Re-rendering Reality
Research Statement

Tali Dekel
tdekel@google.com, talidekel@csail.mit.edu

Computer vision is traditionally focused on *analyzing* images and videos, for example recognizing, segmenting or tracking objects, estimating motion or recovering the 3D geometry of a scene. My focus is to go beyond that by not only obtaining an understanding about our world from digital information but also enhancing the way we perceive it. My goal is to *"Re-render Reality"* such that certain properties that are not available or are difficult to perceive from the input data are better conveyed in the new, re-rendered output. For example, by removing an occluder from an image, we can better perceive the object behind it. Similarly, by separating and muting interference sound in a video, we can focus our audio-visual attention on a desired sound source. My work combines ideas from signal processing, optimization, machine learning and computer graphics for developing new analysis and re-rendering techniques. It is aimed at both scientific discovery and technological innovation for enhancing our visual and auditory perception.

In this statement, I review my past, current and future research work, starting from my Ph.D., through my post-doctoral studies and my current tenure in Google, up to my future research goals. While my Ph.D. research has been focused on multi-view geometry, it planted the seeds to my recent and current work in the scope of *Re-rendering Reality*.

1 My Ph.D. Research – Multiple Views to the Rescue

Below is a very brief overview of my Ph.D. thesis, which includes 4 papers that have been published in top-tier conferences in the field of computer vision, two of which have been presented as oral presentations (with less than 3% acceptance rate). In addition, I have extended and published 3 of these papers in top computer vision journals [14, 12, 13].

My Ph.D. research has been focused on developing geometric-based methodologies that extract and analyze image data captured from multiple cameras. The information gained from multiple views allows us to expand our knowledge beyond a single 2D projection of the scene, hence is important to modeling and recovering the 3D properties of the scene.

During my Ph.D., I have worked with multi-camera systems under different configurations, ranging from fully controlled and dedicated camera arrays, to uncontrolled distributed set of cellphone cameras, which we named *CrowdCam*. I have addressed a variety of tasks, including fundamental computer vision problems such as *scene flow estimation*, i.e., accurately recovering the structure and 3D motion of a dynamic scene [5, 14, 4]; as well as innovative tasks such as *photo sequencing*, i.e., estimating the temporal order of an image collection taken by a CrowdCam [10, 13, 11]. Using the estimated temporal order, we then re-rendered the distributed set of images as a temporally coherent sequence, which allows us to better perceive the action that occurred in the scene.

I demonstrated that information from multiple views can be useful not only for recovering spatiotemporal properties of the scene, but also for modifying them in a 3D-aware manner. For instance,

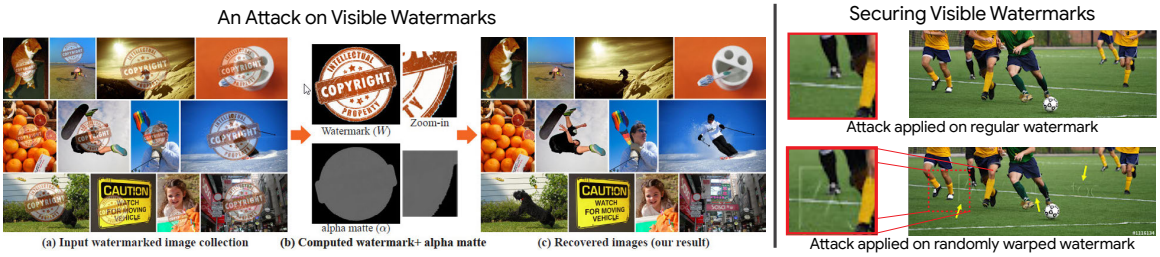

Figure 1: Left: From observing only watermarked images on the Web, the attack algorithm extracts the common watermark pattern and recovers the original images with high accuracy. Right: subtle, random geometric perturbation of the watermark at each image makes the watermark removal significantly more challenging and the attack results in noticeable artifacts.

in [12, 9], I defined and addressed the problem of *stereo retargeting* – changing the aspect ratio of a stereo image pair in order to fit it to a new display. We performed careful 3D analysis and developed a method that guarantees, as we formally proved, that the re-rendered stereo pair is geometrically consistent with a feasible 3D scene, similar to the original one. Hence, the new stereo pair can be viewed on a stereoscopic display or further processed by any computer vision algorithm that uses a stereo pair. This is an example where geometry is key for re-rendering.

My perspective geometry understanding and multi-camera systems experience became the pillars of my recent, current and future research.

2 Recent and Current Research

I next discuss some major research directions that I have been pursing in the past few years, during my post-doc in MIT and my current tenure at Google. These works include 10 papers that have been published at top-tier conferences in computer vision and graphics (CVPR, SIGGRPAH, SIGGRPAH-Asia, ICCP, UIST), most of which have also been extensively covered by the media.

All these research projects are within the scope of *"Re-rendering Reality"*, each with its own application, novel methodology for analyzing the scene and re-rendering it. On the technical side, my work can be generally divided into: (i) *model-driven* algorithms which typically involve rigorously formulating an objective function and solving for the target unknowns using advanced optimization techniques, and (ii) *data-driven* algorithms which utilize recent developments in deep learning and neural networks. Such algorithms aim at optimizing a single complex, nonlinear function with millions of parameters (i.e., a deep neural network) by training on large amounts of data.

I will next briefly describe few of these projects.

Securing Visible Watermarks

Detecting and removing obstructors from images (e.g., rain drops, reflections or fine occluding objects) allows us to better perceive the underlying scene. In this work, I considered a different type of obstructors – *visible watermarks*. Such watermarks are widely-used for marking and protecting copyrights of many millions of stock images that are offered online daily. The watermarking scheme involves overlaying a semi-transparent image containing a logo or branding on a source image. The watermark image typically contains complex structures (e.g., thin lines and shadows) and added to the source image with low, spatially varying opacity. These properties make it extremely challenging to remove a watermark from a an image without user supervision or a-priori information. However,

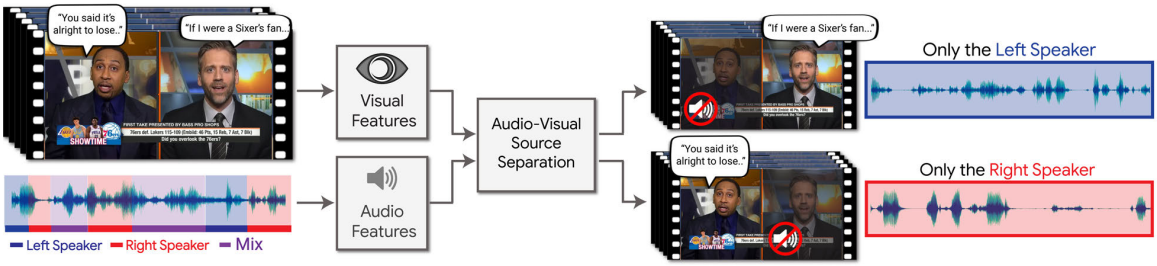

Figure 2: The input to our audio-visual model is a video of two or more people speaking simultaneously. Our model first extracts visual features from the video frames and audio features from the mixed (single) audio track; these features are then fused and processed jointly to estimate the clean audio track for each of the speakers. The video is then re-rendered with the speech of a desired speaker while muting all other sounds.

in this work [16, 1], we revealed an inherent security flaw in this watermarking scheme–*watermarks are typically added in a consistent manner to many images*. This consistency allows to get past this protection and remove watermarks completely automatically (Figure. 2 left).

To reveal this vulnerability, I took a principled approach, i.e., derived the watermarking formation model and designed a new *multi-image matting* algorithm that takes a watermarked image collection as the only input and automatically decompose the observed signals into the "foreground" (watermark), its alpha matte (opacity), and the "background" (original) images. I showed that the coherency of the watermark pattern over many images is a key in resolving this ill-posed decomposition problem and achieving high-quality reconstruction of the watermark-free images.

Since such an attack relies on the consistency of watermarks across image collection, I explored and evaluated how it is affected by various types of inconsistencies in the watermark embedding. We concluded that per-image random geometric perturbations are an effective way to make watermarks more secure (Figure 2 right). Our work had a noticeable practical impact on the stock industry which became aware of the loophole we discovered. Some of the market leaders such as Shutterstock already *deployed our proposed protection on more than 150 million stock images.*[1]

Looking to Listen – Audio-Visual Model for Speech Separation

In our watermark work, an image signal is decomposed into its components and only one of them is re-rendered. Here, we do the same only for an audio signal. We decompose a single audio track with mixed speech into its clean speech tracks; we then *re-render a video in which the speech of specific people is enhanced while all other sounds are suppressed.* This new effect allows us to better perceive a desired speaker in challenging scenarios where people are talking over each other, or talking in a noisy environment. This capability, also known as the *cocktail party effect* [6], comes natural to us humans. However, achieving it computationally remains a difficult challenge.

The key insight in our work was to *utilize visual signals to help process the audio signal.* The additional visual information has two major advantages over an audio only method: (i) *improved separation quality*, especially in the case of mixed speech. (ii) *speaker-association:* the video, by design, provides in "one shot" an association between the separated speech tracks and their speakers' faces in the video. If only audio is used, the separated audio tracks still have to be matched with their corresponding speaker in the video.

[1] https://image.shutterstock.com/z/stock-photo--years-old-boy-walking-on-beach-wooden-pier-ready-to-swim-in-a-sea-summer-outdoor-activities-with-382879477.jpg

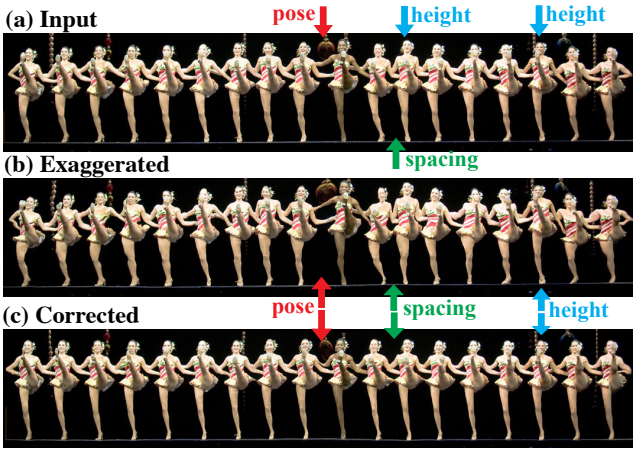

Figure 3: Given a single image as input, our algorithm detects and modify slight differences in the poses, heights, and spacing between a line of dancers performing the same high leg-kick routine (a). In our exaggerated output (b), these various differences clearly stand out. Our corrected output (c) provides an idealized version of the image, in which the non-local variations are reduced.

We used the power of machine learning and data-driven approach to perform this complex separation task and introduced the first audio-visual *speaker independent* separation model[2]. Specifically, we designed and trained a deep audio-visual model that takes a video with a single audio track as input, and outputs the clean speech tracks for each face detected in the video (Figure 2). Our model is trained only once in a supervised manner, i.e., by directly regressing to the ground truth clean speech. This approach requires a large dataset of face videos with mixed speech along with the ground truth clean speech, which did not exist prior to our work. Thus, we built an automatic framework to generate such training data from videos on the Web. The idea is to first collect high-quality, short clips of visible speakers and clean speech and then generate *"synthetic cocktail parties"* by mixing several such clean speech tracks and adding background noise. Using our automatic framework, based on face and audio processing, we extracted over 1500 hours of clean speech from over 100,000 high-quality educational videos mined from YouTube (e.g., TED Talks). This collection forms *the first large-scale audio-visual speech dataset*, which we recently released for public academic use.

Our model works well in a variety of real-world scenarios that involves heated interviews, noisy bars, and screaming children, only requiring the user to select the face of the person in the video whose speech they want to isolate. I envision a wide range of applications for our technology, from speech enhancement and recognition in videos, through video conferencing, to improved hearing aids, especially in situations where there are multiple people speaking. Some of these future directions, which I will discuss in more details in Section 3, have also been featured in Google I/O event by Sundar Picahi, Google's CEO (see [2]).

Geometry based Re-rendered Reality

In the previous two sections I showed how decomposing an audio/image signal into its components can be used to better perceive certain properties of the data. In a different line of work, I have been developing techniques that rely on geometric analysis rather than signal decomposition for revealing physical imperfections that are hard to notice in the input image [8, 22], or for visualizing human motion in video [23].

It is useful to detect and visualize physical imperfections in photos. For example, buildings may seem to be perfectly straight, but in reality may be tilted due to structural damage. Detecting such imperfection ahead of time in everyday photos can be useful for preventing further damage. My goal in this new line of work is to detect such geometric imperfections, given only a single image as input,

[2]Previous deep audio-video speech separation models are speaker dependent, i.e., they train a dedicated model for each speaker they want to isolate. Our model is trained once and can be applied to any speaker.

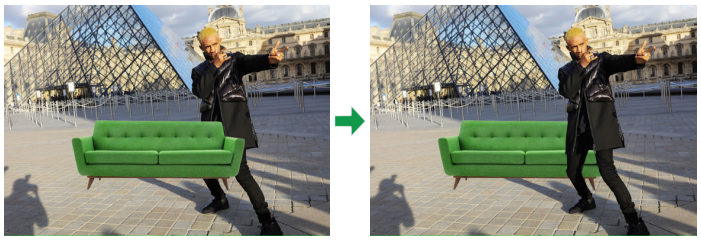

Figure 4: Humans can easily infer the right depth ordering between the sofa and the person, without having precise measurable depth of the scene. Can we make a machine do the same?

and re-render a new image in which the imperfections are exaggerated to highlight them and make them more visible. On the other direction, the imperfections can also be reduced to idealize/beautify images, which can be used as a graphics tool for creating more visually pleasing images (Figure 3).

Deviations are measured from some "perfect model" which I defined as: (i) ideal parametric geometries such as line segments, circles, ellipse, or (ii) perfect repetitions of image structures. Each of these models has lead to a new algorithm, with a wide range of applications in civil engineering, astronomy, and materials defects inspection. I recently extended the single image algorithm to multiple views [21], hence allowing to detect and highlight variations between repeating structures and patterns in multiple images of the same scene.

While the techniques above are focused on studying static photos, there are space-time signals that are important and interesting to study and visualize, such as human motion. We developed a system called *MoSculp*, for visualizing complex human motion. Our system transform a video into a 3D *motion sculpture*–a representation that conveys the 3D structure swept by a human body as it moves through space [23]. To provide an end-to-end, easy to use workflow, I developed an algorithm that estimates the humans 3D geometry over time and developed a 3D-aware image-based rendering approach to preserve the depth ordering between the sculpture and the human as observed in the video. Motion sculptures reveal space-time information that is difficult to perceive from the input video such as how different parts of the object interact over time.

3 Future Work

I plan to continue working towards enhancing the way we perceive our digital world. I strive to make our reformed reality **richer**, more **scene-aware** (better harmonized with image content) and more **accessible** (expanding the range of applications in which re-rendered content is currently being used). As users, we want to see and hear better through modified digital content. We want computer vision to assist us by easily and intuitively displaying information about our world. We want to be creative and express ourselves by designing and editing our digital content. I believe significant progress can be made in those fronts by following two main methodologies: (i) combining domain specific knowledge with recent developments in machine learning, and (ii) developing algorithms that leverage information from multiple modalities.

3.1 Integrating Computer Vision Fundamentals into Machine Learning

Deep learning is revolutionizing computer vision and computer graphics. Deep-network-based models excel at the majority of classic vision tasks and make it possible to tackle new problems, some of which may have been considered unresolvable before. Nevertheless, the common deep models lack two essential properties: *interpretability* and *predictability* of their outputs. Consequently, they often cannot be used in domain expert applications, such as medical diagnosis.

I believe that fusing unlearned, fundamental knowledge into neural networks is the key for the evolution of deep models in the field of computer vision and computer graphics. This general approach came into force in my recent works (e.g., [7]), and I strive to expand and advance it. Below are two major areas of research that can greatly benefit from "classic/modern computer vision" integration.

1. **Deep Perspective Geometry:** Recovering 3D elements of a scene from 2D images is one of the most fundamental problems in computer vision. Nevertheless, it is far from being solved. State-of-the-art multi-view stereo algorithms still rely on a set of assumptions such as brightness constancy (assume textured and Lambertian world), and are applicable only if the scene is static. I believe that replacing such traditional assumptions with data driven priors, using deep learning, is an exciting and promising direction of research. It opens the door to new problems that were intractable before (e.g., recovering depth of moving objects from monocular video). However, "deep geometry" calls for new representations, datasets and learning methodologies that will allow to intelligently integrate the fundamentals of perspective geometry and their constraints into deep networks.

   Augmented Reality (AR) is a rapidly growing area of research in the scope of re-rendering reality that can be developed significantly by combining fundamental geometry with machine learning. AR enhances our environment by embedding virtual elements into our real world. Thus, in many applications, an artifact-free and geometrically-aware compositions is critical for a rich user experience (e.g., Figure 4). Achieving this highly depends on our ability to accurately recover the depth of the scene in real world scenarios. In a current research, I am looking into new ways of predicting depth in a common AR scenario, where both (a single) camera and the humans in the scene are moving. The main idea is to leverage a large corpus of data, in which the world is static in order to learn geometric priors about the scene. These priors allow us to predict depth in cases where traditional methods are inapplicable. I believe that combining geometric knowledge with a machine learning is essential towards making AR content richer and more immersive.

2. **Bridging the Gap between Low-and High Level Vision:** Domain-specific knowledge forms the ground for developing elegant and simple algorithmic principles. For example, based on statistical properties of natural images, I defined the *Best-Buddies Similarity*–a new simple similarity measure between two sets of points [15, 19]. Based on a rigorous mathematical analysis, I demonstrated its power for template matching under occlusions and high level of noise. Adopting this similarity measure to work in a feature space learned by a neural network has recently showed promising results for matching points across different domains for image morphing [3]. This is an example of how fusing high level and low level information can be beneficial. More generally, we would like to benefit from both worlds: low level structured processing and high level information gained from learning.

   Generative models can also benefit from fusing low-level and high-level cues. While these models made a tremendous progress in image synthesis and are able to go beyond just synthesizing repetitive textures, they still struggle to capture global structural patterns. They are often restricted to a certain class of images and a specific image resolution. Furthermore, most of these methods heavily rely on the training loss (e.g., adversarial loss), which often leads to unstable training. I believe that fusing lower level unlearned information about the characteristics of natural images such as distribution of patches within the image and geometric constraints can eliminate many of these limitations.

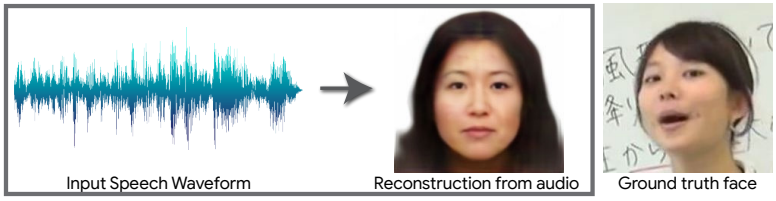

Figure 5: Learning multi-modal correlation via generation–preliminary result: by synthesizing a face image from a speech segment we can reason about the joint speech-face manifold and the hidden correlated latent attributes. We can see that age, gender and ethnicity are well reconstructed, but are there any finer attributes that are being captured (e.g., facial anatomy)?

3.2 Cross-Modal Processing

Another domain which I believe is essential for expanding and enriching re-rendered content is the utilization of multi-modal data. In our audio-visual source separation work [17], we only scratched the surface of joint audio-visual processing in the new era of deep-learning. While there has been a recent surge in developing audio-video algorithms, there are still many fundamental open research questions about the joint manifold of audio and video, their representation, as well as unexplored novel applications. Below are three example research threads I intend to pursue in the future.

1. **Re-rendering and Synthesizing Sound and Video:** Audio and image signals, despite correlated, have quite different characteristics. For instance, while audio signals provide high temporal resolution (e.g., even a cellphone's audio can be sampled at 48kHz) and low spatial resolution (e.g., provide directional information), image data is typically the opposite–the spatial resolution can be extremely high yet the temporal sampling rate is much lower. We can *leverage the converse properties of audio and video for synthesise and re-rendering.* On the *visual-to-rerender-audio* side, image data possess rich information about the scene's motion, structure, materials, object's type and location. Such information that is not contained in the audio signal, has great potential in modeling complex auditory tasks such as cancelling reverberations/echos, or extreme source separation (i.e., isolating sound in a very noisy environment). A novel task that I plan to study is *spatial audio from monocular video*–synthesizing stereo or multi-channel audio from a single microphone input and video of the scene.

2. **Variable Speed Rate Videos:** The digital content that an average person consumes on a daily basis is astonishingly growing. For example, over 5 billion hours of videos are watched on Youtube every single day. How can we watch digital content more efficiently? In YouTube, a user can watch a video at $\times 1.25$, $\times 1.5$, or $\times 2$ speed. In that case, the audio is sped up using advanced techniques that preserve the pitch, however the video content is simply sub-sampled. This leads to temporal aliasing and unnatural motions, especially in case of high frequency motion (e.g., a person speaking to a camera). I plan to develop advanced techniques that leverage both video and audio for "gracefully" speeding up videos such as TED Talks, Video Games, Unboxing videos. The sped up video ought to be: (i) pleasant/easy to watch. (ii) Maintain as much as possible the content of the original video. (iii) be synchronized with the audio. If such a method will be applied on YouTube scale it could save the world millions of hours every day, and potentially can be used to save energy and resources by compressing the videos during upload time.

3. **Learning Correlation via Cross-Modal Generation:** By examining audio and video together, we can infer important correlations between them – what object generates what

sound, what can we tell about an object from the way it sounds, etc. I suggest to *learn correlation via cross-modal generation*, i.e., by synthesizing one signal from the other. For example, by synthesizing an image of a person's face from its voice, we can learn about the specific latent properties that are encapsulated in both voice and the face of a person. Detecting and evaluating these cross-modal attributes can be useful in a variety of applications such as biometric identification, surveillance, and medical/psychological applications. In biology and human perception, studies demonstrate that humans can match a voice to a face with higher accuracy than a chance, and reveal the connection between facial and voice attributes to masculinity/femininity, health, and height [20]. This was also supported by a recent work in the field of computer vision that addressed the problem of machine cross-modal matching [18]. However, addressing the full signal synthesize task, although more challenging, may shed new light on the joint manifold of vision and sound.

References

[1] Securing visible watermarks project page. https://watermark-cvpr17.github.io/.

[2] Sundar picahi (google ceo), google i/o event, 2018. https://www.youtube.com/watch?v=ogfYd705cRs&t=6m53s.

[3] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *arXiv preprint arXiv:1805.04140*, 2018.

[4] Tali Basha, Shai Avidan, Alexander Hornung, and Wojciech Matusik. Structure and motion from scene registration. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*.

[5] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[7] **Tali Dekel**, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] **Tali Dekel**, Tomer Michaeli, Michal Irani, and William T Freeman. Revealing and modifying non-local variations in a single image. *ACM Transactions on Graphics (TOG)*, 34(6):227, 2015.

[9] **Tali Dekel**, Yael Moses, and Shai Avidan. Geometrically consistent stereo seam carving. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1816–1823. IEEE, 2011.

[10] **Tali Dekel**, Yael Moses, and Shai Avidan. Photo sequencing. In *European Conference on Computer Vision*, pages 654–667. Springer, 2012.

[11] **Tali Dekel**, Yael Moses, and Shai Avidan. Space-time tradeoffs in photo sequencing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 977–984, 2013.

[12] **Tali Dekel**, Yael Moses, and Shai Avidan. Stereo seam carving a geometrically consistent approach. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2513–2525, 2013.

[13] **Tali Dekel**, Yael Moses, and Shai Avidan. Photo sequencing. *International journal of computer vision*, 110(3):275–289, 2014.

[14] **Tali Dekel**, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *International journal of computer vision*, 101(1):6–21, 2013.

[15] **Tali Dekel**, Shaul Oron, Michael Rubinstein, Shai Avidan, and William T Freeman. Best-buddies similarity for robust template matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2029, 2015.

[16] **Tali Dekel**, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2146–2154, 2017.

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, **Tali Dekel**, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 2018.

[18] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] Shaul Oron, **Tali Dekel**, Tianfan Xue, Shai Avidan, and William T Freeman. Best-buddies similarity–robust template matching using mutual nearest neighbors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[20] Harriet MJ Smith, Andrew K Dunn, Thom Baguley, and Paula C Stacey. Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1):1474704916630317, 2016.

[21] Tal Tlusty, Tomer Michaeli, **Tali Dekel**, and Lihi Zelnik-Manor. Modifying non-local variations across multiple views. 2018.

[22] Neal Wadhwa, **Tali Dekel**, Donglai Wei, Frédo Durand, and William T Freeman. Deviation magnification: revealing departures from ideal geometries. *ACM Transactions on Graphics (TOG)*, 34(6):226, 2015.

[23] Xiuming Xhang, **Tali Dekel**, Tianfan Xue, Andrew Owen, Jiajun Wu, Stefanie Mueller, and William T. Freeman. MoSculp: Interactive Visualization of Shape and Time. *ACM symposium on User interface software and technology (UIST)*, 2018.