes the partial derivative of L with respect to its first argument.

To compute the derivative in B, we analyze the change of f when B is perturbed by ΔB with $\|\Delta B\| := \sum_i \|\omega_i\| \leq \epsilon$. Suppose ΔZ is the corresponding change of Z up to $o(\epsilon)$. Then letting $M^2 := I + ZZ^\top$ and $G := (\Delta Z) Z^\top + Z (\Delta Z)^\top$, we have

$$\begin{aligned}
(I + (Z + \Delta Z)(Z + \Delta Z)^\top)^{-1/2} &= (I + ZZ^\top + G + o(\epsilon))^{-1/2} && (45)\\
&= [M(I + M^{-1} G M^{-1} + o(\epsilon)) M]^{-1/2} && (46)\\
&= M^{-1/2} (I + M^{-1} G M^{-1} + o(\epsilon))^{-1/2} M^{-1/2} && (47)\\
&= M^{-1/2} (I - \tfrac{1}{2} M^{-1} G M^{-1} + o(\epsilon)) M^{-1/2} && (48)\\
&= M^{-1} - \tfrac{1}{2} M^{-\frac{3}{2}} G M^{-\frac{3}{2}} + o(\epsilon). && (49)
\end{aligned}$$

The change of $\xi(x)$ with respect to ΔB depends on the kernel. Let us use Gaussian kernel and

$$\xi_B(x) = \frac{1}{\sqrt{p}} \begin{pmatrix} \cos(\omega_1^\top x) \\ \sin(\omega_1^\top x) \\ \vdots \\ \cos(\omega_p^\top x) \\ \sin(\omega_p^\top x) \end{pmatrix} \quad \Rightarrow \quad \Delta \xi_B(x) = \frac{1}{\sqrt{p}} \begin{pmatrix} -\sin(\omega_1^\top x) \cdot x^\top \Delta\omega_1 \\ \cos(\omega_1^\top x) \cdot x^\top \Delta\omega_1 \\ \vdots \\ -\sin(\omega_p^\top x) \cdot x^\top \Delta\omega_p \\ \cos(\omega_p^\top x) \cdot x^\top \Delta\omega_p \end{pmatrix} = H \cdot (\Delta B)^\top \cdot x \tag{50}$$

$$\text{where} \quad H = \frac{1}{\sqrt{p}} \begin{pmatrix} -\sin(\omega_1^\top x) & 0 & 0 & \dots & 0 \\ \cos(\omega_1 \top x) & 0 & 0 & \dots & 0 \\ 0 & -\sin(\omega_2^\top x) & 0 & \dots & 0 \\ 0 & \cos(\omega_2^\top x) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \in \mathbb{R}^{2p \times p}. \tag{51}$$

Therefore

$$\begin{aligned}
f(B + \Delta B, g) - f(B, g) &= r^\top O^\top \Big[\Big(M^{-1} - \tfrac{1}{2} M^{-\frac{3}{2}} G M^{-\frac{3}{2}} \Big) \cdot (\xi_B(x) + H \cdot (\Delta B)^\top \cdot x) - M^{-1} \xi_B(x) + o(\epsilon) \Big] && (52)\\
&= r^\top O^\top \Big(M^{-1} H \cdot (\Delta B)^\top \cdot x - \frac{1}{2} M^{-\frac{3}{2}} G M^{-\frac{3}{2}} \xi_B(x) + o(\epsilon) \Big). && (53)
\end{aligned}$$

The contribution of the first part to the gradient in B is easy to derive because

$$r^\top O^\top M^{-1} H \cdot (\Delta B)^\top \cdot x = \left\langle \Delta B, x r^\top O^\top M^{-1} H \right\rangle. \tag{54}$$

So the contribution to the gradient aggregated over the entire dataset is $\sum_x x r_x^\top O^\top M^{-1} H_x$ (note r and H depend on x). M (independent of x) can be computed by first finding the left singular vectors of Z (or a few leading ones), and then adjusting their corresponding singular values to give the eigen-decomposition of M.

The second term in (53) can be expanded as

$$\begin{aligned}
-\frac{1}{2} r^\top O^\top M^{-\frac{3}{2}} G M^{-\frac{3}{2}} \xi_B(x) &= -\frac{1}{2} r^\top O^\top M^{-\frac{3}{2}} ((\Delta Z) Z^\top + Z (\Delta Z)^\top) M^{-\frac{3}{2}} \xi_B(x) && (55)\\
&= -\frac{1}{2} \left(a^\top (\Delta Z) b + c^\top (\Delta Z) d \right), && (56)
\end{aligned}$$

$$\text{where} \quad a = M^{-\frac{3}{2}} O r, \qquad b = Z^\top M^{-\frac{3}{2}} \xi_B(x), \qquad c = M^{-\frac{3}{2}} \xi_B(x), \qquad d = Z^\top M^{-\frac{3}{2}} O r. \tag{57}$$

Here all a, b, c, d depend on x. So if we can compute the contribution of gradient from $a^\top (\Delta Z) b$, then that from $c^\top (\Delta Z) d$ can be computed in exactly the same way. To proceed, we now need to instantiate the invariances z_i.

Suppose z_i models the gradient at y_i in the direction of v_i. Then by (7),

$$Z = \frac{1}{\sqrt{p}} \begin{pmatrix} -(\omega_1^\top v_1)\sin(\omega_1^\top y_1) & \dots & -(\omega_1^\top v_m)\sin(\omega_1^\top y_m) \\ (\omega_1^\top v_1)\cos(\omega_1^\top y_1) & \dots & (\omega_1^\top v_m)\cos(\omega_1^\top y_m) \\ \vdots & \vdots & \vdots \\ -(\omega_p^\top v_1)\sin(\omega_p^\top y_1) & \dots & -(\omega_p^\top v_m)\sin(\omega_p^\top y_m) \\ (\omega_p^\top v_1)\cos(\omega_p^\top y_1) & \dots & (\omega_p^\top v_m)\cos(\omega_p^\top y_m) \end{pmatrix} \tag{58}$$

$$\Rightarrow \Delta Z = \frac{1}{\sqrt{p}} \begin{pmatrix} \alpha_{11}^\top \Delta\omega_1 & \dots & \alpha_{1m}^\top \Delta\omega_1 \\ \beta_{11}^\top \Delta\omega_1 & \dots & \beta_{1m}^\top \Delta\omega_1 \\ \vdots & \vdots & \vdots \\ \alpha_{p1}^\top \Delta\omega_p & \dots & \alpha_{pm}^\top \Delta\omega_p \\ \beta_{p1}^\top \Delta\omega_p & \dots & \beta_{pm}^\top \Delta\omega_p \end{pmatrix}, \text{ where } \begin{cases} \alpha_{ij} = -v_j \sin(\omega_i^\top y_j) - y_j(\omega_i^\top v_j)\cos(\omega_i^\top y_j) \\ \beta_{ij} = v_j \cos(\omega_i^\top y_j) - y_j(\omega_i^\top v_j)\sin(\omega_i^\top y_j) \end{cases}. \tag{59}$$

Denote $a = (a_1^+, a_1^-, \dots, a_p^+, a_p^-)^\top$. Then we can collect the terms in $a^\top(\Delta Z)b$ that involve $\Delta\omega_i$:

$$\frac{1}{\sqrt{p}} \left\langle \Delta\omega_i, a_i^+ \sum_{j=1}^m \alpha_{ij} b_j + a_i^- \sum_{j=1}^m \beta_{ij} b_j \right\rangle = \frac{1}{\sqrt{p}} \left\langle \Delta\omega_i, \sum_{j=1}^m p_{ij} v_j + \sum_{j=1}^m q_{ij} y_j \right\rangle, \tag{60}$$

$$\text{where} \quad p_{ij} = b_j(-a_i^+ \sin(\omega_i^\top y_j) + a_i^- \cos(\omega_i^\top y_j)) \tag{61}$$

$$q_{ij} = -b_j[a_i^+(\omega_i^\top v_j)\cos(\omega_i^\top y_j) + a_i^-(\omega_i^\top v_j)\sin(\omega_i^\top y_j)]. \tag{62}$$

So the gradient in B can be compactly written as $-\frac{1}{2\sqrt{p}}(VP^\top + YQ^\top)$, where $V = (v_1, \dots, v_m)$ and $Y = (y_1, \dots, y_m)$. Furthermore, incorporating the contribution from $c^\top(\Delta Z)d$, we can augment P and Q into:

$$p_{ij} = -\sin(\omega_i^\top y_j)(a_i^+ b_j + c_i^+ d_j) + \cos(\omega_i^\top y_j)(a_i^- b_j + c_i^- d_j) \tag{63}$$

$$q_{ij} = -(\omega_i^\top v_j)\cos(\omega_i^\top y_j)(a_i^+ b_j + c_i^+ d_j) - (\omega_i^\top v_j)\sin(\omega_i^\top y_j)(a_i^- b_j + c_i^- d_j). \tag{64}$$

So finally, the gradient in B can be computed by $-\frac{1}{2\sqrt{p}}(VP^\top + YQ^\top)$. The procedure is

1. Compute all $\omega_i^\top y_j$, followed by their sin and cos. Denote the results by matrices T (for products), S (for sine), and C (for cosine), respectively, all sized p-by-m. Also compute $\omega_i^\top v_j$ as a matrix $R \in \mathbb{R}^{p\times m}$. These cost $O(pmd)$ where d is the dimensionality of the input x.
2. Compute P and Q by

$$P = -S \circ (a^+ b^\top + c^+ d^\top) + C \circ (a^- b^\top + c^- d^\top) \tag{65}$$

$$Q = -R \circ [C \circ (a^+ b^\top + c^+ d^\top) + S \circ (a^- b^\top + c^- d^\top)], \tag{66}$$

   where $\circ$ is the Hadamard product. The total cost is $O(pm)$.
3. Compute $-\frac{1}{2\sqrt{p}}(VP^\top + YQ^\top)$, which costs $O(pmd)$.

If we naively perform this repeatedly for each of the l training examples, the total cost will be $O(pmdl)$. Fortunately this can be reduced to $O(pm(d+l))$ because different training examples only differ in a, b, c, d vectors, while T, S, or C are shared. So overall we can replace step 2 by

$$P = -S \circ F^+ + C \circ F^- \qquad \text{and} \qquad Q = -R \circ \left[C \circ F^+ + S \circ F^-\right], \tag{67}$$

$$\text{where} \quad F^+ = \sum_x a_x^+ b_x^\top + \sum_x c_x^+ d_x^\top, \qquad F^- = \sum_x a_x^- b_x^\top + \sum_x c_x^- d_x^\top. \tag{68}$$

This costs $O(pml)$. Of course it still costs to compute a, b, c, d for all x too. Once we assemble $a_x^+, a_x^-, b_x, c_x^+, c_x^-, d_x$ into matrices A^+, A^-, E (unfortunately the symbol B has been taken), C^+, C^-, D by columns, we get

$$F^+ = A^+ E^\top + C^+ D^\top, \qquad F^- = A^- E^\top + C^- D^\top, \tag{69}$$

again as efficient matrix-matrix multiplications.

C Connection with Convolutional Neural Networks

As demonstrated by [29], CKNs contain a set of convolutional neural networks (CNNs) with smooth and homogeneous activations. We now show that such a relationship is retained when kernel warping is introduced, and the new RKHS norm of the overall function allows CNNs to favor invariance-respecting configurations.

Consider a CNN function f_σ that is defined recursively through the layers. The input image $z_0 = x_0$ is in $L^2(\Omega, \mathbb{R}^{p_0})$ (i.e., p_0 channels). The image z_k at layer k lies in $L^2(\Omega, \mathbb{R}^{p_k})$, constructed from the previous z_{k-1} using convolution and pooling. In particular, it employs p_k filters $\{w_k^i\}_{i=1}^{p_k}$ where each $w_k^i = \{w_k^{ij}\}_{j=1}^{p_{k-1}} \in L^2(S_k, \mathbb{R}^{p_{k-1}})$. Then the convolution and activation yield $\tilde{z}_k^i(u) = n_k(u)\sigma(\langle w_k^i, P_k z_{k-1}(u)\rangle / n_k(u))$ for channel $i \in [p_k] := \{1, 2, \ldots, p_k\}$ and $u \in \Omega$, where σ is a smooth function, and $n_k(u) = \|P_k z_{k-1}(u)\|$. Finally the k-th layer image is obtained by pooling, with $z_k = A_k \tilde{z}_k$. A linear fully connected output/prediction layer is applied to the last layer n by $f_\sigma(x_0) = \langle w_{n+1}, z_n\rangle$.

[29] showed that under smoothness conditions of σ, f_σ with any value of filters $\{w_k^i\}$ can be reconstructed by a CKN with carefully engineered functions lying in the intermediate RKHS $\mathcal{H}_k$. Specifically, the first layer adopts $f_1^i \in \mathcal{H}_1$ and $g_1^i \in \mathcal{P}_1$ for $i \in [p_1]$ such that

$$g_1^i = w_1^i \in L^2(S_1, \mathbb{R}^{p_0}) = L^2(S_1, \mathcal{H}_0) = \mathcal{P}_1, \qquad f_1^i(z) = \|z\|\, \sigma(\langle g_1^i, z\rangle \,/\, \|z\|) \quad \text{for } z \in \mathcal{P}_1.$$

Based on layer $k-1$, the forward function $f_k^i \in \mathcal{H}_k$ and $g_k^i \in \mathcal{P}_k$ for channel $i \in [p_k]$ at layer k are

$$g_k^i(v) = \sum_{j=1}^{p_{k-1}} w_k^{ij}(v) f_{k-1}^j \text{ for } v \in S_k, \quad f_k^i(z) = \|z\|\, \sigma(\langle g_k^i, z\rangle \,/\, \|z\|) \;\text{ for } z \in \mathcal{P}_k.$$

And the linear output layer sets $g_\sigma(u) = \sum_{j=1}^{p_n} w_{n+1}^j(u) f_n^j$ for all $u \in \Omega$, so that $f : x_0 \mapsto \langle g_\sigma, x_n\rangle$ exactly recovers f_σ as shown by [29].

Effect of warping in CKN. Since warping does not change the set of functions in the RKHS at each layer, the CKN constructed above is obviously retained in our new space of CKNs. However, interesting changes occur to the RKHS norm. As shown by [29], $\|f_1^i\|^2 \le C_\sigma^2(\|w_1^i\|_2^2)$ where $\|w_1^i\|_2^2 = \int_{S_1} \|w_1^i(v)\|^2 \, \mathrm{d}\nu_1(v)$, and $C_\sigma : \mathbb{R} \to \mathbb{R}$ is an increasing function depending only on σ and the kernels. Recursively, without kernel warping, we have

$$\|f_k^i\|_{\mathcal{H}_k}^2 \le C_\sigma^2(\|g_k^i\|_{\mathcal{H}_{k-1}}^2), \qquad \|g_k^i\|_{\mathcal{H}_{k-1}}^2 \le p_{k-1} \sum_{j=1}^{p_{k-1}} \|w_k^{ij}\|_2^2 \cdot \|f_{k-1}^j\|_{\mathcal{H}_{k-1}}^2. \tag{70}$$

So the overall the recursion on $\|f_k^i\|_{\mathcal{H}_k}^2$ writes

$$\|f_k^i\|_{\mathcal{H}_k}^2 \le C_\sigma^2 \left(p_{k-1} \sum_{j=1}^{p_{k-1}} \|w_k^{ij}\|_2^2 \cdot \|f_{k-1}^j\|_{\mathcal{H}_{k-1}}^2 \right). \tag{71}$$

And the final prediction $f \in L^2(\Omega, \mathcal{H}_n)$ can be bounded by

$$\|f\|_{\mathcal{H}_n}^2 \le p_n \sum_{j=1}^{p_n} \left(\int_\Omega \left| w_{n+1}^j(u) \right|^2 \mathrm{d}u \right) \|f_n^j\|_{\mathcal{H}_n}^2. \tag{72}$$

With the same functions f_k^i and g_k^i, we note that the RKHS norm of f can only increase because $\|f\|_{\mathcal{H}^\circ}^2 = \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^m \langle z_i, f\rangle_{\mathcal{H}}^2$, if we warp the kernel and RKHS of the last CKN layer (layer n, before the output layer). We note in passing that this argument can be applied only to the last layer because warping is applied to images rather than patches, while the recursion in (70) and (71) works only on patches. We leave it as future work to show that warping an image at intermediate layers will also keep or increase the RKHS norm on all patches.