

DispVoxNets: Non-Rigid Point Set Alignment with Supervised Learning Proxies*

—Supplementary Material—

Soshi Shimada^{1,2} Vladislav Golyanik³ Edgar Tretschk³ Didier Stricker^{1,2} Christian Theobalt³

¹University of Kaiserslautern

²DFKI

³MPI for Informatics, SIC

In this supplement, we provide details on the interpolation of the coarse displacement field (Sec. A) and report training statistics (Sec. B). We show more qualitative comparisons (Sec. C) as well as graphs for further cases with uniform noise (Sec. D).

A. Interpolation of the 3D Displacement Field

Due to the limited resolution of the voxel grid, we apply trilinear interpolation to obtain displacements for every template point at sub-voxel precision. Note that in DE stage, interpolation is applied only in the forward pass. In the refinement stage, it is applied in the forward pass, and the computed trilinear weights are used during backpropagation to weight the gradients.

Suppose $\bar{\mathbf{D}}: \mathbb{Z}^3 \rightarrow \mathbb{R}^3$ is the initial regressed 3D displacement field on a regular lattice induced by the voxel grid. Suppose the template point of interest after the DE stage $\mathbf{y}_j^* = (x_j, y_j, z_j)$, $j \in \{1, \dots, M\}$, falls into a neighbourhood cube between eight displacement values of $\bar{\mathbf{D}}$. We denote these boundary displacements compactly by $\mathbf{d} = \{\mathbf{d}_{abc}\}$, $a, b, c \in \{0, 1\}$ on a unit cube¹ in a local coordinate system, see Fig. I for a schematic visualisation. In the refinement stage, we store for every \mathbf{y}_j^* the index of the voxel it belongs to, the indexes of the eight nearest displacements as well as the corresponding trilinear interpolation weights $\mathbf{w} \in \mathbb{R}^8$ in the point affinity table. The latter is then used in the backward pass of the refinement stage.

Let $x_{\max}, y_{\max}, z_{\max}$ and $x_{\min}, y_{\min}, z_{\min}$ be the maximum and minimum x -, y - and z -values among the eight nearest lattice point coordinates, respectively. To convert \mathbf{y}_j^* from the coordinate system of the lattice to the local coordinate system, we calculate normalised distances l_x, l_y and l_z :

$$l_x = \frac{x_j - x_{\min}}{x_{\max} - x_{\min}}, l_y = \frac{y_j - y_{\min}}{y_{\max} - y_{\min}} \text{ and } l_z = \frac{z_j - z_{\min}}{z_{\max} - z_{\min}}. \quad (1)$$

The individual displacement $\bar{\mathbf{v}}_j$ of \mathbf{y}_j^* is obtained by trilinear

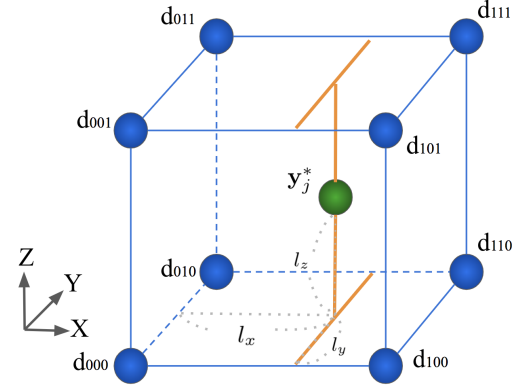


Figure I: Schematic visualisation of trilinear interpolation for a given \mathbf{y}_j^* .

	<i>thin plate</i>	<i>FLAME</i>	<i>DFAUST</i>	<i>cloth</i>
DE stage	530k	400k	715k	500k
refinement	14k	20k	24k	12k

Table I: Number of training iterations for DispVoxNets in the DE and refinement stages for the tested datasets.

interpolation of the eight nearest displacements, *i.e.*, as an inner product of \mathbf{w} and x -, y - and z -components of \mathbf{d} :

$$\bar{\mathbf{v}}_{j,x} = \mathbf{w}^\top \mathbf{d}_x = \begin{bmatrix} (1-l_x)(1-l_y)(1-l_z) \\ (1-l_x)(1-l_y)l_z \\ (1-l_x)l_y(1-l_z) \\ l_x(1-l_y)(1-l_z) \\ (1-l_x)l_y l_z \\ l_x l_y (1-l_z) \\ l_x(1-l_y)l_z \\ l_x l_y l_z \end{bmatrix}^\top \begin{bmatrix} \mathbf{d}_{000,x} \\ \mathbf{d}_{001,x} \\ \mathbf{d}_{010,x} \\ \mathbf{d}_{100,x} \\ \mathbf{d}_{011,x} \\ \mathbf{d}_{110,x} \\ \mathbf{d}_{101,x} \\ \mathbf{d}_{111,x} \end{bmatrix}, \quad (2)$$

$$\bar{\mathbf{v}}_{j,y} = \mathbf{w}^\top \mathbf{d}_y = \begin{bmatrix} (1-l_x)(1-l_y)(1-l_z) \\ (1-l_x)(1-l_y)l_z \\ (1-l_x)l_y(1-l_z) \\ l_x(1-l_y)(1-l_z) \\ (1-l_x)l_y l_z \\ l_x l_y (1-l_z) \\ l_x(1-l_y)l_z \\ l_x l_y l_z \end{bmatrix}^\top \begin{bmatrix} \mathbf{d}_{000,y} \\ \mathbf{d}_{001,y} \\ \mathbf{d}_{010,y} \\ \mathbf{d}_{100,y} \\ \mathbf{d}_{011,y} \\ \mathbf{d}_{110,y} \\ \mathbf{d}_{101,y} \\ \mathbf{d}_{111,y} \end{bmatrix}, \quad (3)$$

*supported by the ERC Consolidator Grant 4DReply (770784) and the BMBF project VIDETE (01IW18002).

¹ \mathbf{d}_{abc} is a shorthand notation for the displacement at point (x, y, z) in the local coordinate system, *i.e.*, at $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, *etc.*

and

$$\vec{v}_{j,z} = \mathbf{w}^\top \mathbf{d}_z = \begin{bmatrix} (1-l_x)(1-l_y)(1-l_z) \\ (1-l_x)(1-l_y)l_z \\ (1-l_x)l_y(1-l_z) \\ l_x(1-l_y)(1-l_z) \\ (1-l_x)l_y l_z \\ l_x l_y(1-l_z) \\ l_x(1-l_y)l_z \\ l_x l_y l_z \end{bmatrix}^\top \begin{bmatrix} \mathbf{d}_{000,z} \\ \mathbf{d}_{001,z} \\ \mathbf{d}_{010,z} \\ \mathbf{d}_{100,z} \\ \mathbf{d}_{011,z} \\ \mathbf{d}_{110,z} \\ \mathbf{d}_{101,z} \\ \mathbf{d}_{111,z} \end{bmatrix}. \quad (4)$$

Note that \mathbf{w} , l_x , l_y and l_z are shared across all dimensions.

B. Training Statistics

Table I shows the number of training iterations until convergence for each dataset. Since *DFAUST* contains relatively large displacements between point sets, it requires the highest number of iterations followed by *thin plate* and *cloth*. On the contrary, *FLAME* contains only small displacements, and the network requires fewer parameter updates to converge compared to other datasets.

C. Qualitative Analysis and Observations

In this section, we provide additional qualitative results. In Fig. II, we show selected registrations by our approach and other tested methods (NR-ICP [5], CPD/CPD with FGT [10], and GMR [7]) on the tested datasets (*thin plate* [6], *FLAME* [8], *DFAUST* [4] and *cloth* [2]).

On the *thin plate* — due to the rather simple object structure — all approaches except NR-ICP align the point sets reasonably accurate. CPD and DispVoxNets produce qualitatively similar results in the shown example. All methods show similar qualitative accuracy on the *cloth* dataset, while differences are noticeable in the corners and areas with large wrinkles. At the same time, only our approach simultaneously captures both small and large wrinkles. Thus, many fine foldings present in the reference surface are not well recognisable in the aligned templates in the case of NR-ICP, CPD/CPD with FGT and GMR. All in all, results of these methods appear to be oversmoothed.

In the absence of large displacements between the point sets — which is the case with *FLAME* dataset — model-based approaches CPD and GMR regress the displacements most accurately. The result of DispVoxNets is of comparable quality, though the deformed template is perceptually rougher and the points are arranged less regularly. This is due to the intermediate conversion steps from the point cloud representation to the voxel grid and back. We see that for small displacements, the limited resolution of the voxel grid is a more influential factor on the accuracy than the deformation prior learned from the data. With an increase of the voxel grid resolution, we expect our approach to come closer to CPD and GMR, up to the complete elimination

of the accuracy gap (this is the matter of future work; currently, our focus is handling of large deformations which is a more challenging problem).

Next, we see that model-based approaches with global regularisers often fail on the *FAUST* dataset, while the proposed approach demonstrates superior quantitative and visual accuracy. Even though the surface produced by DispVoxNets after the refinement stage can still seem coarse at some parts, the overall pose and shape are correctly and realistically inferred as we expect, despite substantial differences between the template and reference in the feet area (a subject standing on one foot and a subject standing on both feet respectively). Thus, model-based methods have difficulty in aligning the feet.

Overall, the qualitative results in Fig. II demonstrate the advantages of DispVoxNets for non-rigid point set alignment over classic, non-supervised learning-based approaches. Since our technique learns class-specific priors implicitly during training, it is successful in registering samples with large displacements and articulations.

D. Additional Experiments with Noisy Data

We present further experimental results with uniform noise in this section. Fig. III shows RMSE graphs for various combinations of uniform noise ratios in the reference and template for all four datasets (*thin plate* [6], *FLAME* [8], *DFAUST* [4] and *cloth* [2]).

For previous methods, we observe the tendency that adding uniform noise to both the template and the reference can result in a lower error than only adding it to one of them. It is reasonable to assume that two point sets certainly differ more if noise is added to only one of them. Thus, when inputs contain a similar amount of noise (we can say that the noise levels correlate), we observe the tendency that the alignment error becomes lower (e.g., see the seventh row, third column), i.e., some graphs roughly show a U-curve bottoming out at around 50% of the added noise in the template. We hypothesise that this is due to what we call the *mutual noise compensation effect*. Further study is required to clarify a more precise reason (it is possible that our observations are dataset-specific). Note that adding noise to both point sets is not a common evaluation setting. Usually, either template or reference is augmented with noise (cf. experimental sections in [7, 10, 9, 1]). With our experiment, we go beyond the prevalent evaluation methodology with noisy point sets.

On the one hand, CPD has the most stable error curve among the four model-based approaches, followed by NR-ICP and GMR. GMR shows higher errors when the noise is only in the template rather than in the reference, and CPD with FGT is the least stable as the noise ratio increases. Moreover, we observe that the relative performance of NR-ICP increases with the added noise. Thus, NR-ICP outper-


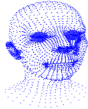

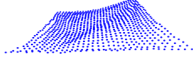







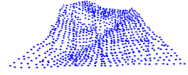
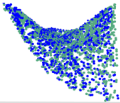
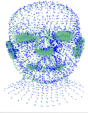

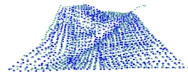



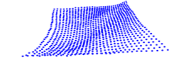
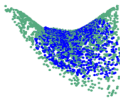
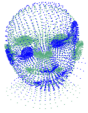

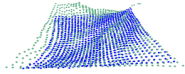

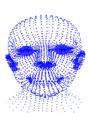

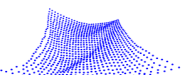
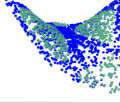


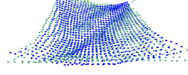

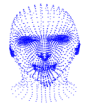

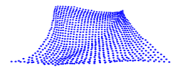
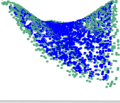
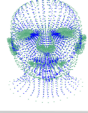
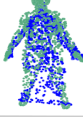
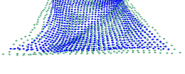



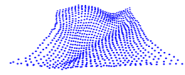
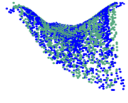

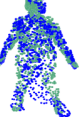
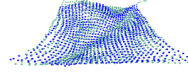
		<i>thin plate</i>	<i>FLAME</i>	<i>DFAUST</i>	<i>cloth</i>
Inputs	Template				
	Reference				
DispVoxNets (Ours)	Output				
	Overlay				
NR-ICP	Output				
	Overlay				
CPD	Output				
	Overlay				
CPD (FGT)	Output				
	Overlay				
GMR	Output				
	Overlay				

Figure II: Qualitative comparison of our DispVoxNets approach and other methods (NR-ICP [3], CPD/CPD with FGT [10] and GMR [7]). The input samples from each dataset are shown in the top rows, followed by the results (aligned templates and overlaid samples) for every method.

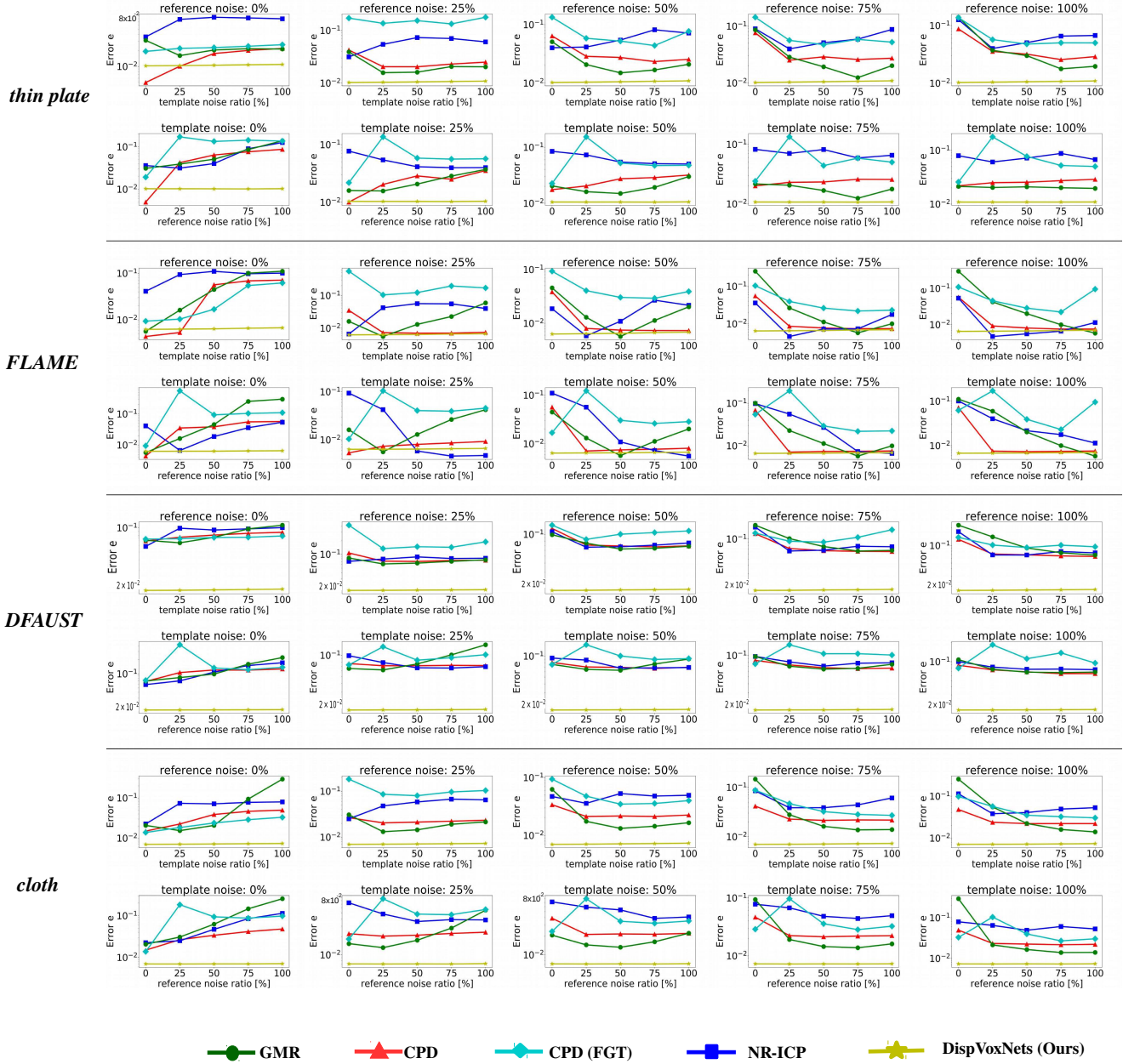


Figure III: RMSE (e) graphs for additional experiments with uniform noise on *thin plate* [6], *FLAME* [8], *DFAUST* [4] and *cloth* [2] datasets. $p\%$ is the ratio between the number of added points and the number of points in the sample. In this experiment, both reference and template are augmented with noise.

forms CPD only on *DFAUST* according to Table 2 of the paper (the experiment with no added noise). In Fig. III, we recognise multiple cases when NR-ICP outperforms CPD also on *FLAME* (the blue curve is below the red curve).

On the other hand, our approach with DispVoxNets shows almost constant error through all noise ratio combinations and all datasets. Compared to the case without noise, it even achieves the lowest RMSE on *FLAME* for

multiple noise level combinations ($\sim 40\%$ of the cases). As our network becomes aware of class-specific features after the training and learns to ignore noise, it can distinguish the meaningful shapes from noise, which contributes to its overall robustness. To the best of our belief, it is for the first time that a NRPSR method is so stable, even in the presence of large amount of noise in the data. Recall that we follow a simple noise augmentation policy for the training

data (Sec. 3.2 of the paper). Thus, our framework seemingly learns to filter uniform noise. Another factor could be that individual unstructured points cause neuron deactivations. In future work, it could be interesting to study augmentation policies for further types of noise (*e.g.*, Gaussian noise along the surfaces or mixed pixel noise).

References

- [1] S. A. Ali, V. Golyanik, and D. Stricker. Nrga: Gravitational approach for non-rigid point set registration. In *International Conference on 3D Vision (3DV)*, pages 756–765, 2018. 2
- [2] J. Bednářík, P. Fua, and M. Salzmann. Learning to reconstruct texture-less deformable surfaces from a single view. In *International Conference on 3D Vision (3DV)*, 2018. 2, 4
- [3] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14(2):239–256, 1992. 3
- [4] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4
- [5] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding (CVIU)*, 89(2-3):114–141, 2003. 2
- [6] V. Golyanik, S. Shimada, K. Varanasi, and D. Stricker. Hdmnet: Monocular non-rigid 3d reconstruction with learned deformation model. In *Virtual Reality and Augmented Reality (EuroVR)*, pages 51–72, 2018. 2, 4
- [7] B. Jian and B. C. Vemuri. A robust algorithm for point set registration using mixture of gaussians. In *International Conference for Computer Vision (ICCV)*, pages 1246–1251, 2005. 2, 3
- [8] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 36(6), 2017. 2, 4
- [9] J. Ma, J. Zhao, J. Jiang, and H. Zhou. Non-rigid point set registration with robust transformation estimation under manifold regularization. In *Conference on Artificial Intelligence (AAAI)*, pages 4218–4224, 2017. 2
- [10] A. Myronenko and X. Song. Point-set registration: Coherent point drift. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010. 2, 3